TMA4145 – Linear Methods
November 28, 2019

Franz Luef

Abstract. These notes are for the course TMA4145 – Linear Methods at NTNU for the fall 2019.

Contents

Introduction

The goal of this course is to present basic facts about vector spaces and mappings between vector spaces in a form suitable for engineers, scientists and mathematicians. The presentation is addressed to students with varying backgrounds.

A special emphasis is put towards general methods and on abstract reasoning. The material in this course is supposed to prepare you for the advanced courses in your respective study program. You might encounter for the first time rigorous reasoning and there will be a particular focus on definitions, statements (=lemmas, propositions, theorems) and proofs.

In the first chapter we discuss basic notions such as sets, functions and the cardinality of a set.

These notes are accompanying the course TMA4145 Linear methods and is based on earlier notes by Sigrid Grespstad and Franz Luef.

CHAPTER 1

Sets and functions

Basic definitions and theorems about sets and functions are the content of this chapter and are presented in the setting of Naive Set Theory. These notions set the stage for tuning our intuition about collections of objects and relations between these objects.

1.1. Sets

Definition 1.1.1. A *set* is a collection of distinct objects, its *elements*. If an object x is an element of a set X, we denote it by $x \in X$. If x is not an element of X, then we write $x \notin X$.

A set is uniquely determined by its elements. Suppose X and Y are sets. Then they are identical, $X = Y$, if they have the same elements. More formalized, $X = Y$ if and only if for all $x \in X$ we have $x \in Y$, and for all $y \in Y$ we have $y \in X$.

Definition 1.1.2. Suppose X and Y are sets. Then Y is a subset of X, denoted by $Y \subseteq X$, if for all $y \in Y$ we have $y \in X$.

If $Y \subseteq X$, one says that Y is contained in X. If $Y \subseteq X$ and $X \neq Y$, then Y is a proper subset of X and we use the notation $Y \subset X$. The most direct way to prove that two sets X and Y are equal is to show that

$$x \in X \iff x \in Y$$

for any element x. (Another way is to prove a double inclusion: if $x \in X$ then $x \in Y$, establishing that $X \subseteq Y$ and if $x \in Y$, then $x \in X$, establishing that $Y \subseteq X$.)

The *empty set* is a set with no elements, denoted by $\emptyset$.

Proposition 1.1. There is only one empty set.

Proof. Suppose E_1 and E_2 are two empty sets. Then for all elements x we have that $x \notin E_1$ and $x \notin E_2$. Hence $E_1 = E_2$. □

Some familiar sets are given by the various number systems:

(1) $\mathbb{N} = \{1, 2, 3, ...\}$ the set of natural numbers, $\mathbb{N}_0 = \{0, 1, 2, 3, ...\}$;
(2) $\mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$ the set of integers;
(3) $\mathbb{Q} = \{p/q : p \in \mathbb{Z}, q \in \mathbb{N}\}$ the set of rational numbers;
(4) $\mathbb{R}$ denotes the set of real numbers;
(5) $\mathbb{C}$ denotes the set of complex numbers.

For real numbers a, b with $a < b < \infty$ we denote by $[a, b]$ the closed bounded interval, and by (a, b) the open bounded interval. The length of these bounded intervals is $b - a$.
Here are a few constructions related to sets.

Definition 1.1.3. Let X and Y be sets.

- The *union* of X and Y, denoted by $X \cup Y$, is defined by
$$X \cup Y = \{z |\, z \in X \quad \text{or } z \in Y\}.$$
- The *intersection* of X and Y, denoted by $X \cap Y$, is defined by
$$X \cap Y = \{z |\, z \in X \quad \text{and } z \in Y\}.$$
- The *difference set* of X from Y, denoted by $X \backslash Y$, is defined by
$$X \backslash Y = \{z \in X : z \in X \quad \text{and } z \notin Y\}.$$
If all sets are contained in one set X, then the difference set $X \backslash Y$ is called the *complement* of Y and denoted by Y^c.
- The *Cartesian product* of X and Y, denoted by $X \times Y$, is the set
$$X \times Y = \{(x, y) |\, x \in X, y \in Y\},$$
i.e the set of all ordered pairs (x, y), with $x \in X$ and $y \in Y$. An ordered pair has the property that $(x_1, y_1) = (x_2, y_2)$ if and only if $x_1 = x_2$ and $y_1 = y_2$.
- $\mathcal{P}(X)$ denotes the set of all subsets of X.

If we two sets X and Y have an empty intersection, $X \cap Y = \emptyset$, we say that X and Y are *disjoint*. Here are some basic properties of sets.

Lemma 1.2. Let X, Y and Z be sets.

(1) $X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$ and $X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$ (*distribution law*)
(2) $(X \cup Y)^c = X^c \cap Y^c$ and $(X \cap Y)^c = X^c \cup Y^c$ (*de Morgan's laws*)
(3) $X \backslash (Y \cup Z) = (X \backslash Y) \cap (X \backslash Z)$ and $X \backslash (Y \cap Z) = (X \backslash Y) \cup (X \backslash Z)$
(4) $(X^c)^c = X$.

Proof. (2) Let us prove one of de Morgan's relations. Let us use the most direct approach. Keep in mind that $x \in E^c \iff x \notin E$. We then have:

$$x \in (X \cup Y)^c \iff x \notin X \cup Y \iff x \notin X \textbf{ and } x \notin Y$$
$$\iff x \in X^c \text{ and } x \in Y^c \iff x \in X^c \cap Y^c.$$

This proves the identity.

(4)
$$x \in (X^c)^c \iff x \notin X^c \iff x \in X.$$

□

Note that if you have a statement involving $\cup$ and $\cap$. Then you get another true statement if you interchange $\cup$ with $\cap$ and $\cap$ with $\cup$, as one can see in the lemma. This is part of the field of Boolean algebra.

1.2. Functions

Let X and Y be sets. A *function* f from X to Y, written $f : X \to Y$, is a relation between the elements of X and Y, i.e. $f \subset X \times Y$, satisfying the following property: for all $x \in X$, there is a unique $y \in Y$ such that $(x, y) \in f$. We denote $(x, y) \in f$ by $f(x) = y$.
X is the *domain* of f, and Y is *codomain* of f. By definition, for each $x \in X$ there is exactly one $y \in Y$ such that $f(x) = y$. We say that y is the *image* of x under f. The *graph* $G(f)$ of a function f is the subset of $X \times Y$ defined by

$$G(f) = \{(x, f(x)) | \, x \in X\}.$$

The *range* of a function $f : X \to Y$, denoted by range(f), or $f(X)$, is the set of all $y \in Y$ that are the image of some $x \in X$:

$$\text{range}(f) = \{y \in Y | \, \text{there exists} \, x \in X \, \text{such that} \, f(x) = y\}.$$

The *pre-image* of $y \in Y$ is the subset of all $x \in X$ that have y as their image. This subset is often denoted by $f^{-1}(y)$:

$$f^{-1}(y) = \{x \in X | \, f(x) = y\}.$$

Note that $f^{-1}(y) = \emptyset$ if and only if $y \in Y \backslash \text{ran}(f)$.

Lemma 1.3. Let $f \colon X \to Y$ be a function and let $C, D \subset Y$. Then

$$f^{-1}(C \cup D) = f^{-1}(C) \cup f^{-1}(D).$$

Proof.

$$\begin{aligned} x \in f^{-1}(C \cup D) &\iff f(x) \in C \cup D \iff f(x) \in C \text{ or } f(x) \in D \\ &\iff x \in f^{-1}(C) \text{ or } x \in f^{-1}(D) \iff x \in f^{-1}(C) \cup f^{-1}(D). \end{aligned}$$

□

Here are some simple examples of functions.

$$|x| = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -x & \text{if } x < 0. \end{cases}$$

Note that $|x| = \max\{x, -x\}$. We define the positive, x^+ and negative part, x^- of $x \in \mathbb{R}$:

$$x^+ = \max\{x, 0\}, \quad \text{and} \quad x^- = \max\{-x, 0\},$$

so we have $x = x^+ - x^-$ and $|x| = x^+ + x^-$.

The following notions are central for the theory of functions.

Definition 1.2.1. Let $f : X \to Y$ be a function.

(1) We call f *injective* or one-to-one if $f(x_1) = f(x_2)$ implies $x_1 = x_2$, i.e. no two elements of the domain have the same image. Equivalently, if $x_1 \neq x_2$, then $f(x_1) \neq f(x_2)$.

(2) We call f *surjective* or *onto* if $\mathrm{ran}(f) = Y$, i.e. each $y \in Y$ is the image of at least one $x \in X$.
(3) We call f *bijective* if f is both injective and surjective.

Note that a bijective function matches up the elements of X with those of Y so that in some sense these two sets have the same number of elements.

Let $f : X \to Y$ and $g : Y \to Z$ be two functions so that the range of f coincides with the domain of g. Then we define the *composition*, denoted by $g \circ f$, as the function $g \circ f : X \to Z$, defined by $x \mapsto g(f(x))$.

For every set X, we define the *identity map*, denoted by id_X or id where $\mathrm{id}(x) = x$ for all $x \in X$.

If one has a function f that maps elements in X to Y, then it is often desirable to reverse this assignment. Let us introduce some notions to address this basic problem.

Definition 1.2.2. Let f be a function from X to Y.

- The mapping f is said to be *left invertible* if there exists a function $g : Y \to X$ such that $g \circ f = \mathrm{id}_X$. We call g a *left inverse* of f and denote it by f_l^{-1}.
- The mapping f is said to be *right invertible* if there exists a function $h : Y \to X$ such that $f \circ h = \mathrm{id}_Y$. We call h a *right inverse* of f and denote it by f_r^{-1}.
- The mapping f is said to be *invertible* if there exists a function $g : Y \to X$ such that $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$, the so-called *inverse* of f denoted f^{-1}.

One may think of a left and right inverse in layman terms: (i) If you map an element of the domain via a function to an element in the target space, then the left inverse tells you how to go back to where you started from;(ii) If one wants to get to a point in the target, then the right inverse tells you a possible place to start in the domain. The inverse of a function has some important properties.

Lemma 1.4. Given an invertible function $f : X \to Y$.
(1) The inverse function $f^{-1} : Y \to X$ is unique.
(2) The inverse function is also invertible and we have $(f^{-1})^{-1} = f$.

PROOF. (1) Suppose there are two inverse functions $g_i : Y \to X$, $i = 1, 2$. By assumption we have that $f \circ g_1 = \mathrm{id}_Y$ and $g_2 \circ f = \mathrm{id}_X$. Hence we have

$$g_2(y) = g_2((f \circ g_1)(y)) = g_2(f(g_1(y))) = g_1(y) \quad \text{for all } y \in Y,$$

i.e. $g_1 = g_2$.
(2) Exercise.

□

Lemma 1.5. Let $f : X \to Y$ and $g : Y \to Z$ be two invertible function. Then $g \circ f$ is also invertible and $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

Let us give a description of left, right invertibility and invertibility in more concrete terms.

Proposition 1.6. Given a function $f : X \to Y$ between two non-empty sets X and Y.

(1) f is left invertible if and only if it is injective.
(2) f is right invertible if and only if it is surjective.
(3) f is invertible if and only if it is injective and surjective, i.e. if f is bijective.

PROOF. (1) Let us assume that f is injective. Then $f : X \to \mathrm{ran}(f)$ is invertible with $f^{-1} : \mathrm{ran}(f) \to X$. Let $g : Y \to X$ be any extension of this inverse. Then $g \circ f = \mathrm{id}_X$.
Suppose f is left invertible. Assume there are $x_1, x_2 \in X$ such that $f(x_1) = f(x_2) = y$. Then

$$x_1 = f_l^{-1}(f(x_1)) = f_l^{-1}(f(x_2)) = x_2,$$

i.e. f is injective.

(2) Let us assume that f is surjective. Pick an arbitrary element $z \in Y$, wich is by assumption an element of $\mathrm{ran}(f)$. Hence z has at least one pre-image in X and thus $f^{-1}(z) \neq \emptyset$.Take $y_1 \neq y_2$. Then the sets $f^{-1}(\{y_1\})$ and $f^{-1}(\{y_2\})$ in X are disjoint. Let us pick from each set $f^{-1}(\{y\})$ an element x and define $x := h(y)$. Then $h : Y \to X$ and $f \circ h = \mathrm{id}_Y$.
Suppose that f is right invertible. Then we have for $y \in Y$ that $f(f_r^{-1}(y)) = f(x)$ where we set x to be $x = f_r^{-1}(y)$. In other words, y is in the range of f.

(3) Follows from the other assertions.

□

A consequence of the characterizations of left and right invertibility is the observation:

Remark 1.2.3. If $f : X \to Y$ is left invertible mapping between non-empty sets such that $\mathrm{ran}(f) \neq Y$, then there are many left inverses. However the restriction of any left inverse of f to $\mathrm{ran}(f)$ is unique.
One the other hand if $f : X \to Y$ is right invertible such that f is surjective but not injective, then f will have many right inverses.

Our study of linear mappings will provide ample examples of the aforementioned notions. Here we just give one example.

Example 1.2.4. Given the linear mapping $T : \mathbb{R}^2 \to \mathbb{R}^3$ given by $T = Ax$ with

$$A = \begin{pmatrix} -3 & -4 \\ 4 & 6 \\ 1 & 1 \end{pmatrix}.$$

Then the matrix

$$A_l^{-1} = \frac{1}{9} \begin{pmatrix} -11 & -10 & 16 \\ 7 & 8 & -11 \end{pmatrix}$$

induces a left inverse T_l^{-1} of T.
This left inverse is not unique, for example

$$\frac{1}{2}\begin{pmatrix} 0 & -1 & 6 \\ 0 & 1 & -4 \end{pmatrix}$$

also gives a left inverse. One can turn this example into one for right inverses as well, see problem set 1.

1.3. Cardinality of sets

Bijective functions provide us with a tool for comparing the sizes of different sets. We start with the case of finite sets.

Definition 1.3.1. Two finite sets X and Y have equal cardinality, if there is a bijective map $f : X \to Y$. If there is an injective map from X to Y, then we say that the cardinality of X is less than or equal to the cardinality of Y.

A set X has n elements if there is a bijection between X and the set $\{0, 1, ..., n-1\}$.

Proposition 1.7. If there is a bijection between the sets $\{0, 1, \dots n-1\}$ and $\{0, 1, \dots m-1\}$, then $n = m$ (i.e. they have the same number of elements).

Proof. We proceed by induction. For $n = 0$ the set $n = \{0, 1, ..., n-1\}$ is the empty set, and thus the only set bijective with it is the empty set. Suppose that $n > 0$ and that the result is true for $n-1$. Moreover, suppose that there is a bijection $f : \{0, 1, ..., n-1\} \to \{0, 1, ..., m-1\}$. Assume first that $f(n-1) = m-1$. Then the restriction of f to the set $\{0, 1, ..., n-2\}$ gives a bijection to $\{0, 1, ..., m-2\}$, and by the induction hypothesis we have $n-1 = m-1$.

Let us now look at the case when $f(n-1) \neq m-1$. We have that $f(n-1) = a$ for some a and $f(b) = m-1$ for some b, and we define a function $\tilde{f}$ by $\tilde{f}(x) = f(x)$ if $x \neq b, n-1$; $\tilde{f}(b) = a$ and $\tilde{f}(n-1) = m-1$. Then $\tilde{f}$ is a bijection and we conclude as above with $n = m$. □

Let us now define *countable* sets.

Definition 1.3.2. A set X is *countable* if there exists an injective map from X to $\mathbb{N}$. In other words, X is countable if we can arrange its elements in a (possibly infinite) sequence $\{x_1, x_2, x_3, ...\}$ where each element occurs exactly once.

Remarks.
(1) Equivalently, X is countable if there exists a surjective map from $\mathbb{N}$ to X.
(2) As illustrated by the examples below, countable sets can be either finite or infinite.

Examples 1.3.3.
(1) Any finite set of elements $X = \{x_1, \dots, x_n\}$ is countable, as the map $f : X \to \mathbb{N}$ defined by $f(x_i) = i$ is injective.
(2) The infinite set of squares $X = \{1, 4, 9, ..., n^2, ...\}$ is countable, as the map $f : \mathbb{N} \to X$ defined by $f(n) = n^2$ is surjective.

(3) The infinite set of odd numbers $X = \{1, 3, 5, ..., 2n-1, ...\}$ is countable, since $f : \mathbb{N} \to X$ defined by $f(n) = 2n - 1$ is a surjective map.

Note that in examples (2) and (3) above, the map f is in fact a bijection.

Proposition 1.8. $\mathbb{N} \times \mathbb{N}$ is countable.

PROOF. The argument starts out with decomposing $\mathbb{N}\times\mathbb{N}$ into finite sets $F_2, ...,$ where

$$F_k = \{(i, j) \in \mathbb{N} \times \mathbb{N} |\, i + j = k\}$$

and the cardinality of F_k is $k - 1$. Now we arrange these sets: first writing the one element of F_2, then the two elements of F_2 and so forth. Hence, we have established the assertion. In other words, we have arranged $\mathbb{N} \times \mathbb{N}$ in a table:

$$\begin{array}{ccccc} (1,1) & (1,2) & (1,3) & (1,4) & \cdots \\ (2,1) & (2,2) & (2,3) & (2,4) & \cdots \\ (3,1) & (3,2) & (3,3) & (3,4) & \cdots \\ (4,1) & (4,2) & (4,3) & (4,4) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

and list the elements along sucessive (anti-)diagonals from bottom-left to top-right as

$$(1, 1), (2, 1)(1, 2), (3, 1), (2, 2), (1, 3), ...$$

We define $f : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ by $f(n) := nth$ pair in this order. Note that f is a bijection. □

Proposition 1.9. We have the following assertions:

(1) The Cartesian product of two countable sets is countable.
(2) The union of countably many countable sets is countable.

PROOF. (1) Exercise.

(2) Let $X_1, ...$ be a countable family of countable sets. We denote the elements of X_i by $\{x_{1i}, x_{2i}, ...\}$ for $i = 0, 1, ...$ and define a map by $f(i, j) = x_{ij}$. Note that $f : \mathbb{N} \times \mathbb{N} \to \cup_{i=1}^{\infty} X_i$ is surjective and thus the union $\cup_{i=1}^{\infty} X_i$. is countable. The map f is not injective in general, because the X_i's need not to be disjoint. The proposition preceding this statement yields the desired claim.

□

Proposition 1.10. The sets $\mathbb{Z}$ of integers and $\mathbb{Q}$ of rational numbers are countable.

PROOF. Exercise. □

Bernstein and Schröder observed an elementary characterization of two sets having the same cardinality. We state it without proof.

Theorem 1.11. Let X and Y be two sets. Suppose there are injective maps $f : X \to Y$ and $g : Y \to X$. Then there exists a bijection between X and Y.

If a set X is **not** countable, then we say that it is *uncountable*.

Theorem 1.12 (Cantor)**.** The set $\mathbb{R}$ of real numbers is uncountable.

Proof. We argue by contradiction and assume that $\mathbb{R}$ is countable. Then a subset of $\mathbb{R}$ is also countable. Thus the open interval $(0, 1)$ is a countable set, i.e.

$$(0, 1) = \{x_0, x_1, ...\}.$$

Any $a_i \in (0, 1)$ has an infinite decimal expansion (possibly terminating, in which case we let it continue forever with zeros):

$$a_i = 0.a_{i0}a_{i1}..., \qquad a_{ij} \in \{0, 1, ..., 9\}.$$

We set b_i to be

$$b_i = \begin{cases} 3 & \text{if } a_{ii} \neq 3 \\ 1 & \text{if } a_{ii} = 3. \end{cases}$$

By construction we have $b_i \neq a_{ii}$ and thus the number

$$a = 0.b_1b_2...$$

differs from all a_i. Note that $a \in (0, 1)$ which is not included in the given enumeration of $(0, 1)$. Hence we have deduced a contradiction to the countability of $(0, 1)$. □

Proposition 1.13. Let X be the set of all binary sequences: $X = \{(a_1, a_2, a_3, ...) : a_i \in \{0, 1\}\}$. Then X is not countable.

Proof. We apply the method from the preceding theorem, aka diagonal argument.
Suppose $X = \{(x_1, x_2, x_3, ...) : x_i \in \{0, 1\}\}$ is countable. Then we have

$$x_1 = 010100....$$
$$x_2 = 101111....$$
$$\vdots$$

Then we define a sequence $x \notin X$ by moving down the diagonal and switching the values from 0 to 1 or from 1 to 0. Hence X is uncountable. □

Proposition 1.14. The power set $\mathcal{P}(\mathbb{N})$ of the natural numbers $\mathbb{N}$ is uncountable.

Proof. Let $C = \cup_{n\in\mathbb{N}} X_n$ be a countable collection of subsets of $\mathbb{N}$. Define $X \subset \mathbb{N}$ by

$$X = \{n \in \mathbb{N} : n \notin X_n\}.$$

Claim: $X \neq X_n$ for every $n \in \mathbb{N}$. Since either $n \in X$ and $n \notin X_n$ or $n \notin X$ and $n \in X_n$.
Thus $X \notin C$ and so no countable collection of subsets of $\mathbb{N}$ includes all of the subsets of $\mathbb{N}$. □

1.4. Supremum and infimum

We introduce two crucial notions: the infimum and supremum of a set. First we provide some preliminaries.

Definition 1.4.1. Let A be a non-empty subset of $\mathbb{R}$

- If there exists $m \in \mathbb{R}$ such that $m \leq a$ for all $a \in A$, then m is a *lower bound* of A. We call A *bounded below*.
- If there exists $M \in \mathbb{R}$ such that $a \leq M$ for all $a \in A$, then M is an *upper bound* of A. We call A *bounded above*.
- If there exist lower and upper bounds, then we say that A is *bounded*.

Definition 1.4.2 (Infimum and Supremum)**.** Let A be a subset of $\mathbb{R}$.

- If m is a lower bound of A such that $m \geq m'$ for every lower bound m', then m is called the *infimum* of A, denoted by $m = \inf A$. Furthermore, if $\inf A \in A$, then we call it the minimum of A, and write $\min A$.
- If M is an upper bound of A such that $M' \geq M$ for every upper bound M', then M is called the *supremum* of A, denoted by $M = \sup A$. Furthermore, if $\sup A \in A$, then we call it the maximum of A, and write $\max A$.

It follows from this definition that the supremum of a set A is its **least upper bound**, whereas the infimum is its **greatest lower bound**. Note that the infimum and supremum of a set A are both unique. The argument is left as an exercise.

If $A \subset \mathbb{R}$ is not bounded above, then we define $\sup A = \infty$. If $A \subset \mathbb{R}$ is not bounded below, then we assign $-\infty$ as its infimum.

We state a different formulation of the notions $\inf A$ and $\sup A$ which is simply a reformulation of the definition above.

Lemma 1.15. Let A be a subset of $\mathbb{R}$.

- Suppose A is bounded above. Then $M \in \mathbb{R}$ is the supremum of A if and only if the following two conditions are satisfied:
  (1) For every $a \in A$ we have $a \leq M$.
  (2) Given $\varepsilon > 0$, there exists $a \in A$ such that $M - \varepsilon < a$.
- Suppose A is bounded below. Then $m \in \mathbb{R}$ is the infimum of A if and only if the following two conditions are satisfied:
  (1) For every $a \in A$ we have $m \leq a$.
  (2) Given $\varepsilon > 0$, there exists $a \in A$ such that $a < m + \varepsilon$.

Here is another way to phrase the statement for the supremum of a set.

Lemma 1.16. Let A be a non-empty subset of $\mathbb{R}$ that is bounded above. Any upper bound M of A is the supremum of A if and only if for every $m < M$, there exists an element $x \in A$ such that $m < x \leq M$.

PROOF. Suppose $M = \sup A$. If $m < M$, then m is not an upper bound of A. Thus there exists an element $x \in A$ such that $x > m$. On the other hand, since M is an upper bound of A we have $x \leq M$.
Conversely, if M is an upper bound of A satisfying the stated condition, then every $m < M$ is not an upper bound of A. Thus $M = \sup A$. □

Lemma 1.17. Suppose A is a bounded subset of $\mathbb{R}$. Then $\inf A \leq \sup A$

PROOF. Let $a \in A$ (we assume that the set A is non-empty, otherwise there is nothing interesting here). Then as a lower bound for A, $\inf A \leq a$. Moreover, as an upper bound for A, $a \leq \sup A$. Using transitivity, we conclude that $\inf A \leq \sup A$. □

For $c \in \mathbb{R}$ we define the *dilate* of a set A by $c\,A := \{b \in \mathbb{R} : b = ca \quad \text{for}\, a \in A\}$ and we define the *sum* $A + B$ of two sets A, B by $A + B = \{c : c = a + b \,\text{for some}\, a \in A, b \in B\}$.

Lemma 1.18 (Properties)**.** Suppose A, B are bounded subsets of $\mathbb{R}$.

(1) For $c > 0$ we have $\sup c\,A = c \sup A$ and $\inf c\,A = c \inf A$.
(2) For $c < 0$ we have $\sup c\,A = c \inf A$ and $\inf c\,A = c \sup A$.
(3) Suppose A is contained in B. If $\sup A$ and $\sup B$ exist, then $\sup A \leq \sup B$. In words, making a set larger, increases its supremum.
(4) Suppose A is contained in B. If $\inf A$ and $\inf B$ exist, then $\inf A \geq \inf B$. In words, making a set smaller increases its infimum.
(5) Suppose $x \leq y$ for all $x \in A$ and $y \in B$. Then $\sup A \leq \inf B$.
(6) If A and B are non-empty subsets of $\mathbb{R}$, then $\sup(A + B) = \sup A + \sup B$ and $\inf(A + B) = \inf A + \inf B$

PROOF. (1) We prove that $\sup cA = c \sup A$ for positive c. Suppose $c > 0$, and let $\sup A = M$. Then $cx \leq cM$ for all $x \in A$. Accordingly, cM is an upper bound for cA.

Let us now see that cM is the *least* upper bound for cA. From the definition of $\sup A$ it follows that for every $\varepsilon > 0$ there exists an element $a \in A$ such that $a \geq M - \varepsilon/c$. Thus, we get $ca \geq cM - \varepsilon$ for every $\varepsilon > 0$, and $ca \in cA$. This shows that

$$\sup c\,A = c\,M = c \sup A.$$

(2) Without loss of generality we set $c = -1$. We will show that $\sup cA = \sup(-A) = c \inf A = -\inf A$.

For any $a \in A$, $\inf A \leq a$, so $-\inf A \geq -a$, showing that $-\inf A$ is an upper bound for $-A$. Therefore, $\boxed{-\inf A \geq \sup(-A)}$.

For any $a \in A$ we have $-a \in -A$, so $-a \leq \sup(-A)$, which implies $a \geq -\sup(-A)$. Therefore, $-\sup(-A)$ is a lower bound for A, meaning $-\sup(-A) \leq \inf A$ and thus $\boxed{\sup(-A) \geq -\inf A}$.

The two boxed inequalities prove the identity $\sup(-A) = -\inf A$.

(3) Since $\sup B$ is an upper bound of B, it is also an upper bound of A, i.e. $\sup A \leq \sup B$.

(4) Analogous to (3).

(5) Since $x \leq y$ for all $x \in A$ and $y \in B$, y is an upper bound of A. Hence $\sup A$ is a lower bound of B and we have $\sup A \leq \inf B$.

(6) By definition $A+B$ is bounded above if and only if A and B are bounded above. Hence $\sup(A+B) < \infty$ if and only if $\sup A$ and $\sup B$ are finite. Take $a \in A$ and $b \in B$, then $a + b \leq \sup A + \sup B$. Thus $\sup A + \sup B$ is an upper bound of $A + B$:

$$\sup(A + B) \leq \sup A + \sup B.$$

The reverse direction is a little bit more involved. Let $\varepsilon > 0$. Then there exists $a \in A$ and $b \in B$ such that

$$a > \sup A - \varepsilon/2, \quad b > \sup B - \varepsilon/2.$$

Thus we have $a+b > \sup A+\sup B-\varepsilon$ for every $\varepsilon > 0$, i.e. $\sup(A+B) \geq \sup A + \sup B$.

The remaining statements are assigned as exercises. □

One reason for the relevance of the notions of supremum and infimum is in the formulation of properties of functions.

Definition 1.4.3. Let f be a function with domain X and range $Y \subseteq \mathbb{R}$. Then

$$\sup_X f = \sup\{f(x) : x \in X\}, \quad \inf_X f = \inf\{f(x) : x \in X\}.$$

If $\sup_X f$ is finite, then f is bounded from above on X, and if $\inf_X f$ is finite we call f bounded from below. A function is bounded if both the supremum and infimum are finite.

Lemma 1.19. Suppose that $f, g : X \to \mathbb{R}$ and $f \leq g$, i.e. $f(x) \leq g(x)$ for all $x \in X$. If g is bounded from above, then $\sup_X f \leq \sup_X g$. Assume that f is bounded from below. Then $\inf_X f \leq \inf_X g$.

Proof. Follows from the definitions. □

Lemma 1.20. Suppose f, g are bounded functions from X to $\mathbb{R}$ and c a positive constant. Then

$$\sup_X(f + cg) \leq \sup_X f + c \sup_X g \quad \inf_X(f + cg) \geq \inf_X f + c \inf_X g.$$

The proof is left as an exercise. Try to convince yourself that these inequalities are often strict, as the functions f and g may take values close to their suprema/infima at different points in X.

Finally, recall that a sequence (x_n) of real or complex numbers is an ordered list of numbers x_n, indexed by the natural numbers. In other words, such a sequence (x_n) may be thought of as a function f from $\mathbb{N}$ to $\mathbb{R}$ (or $\mathbb{C}$) with $f(n) = x_n$. Using this function representation of a sequence, **we can define the supremum and infimum of a sequence** (x_n) **using Definition 1.4.3**.

CHAPTER 2

Normed spaces and innerproduct spaces

In order to measure the length of a vector and to define a distance between vectors we introduce the notion of a norm of a vector. Norms may be a tool to specify properties of a class of vectors in a convenient form. We review basic aspects of vector spaces before we define normed vector spaces.

2.1. Vector spaces

Vector spaces formalize the notion of linear combinations of objects that might be vectors in the plane, polynomials, smooth functions, or sequences. Many problems in engineering, mathematics and science are naturally formulated and solved in this setting due to their linear nature. Vector spaces are ubiquitous for several reasons, e.g. as linear approximation of a non-linear object, or as building blocks for more complicated notions, such as vector bundles over topological spaces. Developing an understanding of vector spaces is one of the main objectives of this course. We will restrict our discussion to complex and real vector spaces.

Definition 2.1.1. A *vector space* over a field $\mathbb{F}$ (normally $\mathbb{R}$ or $\mathbb{C}$) is a set V endowed with an operation called *addition*,

$$V \times V \to V, \quad (u, v) \to u + v,$$

and an operation called *scalar multiplication* $\mathbb{F} \times V \to V$,

$$\mathbb{F} \times V \to V, \quad (\lambda, v) \to \lambda v,$$

where these operations satisfy the following properties:

(1) Commutativity: $u + v = v + u$ for all $u, v \in V$ and $(\lambda\mu)v = \lambda(\mu v)$ for all $\lambda, \mu \in \mathbb{F}$;
(2) Associativity: $(u + v) + w = u + (v + w)$ for all $u, v, w \in V$;
(3) Additive identity: There exists an element $0 \in V$ such that $0 + v = v$ for all $v \in V$;
(4) Additive inverse: For every $v \in V$, there exists an element $w \in V$ such that $v + w = 0$;
(5) Multiplicative identity: $1v = v$ for all $v \in V$;
(6) Distributivity: $\lambda(u + v) = \lambda u + \lambda v$ and $(\lambda + \mu)u = \lambda u + \mu u$ for all $u, v \in V$ and $\lambda, \mu \in \mathbb{F}$.

If $\mathbb{F} = \mathbb{R}$, we say that V is a real vector space; if $\mathbb{F} = \mathbb{C}$, we say that V is complex.

The elements of a vector space are called vectors. Given $v_1, ..., v_n \in V$ and $\lambda_1, ..., \lambda_n \in \mathbb{F}$ we call the vector

$$v = \lambda_1 v_1 + \cdots + \lambda_n v_n$$

a *linear combination*.

Examples 2.1.2. We define some useful vector spaces.

- **Spaces of n-tuples:** The set of tuples $(x_1, ..., x_n)$ of real and complex numbers are vector spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ with respect to component-wise addition and scalar multiplication: $(x_1, ..., x_n) + (y_1, ..., y_n) = (x_1 + y_1, ..., x_n + y_n)$ and $\lambda(x_1, ..., x_n) = (\lambda x_1, ..., \lambda x_n)$.
- The set of functions $\mathcal{F}(X, Y)$ of a set X to a set Y: $\lambda f(x) + \mu g(x) := (\lambda f + \mu g)(x)$ for all $x \in X$.
- The space of polynomials of degree at most n, denoted by $\mathcal{P}_n$, where we define the operations of multiplication and addition coefficient-wise: For $p(x) = a_0 + a_1 x + \cdots a_n x^n$ and $q(x) = b_0 + b_1 x + \cdots b_n x^n$ we define

$$\begin{aligned}(p+q)(x) &= (a_0 + b_0) + (a_1 + b_1)x + \cdots (a_n + b_n)x^n \quad \text{and} \\ &= (\lambda p)(x) = \lambda a_0 + \lambda a_1 x + \cdots \lambda a_n x^n, \quad \lambda \in \mathbb{F}.\end{aligned}$$

  The space of all polynomials $\mathcal{P}$ is the vector space of polynomials of arbitrary degree.
- **Sequence spaces**: s denotes the set of sequences, c the set of all convergent sequences, c_0 the set of all sequences converging to 0, c_f the set of all sequences with finitely many non-zero elements.
- **Function spaces:** The set of continuous functions $C(I)$ on an interval of $\mathbb{R}$. Popular choices for I are $[0, 1]$ and $\mathbb{R}$. We define addition and scalar multiplication as follows: For $f, g \in C(I)$ and $\lambda \in \mathbb{F}$

$$(f + g)(x) = f(x) + g(x) \quad \text{and} \, (\lambda f)(x) = \lambda f(x).$$

  We denote by $C^{(n)}(I)$ the space of n-times continuously differentiable functions on I and by $C^\infty(I)$ the space of functions on I with infinitely many continuous derivatives. More generally, the set $\mathcal{F}(X)$ of functions from a set X to $\mathbb{F}$ is a vector space with the operations defined above. Note that $\mathcal{F}(\{1, 2, ..., n\})$ is just $\mathbb{F}^n$ and hence the first class of examples.
- **Spaces of matrices**: Denote by $\mathcal{M}_{m\times n}(\mathbb{C})$ the space of complex $m \times n$ matrices where we define addition and scalar multiplication entry-wise: For $A = (a_{ij})_{i,j}$ and $B = (a_{ij})_{i,j}$ where $i = 1, ..., m$ and $j = 1, ...n$ we define

$$A + B := (a_{ij} + b_{ij})_{i,j} \qquad \text{and} \qquad \alpha(a_{ij})_{ij} = (\alpha a_{ij})_{ij}, \ \alpha \in \mathbb{F}.$$

There are relations between the vector spaces in the aforementioned list. We start with clarifying their inclusion properties.

Definition 2.1.3. A subset W of a vector space V is called a *subspace* if W is a vector space with respect to addition and scalar multiplication of V, denoted by $W \subseteq V$ and if W is a proper subspace by $W \subset V$.

One way to express this more concretely is stated in the next lemma:

Lemma 2.1. A subset W of a vector space V is a subspace if and only if W is closed under linear combinations: For any $\alpha, \beta \in \mathbb{F}$ and $w_1, w_2 \in W$ we have $\alpha_1 w_1 + \alpha_2 w_2 \in W$. Equivalently, we have that the subset W of a vector space V is a subspace if and only if

(1) $0 \in W$;
(2) $w_1 + w_2 \in W$ for any $w_1, w_2 \in W$;
(3) $\alpha w \in W$ for any $\alpha \in \mathbb{F}$ and any $w \in W$.

Some examples of vector subspaces are:

$$\mathcal{P}_n \subset \mathcal{P} \subset \mathcal{F}, \qquad C^\infty(I) \subset C^{(n)}(I) \subset C(I), \qquad c_f \subset c_0 \subset c \subset s.$$

Definition 2.1.4. A *linear transformation* $T : V \to W$ between the vector spaces V and W is a mapping T that preserves the linear structure of a vector space:

$$T(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 T(v_1) + \alpha_2 T(v_2) \qquad \text{for any } v_1, v_2 \in V, \alpha_1, \alpha_2 \in \mathbb{F}.$$

We denote by $\mathcal{L}(V, W)$ the set of all linear transformations between V and W. This is a subspace of the vector space of all functions $f : V \to W$,

$$\mathcal{L}(V, W) \subseteq \mathcal{F}(V, W).$$

Example 2.1.5. Let D denote the differentiation operator $Df = f'$. Then $D : C^{(1)}(a, b) \to C(a, b)$ is a linear transformation.

Linear transformations have some useful properties.

Lemma 2.2. For any $T \in \mathcal{L}(V, W)$ we have $T(0) = 0$.

Proof. We have that $v + 0 = v$ for any $v \in V$; in particular, for $v = 0$ we get

$$T(0) = T(0 + 0) = T(0) + T(0).$$

Subtracting $T(0)$ from both sides in the equation, we get $T(0) = 0$. □

The *kernel* of $T \in \mathcal{L}(V, W)$ is the set

$$\ker(T) := \{v \in V \,|\, Tv = 0\},$$

i.e. $\ker(T) = T^{-1}(0)$.

Lemma 2.3. For a linear transformation $T : V \to W$ the kernel of T is a subspace of V.

Proof. Suppose $v_1, v_2 \in \ker(T)$. Then for any scalars α_1, α_2 we have

$$T(\alpha v_1 + \alpha_2 v_2) = \alpha_1 T(v_1) + \alpha_2 T(v_2) = \alpha_1 \cdot 0 + \alpha_2 \cdot 0 = 0$$

and thus $\alpha v_1 + \alpha_2 v_2 \in \ker(T)$. □

Lemma 2.4. The range of a linear transformation $T : V \to W$ is a subspace of W.

PROOF. Exercise, see problem set 2. □

Definition 2.1.6. Let V and W be subspaces of Z.

(1) The **sum** of V and W is defined by $V+W := \{z \in Z | \, z = v+w, \; v \in V, w \in W\}$.
(2) The **intersection** of V and W is defined by $V \cap W := \{z \in Z | \, z \in V \text{ and } z \in W\}$.

From the definition we see that $V+W$ and $V \cap W$ are subspaces of Z. We introduce further notions: If the sum of the subspaces V and W equals Z, then we say that Z is the sum of V and W and write $V + W = Z$. Moreover, if the subspaces just have the zero vector in common, $V \cap W = \{0\}$, then we refer to $V + W$ as the **direct sum** of V and W.

Lemma 2.5. Let I be an index set. For any collection of vector spaces $\{V_i\}_{i \in I}$, the intersection $\cap_{i \in I} V_i$ is a vector space.

PROOF. Exercise. □

Definition 2.1.7. Let S be a nonempty subset of a vector space V. Then we define the **span** of S, $\text{span}(S)$, as the intersection of all subspaces of V that contain S.

Lemma 2.6. Let $S \subset V$ be a nonempty subset. Then

$$\text{span}(S) = \{\lambda_1 v_1 + \ldots + \lambda_n v_n \; : \; v_1, \ldots, v_n \in S \text{ and } \lambda_1, \ldots, \lambda_n \in \mathbb{F}\}.$$

PROOF. By definition, $\text{span}(S)$ is the intersection of all subspaces W of V that contain the set S. From the preceding lemma, it follows that $\text{span}(S)$ is a subspace of V, hence it is the *smallest* subspace of V that contains S.

Let us denote

$$W := \{\lambda_1 v_1 + \ldots + \lambda_n v_n \; : \; v_1, \ldots, v_n \in S \text{ and } \lambda_1, \ldots, \lambda_n \in \mathbb{F}\},$$

so W is the set of all linear combinations with elements in S.

Being a subspace of V, $\text{span}(S)$ must contain all such linear combinations, so we must have that

$$W \subset \text{span}(S).$$

All we have left to show is that W is a subspace of V. This is not hard to see, since linear combinations of linear combinations are linear combinations as well.

Indeed, let $a, b \in \mathbb{F}$ and let $w_1, w_2 \in W$, so

$$\begin{aligned} w_1 &= \lambda_1 v_1 + \ldots + \lambda_n v_n \quad \text{with } v_1, \ldots, v_n \in S, \\ w_2 &= \mu_1 u_1 + \ldots + \mu_m u_m \quad \text{with } u_1, \ldots, u_m \in S. \end{aligned}$$

Then

$$a\,w_1 + b\,w_2 = a\lambda_1\,v_1 + \ldots + a\lambda_n\,v_n + b\mu_1\,u_1 + \ldots + b\mu_m\,u_m,$$

and since $v_1, \ldots, v_n, u_1, \ldots, u_m \in S$, it follows that $a\,w_1 + b\,w_2 \in W$.

Therefore, W is a subspace of V that contains S, so we must have

$$\operatorname{span}(S) \subset W.$$

Together with the previous inclusion, this proves the equality of the two sets. □

2.2. Normed spaces and metric spaces

The *norm* on a general vector space generalizes the notion of length of a vector in $\mathbb{R}^2$ and $\mathbb{R}^3$.

Definition 2.2.1. Let $\mathbb{F}$ be either $\mathbb{R}$ or $\mathbb{C}$. A *normed space* is a vector space X over $\mathbb{F}$ endowed with a function $\|.\| : X \to \mathbb{R}$, the *norm* on X, such that for all $x, y \in X$:

(1) *Positivity:* $0 \leq \|x\| < \infty$ and $\|x\| = 0$ if and only if $x = 0$;
(2) *Homogeneity:* $\|\alpha x\| = |\alpha|\|x\|$ for $\alpha \in \mathbb{F}$;
(3) *Triangle inequality:* $\|x + y\| \leq \|x\| + \|y\|$.

We denote this normed space by $(X, \|.\|)$

Example 2.2.2. The vector space $\mathbb{R}^n$ with usual addition and scalar multiplication is a normed space when endowed with the summation norm

$$\|(x_1, \ldots, x_n)\|_1 = |x_1| + \cdots + |x_n|.$$

This is a special case of the p-norm, which will be introduced below.

A norm provides us with a way to measure the distance between two vectors in X. We say that the distance between $x \in X$ and $y \in X$ is given by $d(x, y) := \|x - y\|$. This is an example of a *metric* on the vector space X.

Definition 2.2.3. Let X be a set. A *metric* $d : X \times X \to [0, \infty)$ is a function such that for every $x, y, z \in X$

(i) $d(x, y) = 0$ if and only if $x = y$ (positivity);
(ii) $d(x, y) = d(y, x)$ (symmetry);
(iii) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality).

The pair (X, d) is called a *metric space*.

Example 2.2.4.

i) Any set X becomes a metric space when endowed with the *discrete metric*

$$d(x, y) = \begin{cases} 1, & x \neq y \\ 0, & x = y. \end{cases}$$

ii) The summation norm $\|\cdot\|_1$ gives rise to a metric $d(x, y) := \|x - y\|_1$ on $\mathbb{R}^n$. Similarly, any norm $\|\cdot\|$ on a vector space X induces a metric d on X.

Proposition 2.7. If $\|\cdot\|$ is a norm on X then $d(x,y) := \|x-y\|$ is a metric on X.

PROOF. The properties in Definition 2.2.3 are direct consequences of the axioms for a norm. In particular, (i) follows from property (1) of a norm, (ii) is derived from property (2) of a norm for $\lambda = -1$ and (iii) is deduced from property (3) of a norm. □

Proposition 2.7 shows that any normed space may be viewed as a metric space. Note, however, that metric spaces need not even be vector spaces. For instance, the positive real numbers $\mathbb{R}_+ = (0,\infty)$ with the metric $d(x,y) := |x-y|$ is a metric space, but it is not a vector space, as it contains neither an additive identity nor additive inverses.

In this course, we will mainly be interested in normed spaces and inner product spaces. Nevertheless, we introduce certain topological properties in the more general setting of metric spaces.

Definition 2.2.5. For $r > 0$ and $x \in X$ we define the open ball $B_r(x)$ of radius r centered at x as the set

$$B_r(x) = \{y \in X : d(x,y) < r\}.$$

Balls generalize the concept of an interval in $\mathbb{R}$ to any metric space (X,d).

Definition 2.2.6. For $p \in [1,\infty)$ we define the **p-norm**, denoted by $\|\cdot\|_p$, on $\mathbb{R}^n$ by assigning to $x = (x_1,...,x_n) \in \mathbb{R}^n$ the number $\|x\|_p$:

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

For $p = \infty$ we define the ℓ^∞-norm $\|\cdot\|_\infty$ on $\mathbb{R}^n$ by

$$\|x\|_\infty = \max|x_1|,...,|x_n|.$$

The notation $\|.\|_\infty$ is justified by the fact that it is the limit of the $\|.\|_p$ norms.

Lemma 2.8. For $x \in \mathbb{R}^n$ we have

$$\|x\|_\infty = \lim_{p\to\infty} \|x\|_p.$$

PROOF. Without loss of generality we may assume that $\|x\|_\infty = |x_n|$. For $1 \le p < \infty$ we then have

$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots |x_n|^p)^{1/p} = \|x\|_\infty \left(\left(\frac{|x_1|}{\|x\|_\infty} \right)^p + \left(\frac{|x_2|}{\|x\|_\infty} \right)^p + \cdots + 1 \right)^{1/p}.$$

Finally, since $\frac{|x_i|}{\|x\|_\infty} < 1$ for each $i = 1,\ldots,n-1$, we get $\lim_{p\to\infty}(\frac{|x_i|}{\|x\|_\infty})^p = 0$, and it follows that

$$\lim_{p\to\infty} \|x\|_p = \|x\|_\infty.$$

If the maximum is attained by k components of x, we get that the expression in the bracket behaves like $k^{1/p}$, which for $p \to \infty$ still converges to 1. □

Proposition 2.9. For any $1 \leq p \leq \infty$ the vector space $\mathbb{R}^n$ endowed with the p-norm $\|.\|_p$ is a normed space.

Confirming that the p-norm $\|\cdot\|_p$ satisfies positivity and homogeneity is straightforward. However, showing that it also satisfies the triangle inequality is more involved, and requires the following preliminary results:

For $p \in (1, \infty)$, define its *conjugate* as the number q such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

If $p = 1$, then we define q to be ∞ and vice versa (i.e. for $p = \infty$ we set $q = 1$).

Lemma 2.10 (Young's inequality)**.** For $p \in (1, \infty)$ and q its conjugate we have

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q},$$

for any non-negative real numbers a, b. We have equality if and only if $a^p = b^q$.

Here are a few identities about conjugate exponents:

- $q = p/(p-1)$, since $\frac{1}{p} + \frac{1}{q} = 1$ if and only if $\frac{1}{q} = \frac{p-1}{p}$.
- $(p-1)(q-1) = 1$, since $(p-1)q = p$ is equal to $(p-1)q - (p-1) = 1$.

Proof. Consider the function $f(x) = x^{p-1}$ and integrate this with respect to x from zero to a. Now take the inverse function of f given by $f^{-1}(y) = y^{q-1}$, where we used that $1/(p-1) = q-1$ for conjugate exponents p and q. Let us integrate f^{-1} from zero to b. Then the sum of these two integrals always exceeds the product ab. Note that the two integrals are given by a^p/p and b^q/q. Hence we have established Young's inequality.
Equality occurs when $f(a) = f^{-1}(b)$, i.e. if $a^p = b^q$. □

A consequence of Young's inequality is Hölder's inequality.

Lemma 2.11 (Hölder's inequality)**.** Suppose $p \in (1, \infty)$ and q its conjugate, and $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ are vectors in $\mathbb{R}^n$. Then

$$\left|\sum_{i=1}^n x_i y_i\right| \leq \sum_{i=1}^n |x_i|\,|y_i| \leq \Big(\sum_{i=1}^n |x_i|^p\Big)^{1/p} \Big(\sum_{i=1}^n |y_i|^q\Big)^{1/q}.$$

Proof. Set $a_i = |x_i|/(\sum_{i=1}^n |x_i|^p)^{1/p}$ and $b_i = |y_i|/(\sum_{i=1}^n |y_i|^q)^{1/q}$. Then we have $\sum_i a_i^p = 1$ and $\sum_i b_i^q = 1$. By Young's inequality, we get

$$\sum_{i=1}^n |x_i||y_i| \leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |y_i|^q\right)^{1/q}.$$

□

Proof of Proposition 2.9. Positivity and homogeneity of $\|\cdot\|_p$ are consequences of the corresponding properties for the absolute value of a real number.

The triangle inequality is non-trivial, and we split the proof into three cases: $p = 1$, $p = \infty$ and $p \in (1, \infty)$. Let $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ be points in $\mathbb{R}^n$.

(1) For $p = 1$ we have

$$\|x + y\|_1 = |x_1 + y_1| + \cdots + |x_n + y_n| \leq |x_1| + |y_1| + \cdots + |x_n| + |y_n| = \|x\|_1 + \|y\|_1$$

.

(2) For $p = \infty$ the argument is similar:

$$\begin{aligned} \|x + y\|_\infty &= \max\{|x_1 + y_1|, ..., |x_n + y_n|\} \\ &\leq \max\{|x_1| + |y_1|, ..., |x_n| + |y_n|\} \\ &\leq \max\{|x_1|, ..., |x_n|\} + \max\{|y_1|, ..., |y_n|\} = \|x\|_\infty + \|y\|_\infty. \end{aligned}$$

(3) The general case $p \in (1, \infty)$: The triangle inequality follows from Hölder's inequality.

$$\begin{aligned} \|x + y\|_p^p &= \sum_{i=1}^n |x_i + y_i|^p \\ &\leq \sum_{i=1}^n |x_i + y_i|^{p-1}(|x_i| + |y_i|) \\ &= \sum_{i=1}^n |x_i + y_i|^{p-1}|x_i| + \sum_{i=1}^n |x_i + y_i|^{p-1}|y_i| \\ &\leq \Big(\sum_{i=1}^n |x_i + y_i|^p\Big)^{1/q}\Big(\Big(\sum_{i=1}^n |x_i|^p\Big)^{1/p} + \Big(\sum_{i=1}^n |y_i|^p\Big)^{1/p}\Big) \\ &= \|x + y\|_p^{p/q}(\|x\|_p + \|y\|_p) \end{aligned}$$

Dividing by $\|x + y\|_p^{p/q}$ and using $1 - 1/q = 1/p$ we obtain the triangle inequality:

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p.$$

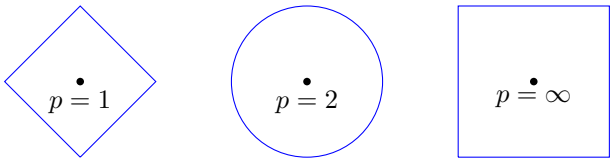
Thus the space $\mathbb{R}^n$ with the p-norm $\|.\|_p$ is a normed space for $p \in [1, \infty]$.

□

The triangle inequality for p-norms on $\mathbb{R}^n$ is also known as **Minkowski's inequality**:

$$\Big(\sum_{i=1}^n |x_i + y_i|^p\Big)^{1/p} \leq \Big(\sum_{i=1}^n |x_i|^p\Big)^{1/p} + \Big(\sum_{i=1}^n |y_i|^p\Big)^{1/p}.$$

Finally, observe that the unit balls of the vector spaces $(\mathbb{R}^2, \|.\|_1)$, $(\mathbb{R}^2, \|.\|_2)$ and $(\mathbb{R}^2, \|.\|_\infty)$ illustrate how the choice of norm affects the nature of the resulting normed space:



The unit sphere $\|x\|_p = 1$ for different values of p.

We may replace the vector space $\mathbb{R}^n$ by $\mathbb{C}^n$ in $(\mathbb{R}^n, \|\cdot\|_p)$, and still obtain a normed space.

Proposition 2.12. Let $\mathbb{C}^n$ be the vector space of complex n-tuples $z = (z_1, ..., z_n)^T$, $z_i \in \mathbb{C}$ for $i = 1, ..., n$. For $1 \leq p < \infty$ we define

$$\|z\|_p = \Big(\sum_{i=1}^{n} |z_i|^p\Big)^{1/p}, \ z \in \mathbb{C}^n$$

and for $p = \infty$ we have $\|z\|_\infty := \max|z_i| : i = 1, ..., n.$ where $z_i \in \mathbb{C}$ and $|z_i| = (z_i\overline{z_i})^{1/2}$ denotes the modulus of z_i. Then $(\mathbb{C}^n, \|.\|_p)$ is a normed space for $1 \leq p \leq \infty$.

The proof for the case $\mathbb{R}^n$ extends to $\mathbb{C}^n$ without significant changes. Similarly, we can define a p-type norm on the vector space of $m \times n$ matrices $\mathcal{M}_{m\times n}(\mathbb{F})$ by using the p-norm on the space $\mathbb{F}^{nm}$: For $1 \leq p < \infty$ we define $\|A\|_{(p)} = (\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^p)^{1/p}$ or $\|A\|_{(\infty)} = \max|a_{ij}|$ for $A \in \mathcal{M}_{m\times n}(\mathbb{F})$. The case $p = 2$ is of special interest and is known as the *Frobenius norm.*

Proposition 2.13. For $1 \leq p \leq \infty$ we have that $(\mathcal{M}_{m\times n}(\mathbb{F}), \|.\|_{(p)})$ is a normed space.

We move on to spaces of sequences.

Definition 2.2.7. For $p \in [1, \infty)$ and a sequence $x = (x_1, x_2, ...)$ of real or complex numbers, we define the p-norm on x by

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

For $p = \infty$, we define the ℓ^∞-norm by

$$\|x\|_\infty := \sup_{i\in\mathbb{N}} |x_i|.$$

We denote by $\ell^p = \ell^p(\mathbb{R})$ (resp. $\ell^p(\mathbb{C})$) the set of all real (resp. complex) sequences with bounded p-norm.

Proposition 2.14. The set ℓ^p is a normed vector space for every $1 \leq p \leq \infty$.

In the proof of Proposition 2.14 we rely again on Hölder's inequality.

Lemma 2.15 (Hölder's inequality)**.** For $1 \leq p \leq \infty$ and q its conjugate we have for $x \in \ell^p$ and $y \in \ell^q$

$$\sum_{i=1}^{\infty} |x_i||y_i| \leq \Big(\sum_{i=1}^{\infty} |x_i|^p\Big)^{1/p} \Big(\sum_{i=1}^{\infty} |y_i|^q\Big)^{1/q}.$$

PROOF. Since Hölder's inequality (Lemma 2.11) is true for all $n \in \mathbb{N}$ we deduce that the limits of the partial sums in question also satisfy these inequalities. Hence we deduce the desired inequality for sequences instead of n-tuples. □

PROOF OF PROPOSITION 2.14. We show first that ℓ^p is a vector space for $p \in [1, \infty)$: For $\alpha \in \mathbb{F}$ and $x \in \ell^p$ it is clear that $\alpha x \in \ell^p$. Confirming that $x + y \in \ell^p$ when $x, y \in \ell^p$ requires an argument:

$$\begin{aligned}
\|x + y\|_p^p &= \sum_{i=1}^{\infty} |x_i + y_i|^p \\
&\leq \sum_{i=1}^{\infty} (|x_i| + |y_i|)^p \\
&\leq 2^p \sum_{i=1}^{\infty} \max\{|x_i|, |y_i|\}^p \\
&= 2^p \sum_{i=1}^{\infty} |\max\{|x_i|^p, |y_i|^p\}| \\
&\leq 2^p (\sum_{i=1}^{\infty} |x_i|^p + \sum_{i=1}^{\infty} |y_i|^p) = 2^p (\|x\|_p^p + \|y\|_p^p) < \infty.
\end{aligned}$$

The norm properties may be deduced as in the case of $\mathbb{R}^n$, as we have Hölder's inequality at our disposal.

The case $p = \infty$ is easier and left as an exercise. □

For $1 \leq p < \infty$ the spaces $(\ell^p, \|.\|_p)$ are subspaces of the vector space of sequences converging to zero, c_0. In contrast $(\ell^\infty, \|.\|_\infty)$ is the space of bounded sequences and is much larger than the other ℓ^p-spaces. We have the following inclusions:

Lemma 2.16. For $p_1 < p_2$ the space ℓ^{p_1} is a proper subspace of ℓ^{p_2}, i.e.

$$\ell^{p_1} \subset \ell^{p_2} \subset \ell^\infty.$$

PROOF. Exercise. □

For example $(1/n)_n$ is in ℓ^p for $p \geq 2$, but not in ℓ^1.

We complete this section by discussing normed spaces of continuous functions.

Definition 2.2.8. For $f \in C[a,b]$ we define its p-norm for $1 \leq p < \infty$ by

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx \right)^{1/p}$$

and $\|f\|_\infty = \sup_{x \in [a,b]} |f(x)|$. We denote by $(C[a,b], \|.\|_p)$ the set of all functions satisfying $\|f\|_p < \infty$.

Proposition 2.17. The space $(C[a,b], \|.\|_p)$ is a normed space for $p \in [1, \infty]$.

We do not include the proof of this proposition, but strongly encourage the reader to go through it. The following version of Hölder's inequality will be useful in confirming that $\|\cdot\|_p$ is indeed a norm on $C[a,b]$.

Lemma 2.18 (Hölder's inequality)**.** For $1 \leq p \leq \infty$ and its conjugate q we have

$$\int_a^b |f(x)||g(x)|dx \leq \|f\|_p \|g\|_q.$$

Proof. We assume without loss of generality that $\|f\|_p = 1 = \|g\|_q$. By Young's inequality we have

$$|f(x)||g(x)| \leq |f(x)|^p/p + |g(x)|^q/q$$

and thus

$$\int_a^b |f(x)||g(x)|\, dx \leq \frac{1}{p} \int_a^b |f(x)|^p\, dx + \frac{1}{q} \int_a^b |g(x)|^q\, dx = 1 = \|f\|_p \|g\|_q.$$

□

One considers as well $(C(\mathbb{R}), \|.\|_p)$ and in this case $\|.\|_p < \infty$ behaves differently as for bounded intervals. Namely, we have for $f \in C(\mathbb{R})$ that $\|.\|_p < \infty$ implies for $p = \infty$ that f is bounded, and for $1 \leq p < \infty$ one has in some sense that f has to converge to zero at a certain rate.

2.3. Inner product spaces

In this section we introduce inner product spaces. For vectors $z, z' \in \mathbb{C}^n$, we are familiar with the 'dot product'

$$z \cdot z' = z_1 \overline{z_1'} + \cdots + z_n \overline{z_n'},$$

where, if we take the dot product of z with itself, we get the length of the vector z squared:

$$z \cdot z = |z_1|^2 + \cdots + |z_n|^2.$$

For real vectors $x \in \mathbb{R}^n$, this can be expressed without conjugates as

$$x \cdot x' = x_1 x_1' + \cdots + x_n x_n',$$

and

$$x \cdot x = |x_1|^2 + \cdots + |x_n|^2.$$

It is this concept that we now extend to general vector spaces by introducing the notion of an inner product.

Definition 2.3.1. Let X be a vector space over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. An *inner product* $\langle \cdot, \cdot \rangle$ on X is a map $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{F}$ which is

(1) conjugate symmetric,
$$\langle x, y \rangle = \overline{\langle y, x \rangle},$$
(2) linear in its first argument,
$$\langle \alpha x, y \rangle = \alpha \langle x, y \rangle ,$$
$$\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle ,$$
(3) and positive definite,
$$\langle x, x \rangle \geq 0 \quad \text{with equality only if} \quad x = 0,$$
for any $x, y, z \in X$ and $\alpha \in \mathbb{F}$. The pair $(X, \langle \cdot, \cdot \rangle)$ is called an *inner product space.*

Example 2.3.2. The familiar dot product on $\mathbb{R}^n$ defines an inner product on this vector space:
$$\langle x, y \rangle := x \cdot y = \sum_{j=1}^{n} x_j y_j,$$
and the familiar dot product on $\mathbb{C}^n$ defines an inner product on $\mathbb{C}^n$:
$$\langle z, z' \rangle := z \cdot z' = \sum_{j=1}^{n} z_j \overline{z'_j}.$$

Proposition 2.19 (Properties of the inner product)**.** An inner product $\langle \cdot, \cdot \rangle$ on X satisfies

i) $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle ,$
ii) $\langle x, \alpha y \rangle = \overline{\alpha} \langle x, y \rangle ,$
iii) $\langle x, 0 \rangle = \langle 0, x \rangle = 0,$
iv) If $\langle x, z \rangle = 0$ for all $z \in X$, then $x = 0$.

Proof. Proof of *i)*: We have that
$$\langle x, y + z \rangle = \overline{\langle y + z, x \rangle} = \overline{\langle y, x \rangle + \langle z, x \rangle} = \overline{\langle y, x \rangle} + \overline{\langle z, x \rangle} = \langle x, y \rangle + \langle x, z \rangle .$$
The proofs of the remaining statements are left as an exercise. □

An inner product space $(X, \langle \cdot, \cdot \rangle)$ carries a natural norm given by
$$\|x\| = \langle x, x \rangle^{1/2} .$$
To prove that this is indeed a norm on X we need the following inequality.

Proposition 2.20 (Cauchy-Schwarz inequality)**.** For all $x, y \in (X, \langle\cdot,\cdot\rangle)$, we have

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

If both x and y are non-zero, we have equality if and only if $x = \alpha y$ for some $\alpha \in \mathbb{F}$.

PROOF. Assume first that $x = \alpha y$ for some $\alpha \in \mathbb{F}$. Then

$$\begin{aligned}|\langle x, y\rangle| &= |\langle \alpha y, y\rangle| = |\alpha|\,|\langle y, y\rangle| \\ &= |\alpha|\|y\|^2 = |\alpha|\|y\|\,\|y\| = \|x\|\|y\|.\end{aligned}$$

Now assume that there is no $\alpha \in \mathbb{F}$ for which $x = \alpha y$. Then $x - \alpha y \neq 0$ for all $\alpha \in \mathbb{F}$, and thus

$$\begin{aligned}0 &< \langle x - \alpha y, x - \alpha y\rangle \\ &= \langle x, x\rangle - \overline{\alpha}\langle x, y\rangle - \alpha\langle y, x\rangle + |\alpha|^2\langle y, y\rangle \\ &= \|x\|^2 - \overline{\alpha}\langle x, y\rangle - \alpha\overline{\langle x, y\rangle} + |\alpha|^2\|y\|^2 \\ &= \|x\|^2 - 2\mathrm{Re}(\overline{\alpha}\langle x, y\rangle) + |\alpha|^2\|y\|^2.\end{aligned}$$

The Cauchy-Schwarz inequality is trivially true if $\|y\| = 0$, so we may assume $\|y\| \neq 0$. We can then choose to insert $\alpha = \langle x, y\rangle / \|y\|^2$ in the inequality above to obtain

$$\begin{aligned}0 &< \|x\|^2 - \frac{2}{\|y\|^2}|\langle x, y\rangle|^2 + \frac{|\langle x, y\rangle|^2}{(\|y\|^2)^2}\cdot\|y\|^2 \\ &= \|x\|^2 - \frac{1}{\|y\|^2}|\langle x, y\rangle|^2.\end{aligned}$$

Finally multiplying by $\|y\|^2 > 0$ on both sides in this inequality, we get

$$0 < \|x\|^2\|y\|^2 - |\langle x, y\rangle|^2.$$

The Cauchy-Schwarz inequality follows. □

Proposition 2.21. If $(X; \langle\cdot,\cdot\rangle)$ is an inner product space, then $\|x\| = \langle x, x\rangle^{1/2}$ defines a norm on X.

PROOF. We see immediately that

$$\|x\| = 0 \quad\Leftrightarrow\quad \|x\|^2 = 0 \quad\Leftrightarrow\quad \langle x, x\rangle = 0 \quad\Leftrightarrow\quad x = 0,$$

and the positive homogeneity is also straightforward:

$$\|\alpha x\| = \langle \alpha x, \alpha x\rangle^{1/2} = (\alpha\overline{\alpha}\langle x, x\rangle)^{1/2} = \left(|\alpha|^2\|x\|^2\right)^{1/2} = |\alpha|\|x\|.$$

Finally, we have

$$\begin{aligned}\|x + y\|^2 &= \|x\|^2 + 2\mathrm{Re}(\langle x, y\rangle) + \|y\|^2 \\ &\leq \|x\|^2 + 2|\langle x, y\rangle| + \|y\|^2 \\ &\leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 \\ &= (\|x\| + \|y\|)^2,\end{aligned}$$

where we have used the Cauchy-Schwarz inequality in the third line. This shows that the triangle inequality holds for $\|x\| = \langle x, x\rangle^{1/2}$. □

Example 2.3.3. (1) The sequence space ℓ^2 of square-summable (real- or) complex-valued sequences $(z_i), (z_i')$ with the inner product

$$\langle z, z'\rangle = \sum_{i=1}^{\infty} z_i \overline{z_i'},$$

is an inner product space. Moreover, the $\|\cdot\|_2$-norm on ℓ^2 is induced by this inner product, so the Cauchy-Schwarz inequality yields

$$|\langle z, z'\rangle| \leq \|z\|_2 \|z'\|_2.$$

The sequence space ℓ^2 was the first example of an inner product space, studied by D. Hilbert in 1901 in his work on Fredholm operators.

(2) The vector space $C[a, b]$ of continuous (real- or) complex-valued functions on an interval $[a, b]$ with the inner product

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}dx$$

is an inner product space. Moreover, the $\|\cdot\|_2$-norm on $C[a, b]$ is induced by this inner product, so the Cauchy-Schwarz inequality yields

$$|\langle f, g\rangle| \leq \|f\|_2 \|g\|_2.$$

We have seen that any inner product $\langle\cdot,\cdot\rangle$ carries a norm $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$. Let us now turn this around: How can we determine if a given norm is induced by an underlying inner product? Jordan and von Neumann gave the following characterization of these norms.

Theorem 2.22 (Jordan-von Neumann)**.** Let $(X, \|\cdot\|)$ be a normed space. If the norm $\|\cdot\|$ satisfies the *parallelogram law*

$$\|x - y\|^2 + \|x + y\|^2 = 2\|x\|^2 + 2\|y\|^2 \quad \text{for all } x, y \in X,$$

then there exists an inner product $\langle\cdot,\cdot\rangle$ on X such that $\langle\cdot,\cdot\rangle^{1/2} = \|\cdot\|$. If X is a complex vector space, then the inner product is given by

$$\langle x, y\rangle = \frac{1}{4}\sum_{k=1}^{4} i^k \|x + i^k y\|^2.$$

If X is a real vector space, then

$$\langle x, y\rangle = \frac{1}{4}\left(\|x + y\|^2 - \|x - y\|^2\right).$$

The last two equations are both known as the *polarization identity*.

We state Theorem 2.22 without proof in this course. Checking the inner product axioms for the suggested inner product is not difficult, but non-trivial, and involves repeated use of the parallelogram law. Using Theorem 2.22, we can argue that *not* every normed space is necessarily induced by an inner product space.

Example 2.3.4. By showing that the following norms do not satisfy the parallelogram law for all vectors $x, y \in X$, one can verify that:

i) The supremum norm $\|\cdot\|_\infty$ on $C[0,1]$ is not induced by an inner product. Neither is the p-norm on $C[0,1]$ for any $p \neq 2$.
ii) The p-norm on the sequence space ℓ^p is not induced by an inner product if $p \neq 2$.

Finally, we will see that inner products provide us with a generalization of the notion of *orthogonality* of elements.

Definition 2.3.5. Two elements x, y in an inner product space $(X, \langle\cdot,\cdot\rangle)$ are *orthogonal* if $\langle x, y\rangle = 0$

The theorem of Pythagoras is true for any innerproduct space $(X, \langle .,.\rangle)$.

Proposition 2.23 (Pythagoras' Theorem)**.** Let $(X, \langle\cdot,\cdot\rangle)$ be an inner product space. For two orthogonal elements $x, y \in X$ we have

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2.$$

Proof. By assumption we have $\langle x, y\rangle = 0$. It follows that

$$\|x+y\|^2 = \|x\|^2 + 2\mathrm{Re}\,\langle x, y\rangle + \|y\|^2 = \|x\|^2 + \|y\|^2.$$

□

Example 2.3.6. Consider the collection of exponential functions

$$e_m(x) = e^{2\pi i m x}, \quad m \in \mathbb{Z},$$

in the inner product space $(C[0,1], \langle\cdot,\cdot\rangle)$ of continuous, complex-valued functions on the unit interval $[0,1]$. For $m \neq n$, we see that e_m and e_n are orthogonal, as

$$\langle e_m, e_n\rangle = \int_0^1 e^{2\pi i(m-n)x} dx = \frac{1}{2\pi i(m-n)}(e^{2\pi i(m-n)} - 1) = 0.$$

Note that $\langle e_n, e_n\rangle = 1$. We can express this using Kronecker's delta function as

$$\langle e_m, e_n\rangle = \delta_{m,n}.$$

Definition 2.3.7. A set of vectors $\{e_i\}_{i\in I}$ in an inner product space $(X, \langle .,,\rangle)$ is called an *orthogonal family* if $\langle e_i, e_j\rangle = 0$ for all $i \neq j$. If additionally $\|e_i\| = 1$ for all $i \in I$, then we refer to it as an *orthonormal family*.

We see that the set of exponentials $\{e^{2\pi i n x}\}_{n\in\mathbb{Z}}$ in the previous example is an orthonormal family in $(C[0,1], \langle\cdot,\cdot\rangle)$. This particular system lies at the very heart of Fourier analysis, or more generally harmonic analysis.

CHAPTER 3

Banach and Hilbert spaces

With normed spaces and inner product spaces introduced, we focus on *completeness* in this chapter. This requires that we first define what we mean by convergent sequences in these spaces. We will learn that a complete normed space is called a *Banach space*, whereas a complete inner product space is called a *Hilbert space*. Finally, we prove *Banach's fixed point theorem*, and discuss some applications thereof.

3.1. Sequences in metric spaces and normed spaces

Definition 3.1.1. Let (X, d) be a metric space. A sequence $(x_n)_{n\in\mathbb{N}}$ in X is said to **converge to** $x \in X$ if for every $\varepsilon > 0$ one can find $N = N(\varepsilon) \in \mathbb{N}$ such that

$$d(x_n, x) < \varepsilon$$

whenever $n \geq N$. The element x is called the **limit** of the sequence $(x_n)_{n\in\mathbb{N}}$.

In particular, if $(X, \|.\|)$ is a normed space, then $(x_n)_{n\in\mathbb{N}}$ converges to $x \in X$ if for every $\varepsilon > 0$ one can find $N = N(\varepsilon) \in \mathbb{N}$ such that

$$\|x - x_n\| < \varepsilon$$

whenever $n \geq N$.

This notion of convergence for a sequence is a natural generalization of the notion of convergence for sequences of real or complex numbers. Note, however, that the elements of our sequence can now belong to any (metric or) normed space; for example, a sequence in ℓ^2 is a sequence of sequences, i.e. a sequence (x_n) where each element $x_n = (x_{n1}, x_{n2}, \ldots)$ is itself a (square-summable) sequence. A more geometric view on convergence is to observe that for any $\varepsilon > 0$ there must exist an $N = N(\varepsilon)$ such that $(x_N, x_{N+1}, ...)$ lies inside the ball $B_\varepsilon(x)$ of radius ε around the limit point x. Note that $(x_N, x_{N+1}, ...)$ is often called the **tail** of the sequence $(x_n)_{n\in\mathbb{N}}$. Hence convergence of $x_n \to x$ means that for arbitrary small balls around the limit point x, the tail of $(x_n)_{n\in\mathbb{N}}$ is contained in $B_\varepsilon(x)$.

Proposition 3.1 (Properties)**.** Suppose that the sequence $(x_n)_{n\in\mathbb{N}}$ in the normed space $(X, \|\cdot\|)$ is convergent.

i) The limit $x \in X$ of the sequence $(x_n)_{n\in\mathbb{N}}$ is unique.

ii) The norm of x_n converges to the norm of x:

$$\Big|\, \|x_n\| - \|x\| \,\Big| \to 0.$$

PROOF. i) Suppose there exist two limits $x, y \in X$ of $(x_n)_{n\in\mathbb{N}}$. Then for any $\varepsilon > 0$ there exist $N_1, N_2 \in \mathbb{N}$ such that $\|x_n - x\| \leq \varepsilon/2$ for all $n \geq N_1$, and $\|x_n - y\| \leq \varepsilon/2$ for all $n \geq N_2$. Hence for all $n \geq \max\{N_1, N_2\}$, we have

$$\|x - y\| = \|x - x_n + x_n - y\| \leq \|x - x_n\| + \|x_n - y\| \leq \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

It follows that $\|x - y\| = 0$, and thus $x = y$.

ii) By the reversed triangle inequality we have that

$$\Big|\, \|x_n\| - \|x\| \,\Big| \leq \|x_n - x\|,$$

and by assumption $\|x_n - x\| \to 0$. The claim follows.

□

The notion of convergence depends on the metric that the set X is equipped with, or similarly on the norm that the vector space is equipped with. This is illustrated by the following example.

Example 3.1.2. Consider the sequence $(f_n)_{n\in\mathbb{N}}$ in $C[0,1]$ given by

$$f_n(t) = e^{-nt}.$$

It is clear that f_n converges to the zero function in $(C[0,1], \|.\|_1)$, as

$$\|f_n - 0\|_1 = \int_0^1 e^{-nt} dt = \frac{1}{n}(1 - e^{-n}) \to 0$$

as $n \to \infty$. However, the sequence $(f_n)_{n\in\mathbb{N}}$ does not converge to the zero function in $(C[0,1], \|.\|_\infty)$, since

$$\|f_n - 0\|_\infty = \sup_{t\in[0,1]} |e^{-nt}| = 1$$

for any $n \in \mathbb{N}$. In fact, one can show that there is no $g \in C[0,1]$ such that $f_n \to g$ in $(C[0,1], \|.\|_\infty)$.

Let A be a subset of the metric space (X, d).

- A point $x_0 \in A$ is called an *interior point* of A if there is a small ball centered at x_0 which is contained entirely in A, i.e. if there exists an $\varepsilon > 0$ such that $B_\varepsilon(x_0) \subseteq A$.
- A point $x_0 \in X$ is called a *boundary point* of A if any small ball centered at x_0 has nonempty intersections with both A and its complement, i.e. if for all $\varepsilon > 0$ there exist $x, y \in B_\varepsilon(x_0)$ such that $x \in A$ and $y \in X\backslash A$.
- The set of interior points of A constitutes its *interior*, $\mathrm{int}(A)$, and the set of boundary points its *boundary*, ∂A. We say that A is *open* if any point in A is an interior point, and A is *closed* if its boundary ∂A is contained in A. The *closure of* A is the union of A and its boundary, and is denoted $\overline{A} := A \cup \partial A$.

Note: The open ball $B_r(x_0)$ is an open set. Its closure is contained in the closed ball

$$\overline{B}_r(x_0) = \{y \in X : d(x, y) \leq r\},$$

but for a normed space $(X, \|.\|)$ we have that the closure of $B_r(x_0)$ may be identified with the closed ball of radius r centered at x_0.

A useful reformulation of the definition of closed sets is stated in the following lemma:

Lemma 3.2. Let A be a subset of (X, d). Then A is closed if and only if its complement $X\backslash A$ is open.

A few remarks concerning closed and open sets: A set in a metric space does not need to be either open or closed, e.g. $[0, 1)$ in $(\mathbb{R}, |.|)$. Furthermore, a set might be both open and closed, e.g. in (X, d) the set X and the empty set are both open and closed.

Example 3.1.3. The ball $B_r(x)$ in the metric space $(\mathbb{R}, |\cdot|)$ is the open interval $(x-r, x+r)$. The boundary points of $B_r(x)$ are $x-r$ and $x+r$, and accordingly the closure $\overline{B}_r(x)$ of $B_r(x)$ is the closed interval $[x-r, x+r]$. The intervals $[x-r, x+r)$ and $(x-r, x+r]$ are neither open nor closed in the metric space $(\mathbb{R}, |\cdot|)$.

Now suppose A is a subset of X, and let $(a_n)_{n\in\mathbb{N}}$ in A be a convergent sequence such that $a_n \to x$, with $x \in X\backslash A$. That is, the sequence a_n converges to an element $x \in X$ which is *not* contained in A. Then we call x a **limit point** of A. We write $\overline{A}$ for the union of A and all its limit points. This agrees with our definition of the *closure* of a set A . We see that $x \in \overline{A}$ if there exists a sequence $(a_n)_{n\in\mathbb{N}}$ in A such that $a_n \to x$.

Lemma 3.3. Let A be subset of a metric space (X, d). Then the closure of A, $\overline{A}$ consists of all limits of convergent sequences $(x_n)_n \subset A$. Consequently, a set A in (X, d) is closed if and only if the limit of any convergent sequence of elements in A is also in A.

Proof. Assume A is closed in (X, d) and let $(x_n)_n \subset A$ converge to $x \in X$. Suppose x is not in the closed set A. Hence x is an element of the open set $X\backslash A$, i.e. there exists a ball of radius ε around x which does not contain any elements of the sequence $(x_n)_n \subset A$. This contradicts our assumption that $x_n \to x$.
Suppose now that $x \in \overline{A}$. Then $x \in A$ or $x \in \partial A$. Hence we have for any $\varepsilon > 0$ that $B_\varepsilon(x) \cap A \neq \emptyset$. Thus if we pick for $\varepsilon = 1/n$ for $n \in \mathbb{N}$, then we obtain a sequence $(x_n) \subset A$. For any $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ such that $\varepsilon > 1/N$. By the construction of the sequence $(x_n)_n$ we have $|x - x_n| < 1/N < \varepsilon$. □

Example 3.1.4. We revisit Example 3.1.2. Let A be the subset of $C[0, 1]$ containing all positive-valued functions, i.e.

$$A = \{f \in C[0, 1] \,:\, f(t) > 0 \text{ for all } t \in [0, 1]\}\,.$$

We have seen that f_n converges to the zero function in $(C[0, 1], \|\cdot\|_1)$. Thus, although $f_n \in A$ for every $n \in \mathbb{N}$, the sequence $(f_n)_{n\in\mathbb{N}}$ converges to an element (the zero function) which is not contained in A. This shows that the function $f(t) = 0$ for all $t \in [0, 1]$ is a *limit point* of A.

We know that if $(a_n)_{n\in\mathbb{N}}$ is a convergent sequence of real numbers, then $(a_n)_{n\in\mathbb{N}}$ is a bounded sequence, meaning that there exists a constant $M > 0$ such that $|a_n| < M$ for all $n \in \mathbb{N}$. The same is true for convergent sequences in metric spaces and normed spaces, provided that we define bounded subsets appropriately.

Definition 3.1.5. A subset A of a metric space (X, d) is **bounded** if there exists a radius $r_0 > 0$ and a vector $x_0 \in X$ such that $A \subseteq B_{r_0}(x_0)$. In this case we define the **diameter** of A, $\operatorname{diam}(A)$, to be the real number

$$\operatorname{diam}(A) = \sup\{d(x, y) \,:\, x, y \in A\}.$$

In particular, if A is a bounded subset of a normed space $(X, \|\cdot\|)$, then the diameter of A is

$$\operatorname{diam}(A) = \sup\{\|x - y\| \,:\, x, y \in A\}.$$

Lemma 3.4. For a subset A of a normed space $(X, \|\cdot\|)$, the following statements are equivalent:

i) A is bounded.
ii) There exists a constant $M > 0$ such that $\|x-y\| \leq M$ for all $x, y \in A$.
iii) $\operatorname{diam}(A) < \infty$
iv) For every $x \in X$ one can find a radius $r > 0$ such that $A \subseteq B_r(x)$.
v) There exists a $m > 0$ such that $\|x\| \leq m$ for all $x \in A$.

Proof. We show that $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv) \Rightarrow (i)$, and finally that $(i) \Rightarrow (v) \Rightarrow (i)$.

If (i) holds, we have $A \subset B_{r_0}(x_0)$ for some $x_0 \in X$ and $r_0 > 0$. Then for any $x, y \in A$, we have $x, y \in B_{r_0}(x_0)$, and thus

$$\|x - y\| \leq \|x - x_0\| + \|x_0 - y\| \leq 2r_0.$$

Hence, $\|x - y\| \leq M = 2r_0$ for all $x, y \in A$.

If (ii) holds, then $\operatorname{diam}(A) \leq M < \infty$ by the definition of supremum.

If (iii) holds, then for all $x, y \in A$ we have $\|x - y\| \leq \operatorname{diam}(A) < \infty$. Fix an element $a_1 \in A$. Given any $x \in X$ and $a \in A$, we have $\|x - a\| \leq \|x - a_1\| + \|a_1 - a\| \leq d(x, a_1) + \operatorname{diam}(A)$ (where d is the metric induced by $\|\cdot\|$). Now define $r := d(x, a_1) + \operatorname{diam}(A)$. Then $A \subseteq B_r(x)$, so (iv) is satisfied.

The implication $(iv) \to (i)$ is immediate from our definition of boundedness of A.

To see that $(i) \to (v)$, we simply observe that

$$A \subseteq B_{r_0}(x_0) \subseteq B_{\|x_0\|+r_0}(0),$$

and thus $\|x\| \leq m := \|x_0\| + r$ for all $x \in A$.

Finally, if (v) is satisfied, then $A \subseteq B_m(0)$, and again the condition defining boundedness of A is immediately satisfied (with $x_0 = 0$ and $r_0 = m$). □

Lemma 3.5. A convergent sequence in a metric space (X, d) is bounded.

Proof. Exercise. □

The definition of convergence of a sequence has one obvious disadvantage; it involves a specified limit value. In order to prove that a sequence $(x_n)_{n\in\mathbb{N}}$ is indeed convergent, we must first have a candidate x for the limit value. This shortcoming in the definition motivates the introduction of the following notion.

Definition 3.1.6. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in the metric space (X, d). We say that $(x_n)_{n\in\mathbb{N}}$ is a **Cauchy sequence** if for any $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ such that for all $m, n \geq N$ we have

$$d(x_n, x_m) < \varepsilon.$$

In particular, if $(x_n)_{n\in N}$ is a sequence in the normed space $(X, \|\cdot\|)$, then $(x_n)_{n\in\mathbb{N}}$ is Cauchy if for any $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ such that

$$\|x_n - x_m\| < \varepsilon, \quad n, m \geq N.$$

In an inner product space $(X, \langle\cdot,\cdot\rangle)$, we say that a sequence $(x_n)_{n\in N}$ is Cauchy if the sequence is Cauchy with respect to the induced norm $\|x\| := \langle x, x\rangle^{1/2}$.

Later, we will also discuss Cauchy sequences in an inner product space

Lemma 3.6. Any Cauchy sequence in (X, d) is bounded.

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a Cauchy sequence. Then there exists $N \in \mathbb{N}$ such that for all $m, n \geq N$ we have

$$d(x_n, x_m) < 1.$$

In particular, we have

$$d(x_N, x_m) < 1 \quad \forall\, m \geq N,$$

or equivalently $x_m \in B_1(x_N)$ for all $m \geq N$. Now let

$$r = \max\left\{1, d(x_1, x_N), d(x_2, x_N), \ldots, d(x_{N-1}, x_N)\right\}.$$

Then for any $n \in \mathbb{N}$, we have $x_n \in B_{r+1}(x_N)$, so $(x_n)_{n\in\mathbb{N}}$ is bounded. □

Lemma 3.7. Every convergent sequence in $(X, \|.\|)$ (or in (X, d)) is a Cauchy sequence.

Proof. Exercise. □

The reversed implication does not hold; Cauchy sequences need not necessarily be convergent.

Example 3.1.7. The sequence of functions $(f_n)_{n\in N}$ in $(C[a, b], \|\cdot\|_1)$ given by

$$f_n(t) = \begin{cases} 0 & \text{for } a \leq t \leq \frac{a+b}{2}, \\ n(t - \frac{a+b}{2}) & \text{for } \frac{a+b}{2} < t \leq \frac{a+b}{2} + \frac{1}{n}, \\ 1 & \text{for } \frac{a+b}{2} + \frac{1}{n} \leq t \leq b. \end{cases}$$

is a Cauchy sequence. For $m > n$, we have

$$\|f_m - f_n\|_1 = \frac{1}{2}\left(\frac{1}{n} - \frac{1}{m}\right) < \frac{1}{2n} \to 0 \quad \text{as } n \to \infty.$$

(To see this, observe that the norm of the difference is necessarily equal to the area of the red triangle in Figure 1.)

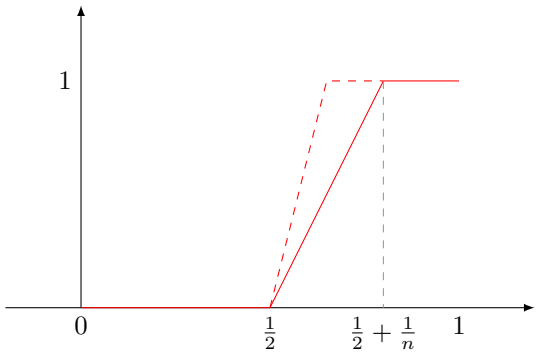


FIGURE 1. The function f_n in Example 3.1.7 when $[a,b]=[0,1]$.

However, the sequence f_n does not converge in $(C[a,b],\|\cdot\|_1)$. Suppose to the contrary that there exists $f\in C[a,b]$ such that $f_n\to f$. Let us analyze the implications of $\|f_n-f\|_1\to 0$ as $n\to\infty$ by splitting the integral in three:

$$\int_a^b |f_n(t)-f(t)|dt=\Big[\int_a^{\frac{a+b}{2}}+\int_{\frac{a+b}{2}}^{\frac{a+b}{2}+\frac{1}{n}}+\int_{\frac{a+b}{2}+\frac{1}{n}}^{b}\Big]|f_n(t)-f(t)|dt$$

In order for $\|f_n-f\|_1$ to tend to zero, each of the integrals on the right hand side must necessarily tend to 0. We have that:

(1) $\int_a^{\frac{a+b}{2}}|f_n(t)-f(t)|dt\to 0$ only if $f=0$ on $[a,\frac{a+b}{2}]$;

(2) Since f_n is continuous for all $n\in\mathbb{N}$ and f is continuous on $[a,b]$ we have

$$\int_{\frac{a+b}{2}}^{\frac{a+b}{2}+\frac{1}{n}}|f_n(t)-f(t)|dt\le(\max_{t\in[a,b}|f(t)|+1)\frac{1}{n}\to 0$$

as $n\to\infty$. Hence, requiring $\int_{\frac{a+b}{2}}^{\frac{a+b}{2}+\frac{1}{n}}|f_n(t)-f(t)|dt\to 0$ imposes no condition on the limit function f.

(3) By the continuity of f we have that

$$\int_{\frac{a+b}{2}+\frac{1}{n}}^{b}|f_n(t)-f(t)|dt=\int_{\frac{a+b}{2}+\frac{1}{n}}^{b}|1-f(t)|dt\to\int_{\frac{a+b}{2}}^{b}|1-f(t)|dt,$$

as $n\to\infty$. Hence, for this limit to be zero, we must have $1-f(t)=0$ for all $t\in((a+b)/2,b]$, or equivalently $f(t)=1$ for $t\in((a+b)/2,b]$. This shows that if f_n has a limit f, it must necessarily satisfy

$$f(t)=\begin{cases}0 & \text{for } a\le t\le\frac{a+b}{2},\\ 1 & \text{for } \frac{a+b}{2}<t\le b.\end{cases}$$

But this is a discontinuous function $f\notin C[a,b]$.

Example 3.1.7 illustrates that the vector space $(C[a,b], \|\cdot\|_1)$ is not *complete*; this is the topic of our next subsection.

3.2. Completeness

We have seen that a convergent sequence $(x_n)_{n\in\mathbb{N}}$ must be Cauchy. To the contrary, a Cauchy sequence $(y_n)_{n\in\mathbb{N}}$ need not necessarily be convergent. However, this is indeed the case in the vector space of real (or complex) numbers endowed with the norm $\|\cdot\| = |\cdot|$.

Theorem 3.8. Suppose $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $(\mathbb{R}, |\cdot|)$. Then $(a_n)_{n\in\mathbb{N}}$ is convergent in $\mathbb{R}$, meaning there exists $a \in \mathbb{R}$ such that $a_n \to a$.

Proof. We are assuming $(a_n)_{n\in\mathbb{N}}$ is Cauchy, so by Lemma 3.6 the sequence $(a_n)_{n\in\mathbb{N}}$ is bounded, and we can find $M \in \mathbb{R}$ such that $a_n \in [-M, M]$ for all n. Now let

$$S := \{s \in [-M, M] : \text{ there exist infinitely many } n \in \mathbb{N} \text{ such that } a_n \geq s\}.$$

It is clear that $-M \in S$ and that S is bounded above by M. We now define $a := \sup S$. Then a is necessarily an element of $\mathbb{R}$.[1]

Claim: $a_n \to a$ as $n \to \infty$.

Observe that for any $\varepsilon > 0$, the Cauchy condition ensures that we can find N_1 such that

$$|a_m - a_n| < \varepsilon/2, \quad m, n > N_1. \tag{3.1}$$

By the definition of supremum as the least upper bound, it is clear that $a+\varepsilon/2 \notin S$, meaning only finitely many elements a_n exceed $a + \varepsilon/2$. In other words, we can find N_2 such that

$$a_n \leq a + \varepsilon/2, \quad n > N_2.$$

On the other hand, since a is the least upper bound of S, the smaller number $a - \varepsilon/2$ cannot be an upper bound of S. Thus, there exists $s \in S$ such that $s \geq a - \varepsilon/2$. Consequently, we have infinitely many elements a_n satisfying

$$a_n > s \geq a - \varepsilon/2,$$

and in particular there exists $N > \max\{N_1, N_2\}$ such that

$$a - \varepsilon/2 < a_N \leq a + \varepsilon/2. \tag{3.2}$$

Finally, combining (3.1) and (3.2) we get

$$|a_n - a| \leq |a_n - a_N| + |a_N - a| < \varepsilon$$

for all $n \geq N$. This completes the proof that $a_n \to a$. □

When we define $a := \sup S$ in the proof above, we are using a *fundamental* property of $\mathbb{R}$, namely that *any non-empty subset $S \subset \mathbb{R}$ that is bounded from above must have a least upper bound (or supremum) in the set of real numbers.* This is known as the **completeness property** or **least upper bound property**

[1] We elaborate on this once the proof is completed.

of $\mathbb{R}$. In contrast, the set $\mathbb{Q}$ of rational numbers does not have the least upper bound property. An example that illustrates this is

$$S := \left\{ x \in \mathbb{Q} \,:\, x^2 < 3 \right\}.$$

This set is non-empty and bounded from above, but it does not have a least upper bound *in* $\mathbb{Q}$; its least upper bound as a subset of the reals would be $\sqrt{3}$, but this is not an element in $\mathbb{Q}$. For any upper bound $y \in \mathbb{Q}$, we can always find another upper bound $x \in \mathbb{Q}$ such that $x < y$.

EXERCISE 3.2.1. *Show that the equation $x^2 = 3$ has no solutions in $\mathbb{Q}$.*

The property of $\mathbb{R}$ that any Cauchy sequence converges is such a desirable property that it has a name when it occurs in general metric and normed spaces.

Definition 3.2.2. A metric space (X, d) is said to be *complete* if every Cauchy sequence $(x_k)_{k\in\mathbb{N}}$ in X converges to a limit $x \in X$. A complete normed space $(X, \|\cdot\|)$ is called a *Banach space*. Similarly, a complete inner product space $(X, \langle\cdot,\cdot\rangle)$ is called a *Hilbert space*.

We will return to Hilbert spaces in Chapters 5 and 6, and focus now on examples of Banach spaces. We have seen already that $(\mathbb{R}, |\cdot|)$ is a Banach space. It is an almost immediate consequence that so is $(\mathbb{R}^d, \|\cdot\|_\infty)$.

Theorem 3.9. $(\mathbb{R}^d, \|.\|_\infty)$ is a Banach space.

PROOF. Suppose $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}^d$. Then for any $\varepsilon > 0$ we can find N such that

$$\|x_n - x_m\|_\infty = \max_{1\le j\le d} |x_{nj} - x_{mj}| < \varepsilon, \quad n, m > N,$$

where x_{nj} is the jth entry in the vector x_n. It follows that for every $j = 1, \ldots, d$, we have

$$|x_{nj} - x_{mj}| < \varepsilon, \quad n, m > N.$$

In other words, the sequence $(x_{nj})_{n\in\mathbb{N}}$ is Cauchy in $(\mathbb{R}, |\cdot|)$. Since $(\mathbb{R}, |\cdot|)$ is complete, it follows that there exists $y_j \in \mathbb{R}$ such that $x_{nj} \to y_j$, and we can find N_j such that

$$|x_{nj} - y_j| < \varepsilon, \quad n > N_j.$$

Now let $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$ and $N = \max_{1\le j\le d} N_j$. Then

$$\|x_n - y\|_\infty = \max_{1\le j\le d} |x_{nj} - y_j| < \varepsilon, \quad n > N.$$

Thus $x_n \to y \in \mathbb{R}^d$, and this completes the proof that $(\mathbb{R}^d, \|\cdot\|_\infty)$ is complete. □

We move on to examples of sequence spaces.

Theorem 3.10. $(\ell^1, \|\cdot\|_1)$ is a Banach space.

PROOF. A useful strategy when proving that a metric space is complete is to split the proof into three steps.
Step 1: Find a candidate for the limit.

Let $(x_n)_n$ be a Cauchy sequence in ℓ^1. We denote the nth element of the sequence by

$$x_n = (x_1^{(n)}, x_2^{(n)}, ...).$$

Fix $\varepsilon > 0$. Note that for any fixed coordinate j, we have

$$|x_j^{(m)} - x_j^{(n)}| \leq \|x_m - x_n\|_1 < \varepsilon$$

for sufficiently large m and n, so the sequence $(x_j^{(n)})_{n\in\mathbb{N}}$ is Cauchy in $(\mathbb{R}, |\cdot|)$. Since $(\mathbb{R}, |\cdot|)$ is complete, we can find $z_j \in \mathbb{R}$ such that $x_j^{(n)} \to z_j$ as $n \to \infty$. Hence, a reasonable candidate for the limit of (x_n) (should it converge) is the sequence

$$z = (z_1, z_2, z_3, \ldots).$$

Step 2: Show that $z \in \ell^1$.

We have that

$$\sum_{j=1}^{N} |z_j| = \sum_{j=1}^{N} |\lim_{n\to\infty} x_j^{(n)}| = \sum_{j=1}^{N} \lim_{n\to\infty} |x_j^{(n)}| = \lim_{n\to\infty} \sum_{j=1}^{N} |x_j^{(n)}|,$$

where the interchange of limit and (finite) sum in $\mathbb{R}$ does not pose a problem (make sure you agree with this!). Treating the sum on the right hand side above, we see that

$$\sum_{j=1}^{N} |x_j^{(n)}| \leq \sum_{j=1}^{\infty} |x_j^{(n)}| \leq C, \tag{3.3}$$

since $x_n \in \ell^1$. Moreover, as the sequence (x_n) is Cauchy, it is bounded by Lemma 3.6, so there exists a universal C such that (3.3) holds for all $n \in \mathbb{N}$. By taking the limit $n \to \infty$ on both sides in this inequality, we get

$$\sum_{j=1}^{N} |z_j| = \lim_{n\to\infty} \sum_{j=1}^{N} |x_j^{(n)}| \leq C,$$

and since this holds for arbitrary N, it also holds as $N \to \infty$. We get

$$\sum_{j=1}^{\infty} |z_j| \leq C,$$

and conclude that $z \in \ell^1$.

Step 3: Show that $x_n \to z$.

Finally, we prove that $\|x_n - z\|_1 \to 0$ as $n \to \infty$. Fix $\varepsilon > 0$, and find N_1 such that

$$\|x_n - x_m\|_1 < \varepsilon, \quad n, m > N_1.$$

For any fixed N, we then have

$$\sum_{j=1}^{N} |x_j^{(n)} - x_j^{(m)}| \leq \|x_n - x_m\|_1 < \varepsilon, \quad n, m > N_1.$$

In particular, this holds as $m \to \infty$. We get

$$\lim_{m\to\infty} \sum_{j=1}^{N} |x_j^{(n)} - x_j^{(m)}| = \sum_{j=1}^{N} |x_j^{(n)} - \lim_{m\to\infty} x_j^{(m)}| = \sum_{j=1}^{N} |x_j^{(n)} - z_j| \leq \varepsilon.$$

As this holds for arbitrary N, it must also hold in the limit $N \to \infty$. We get

$$\|x_n - z\|_1 = \sum_{j=1}^{\infty} |x_j^{(n)} - z_j| \leq \varepsilon, \quad n \geq N_1,$$

and $x_n \to z$ as $n \to \infty$. □

By similar reasoning, one can show the following.

Theorem 3.11. $(\ell^p, \|\cdot\|_p)$ is a Banach space for any $1 \leq p \leq \infty$.

EXERCISE 3.2.3. *Show that* $(\ell^\infty, \|\cdot\|_\infty)$ *is a Banach space.*

Inspired by Theorem 3.11, it is tempting to suggest that also the function spaces $(C[a,b], \|\cdot\|_p)$ are complete. However, we have seen that this is not the case for $p = 1$ (recall Example 3.1.7). In fact, it is not the case for any $1 \leq p < \infty$.

Theorem 3.12. $(C[a,b], \|\cdot\|_\infty)$ is a Banach space.

Before we pursue the proof of Theorem 3.12, let us briefly review some notions and results from the theory of continuous functions. Recall that a function f defined on an interval I is continuous at a point $x_0 \in I$ if for every $\varepsilon > 0$ we can find $\delta > 0$ such that

$$|x - x_0| < \delta \quad \Rightarrow \quad |f(x) - f(x_0)| < \varepsilon.$$

We say that the function is continuous on the interval I if it is continuous at every point $x_0 \in I$.

Definition 3.2.4. Let (f_n) be a sequence of functions on an interval $I \subset \mathbb{R}$.

i) We say that (f_n) *converges pointwise* to a limit function f if for any given $\varepsilon > 0$ and $x \in I$ there exists N such that

$$|f_n(x) - f(x)| < \varepsilon, \quad n \geq N.$$

ii) We say that (f_n) *converges uniformly* to a limit function f if for any given $\varepsilon > 0$ there exists N such that

$$|f_n(x) - f(x)| < \varepsilon, \quad n \geq N,$$

for all $x \in I$.

There is a substantial difference between pointwise and uniform convergence. While pointwise convergence allows you to find an N, which might depend on both ε and x, such that $|f_n(x) - f(x)| < \varepsilon$ whenever $n \geq N$, uniform convergence implies that there exists a *universal* N such that $|f_n(x) - f(x)| < \varepsilon$ *for all* $x \in I$, $n \geq N$. The latter implies the former.

Remark 3.2.5. It is clear from the definition of $\|\cdot\|_\infty$ on $C[a,b]$ that the uniform convergence of a sequence (f_n) in $C[a,b]$ to a limit function f is equivalent to saying that $\|f_n - f\|_\infty \to 0$ as $n \to \infty$.

Finally, we recall the following.

Theorem 3.13. Let (f_n) be a sequence of continuous functions on $[a,b]$ which converges uniformly to a limit function f. Then f is continuous on $[a,b]$.

PROOF. We want to show that for any fixed $y \in [a,b]$ and $\varepsilon > 0$ we can find $\delta > 0$ such that

$$|x-y| < \delta \quad \Rightarrow \quad |f(x)-f(y)| < \varepsilon.$$

By the uniform convergence of (f_n) to f, there exists an N such that

$$|f_n(x)-f(x)| < \frac{\varepsilon}{3} \quad \text{for all } x \in [a,b], n \geq N.$$

Moreover, the function f_N is continuous, so there exists a $\delta > 0$ such that

$$|x-y| < \delta \quad \Rightarrow \quad |f_N(x)-f_N(y)| < \frac{\varepsilon}{3}.$$

It follows that

$$\begin{aligned} |f(x)-f(y)| &\leq |f(x)-f_N(x)| + |f_N(x)-f_N(y)| + |f_N(y)-f(y)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \end{aligned}$$

whenever $|x-y| < \delta$. □

PROOF OF THEOREM 3.12.

Step 1: Find a candidate for the limit.

Fix $x \in [a,b]$ and note that

$$|f_n(x)-f_m(x)| \leq \|f_n - f_m\|_\infty = \max_{a \leq x \leq b} |f_n(x)-f_m(x)|.$$

Thus if (f_n) is a Cauchy sequence in $(C[a,b], \|\cdot\|_\infty)$, then $(f_n(x))_{n\in\mathbb{N}}$ is a Cauchy sequence in $(\mathbb{R}, |\cdot|)$. Since $(\mathbb{R}, |\cdot|)$ is complete, there exists a point $f(x) \in \mathbb{R}$ such that $f_n(x) \to f(x)$. A reasonable candidate for the limit is the function f given by pointwise limits.

Step 2: Show that $f \in C[a,b]$.

We observe that the convergence of f_n to f is not only pointwise, but in fact uniform; Since (f_n) is Cauchy, there is for every $\varepsilon > 0$ an integer N such that

$$\|f_n - f_m\|_\infty = \max_{a \leq x \leq b} |f_n(x)-f_m(x)| < \frac{\varepsilon}{2}, \quad n,m \geq N.$$

In particular, this holds as $m \to \infty$, and we get

$$\max_{a \leq x \leq b} |f_n(x)-f(x)| \leq \frac{\varepsilon}{2} < \varepsilon, \quad n \geq N. \tag{3.4}$$

Thus, f_n converges uniformly to f on the interval $[a,b]$, and it follows by Theorem 3.13 that $f \in C[a,b]$.

Step 3: Show that $f_n \to f$.

This is immediate from (3.4). □

Remark. In the proof of Theorem 3.12, we are assuming for simplicity that the functions in $C[a,b]$ are real-valued. Note, however, that Theorem 3.12 is true also for complex-valued functions.

We close this subsection with a useful result on completeness in subsets.

Theorem 3.14.

i) A subset Y of a complete metric space (X, d) is itself a complete metric space (with the inherited metric) if and only if Y is closed.

ii) A subspace Y of a Banach space $(X, \|\cdot\|)$ is itself a Banach space (with the inherited norm) if and only if Y is closed.

PROOF OF THEOREM 3.14 I). Suppose Y is a complete metric space (with the inherited metric d), and choose any point x in the closure $\overline{Y}$. Then there exists a sequence (x_n) in Y such that $x_n \to x$. Since the sequence (x_n) is convergent (in X), it is necessarily Cauchy by Lemma 3.7. Finally, since Y is complete, we must have $x \in Y$. This shows that $\overline{Y} = Y$, and thus Y is closed.

Conversely, suppose Y is closed, and let (x_n) be a Cauchy sequence in Y. Since (X, d) is complete, we have $x_n \to x$ with $x \in X$. In other words, x is a limit point of Y. Since $Y = \overline{Y}$ by assumption, it follows that $x \in Y$, so the Cauchy sequence (x_n) converges in Y, and Y is complete. □

3.3. Completions

We have seen that not all metric (or normed) spaces are complete, and in particular $(C[a, b], \|\cdot\|_p)$ is not complete for any $1 \leq p < \infty$. However, every metric space and every normed space can be "made complete" by adding some more elements to the space. In this section we introduce the concepts needed to give a precise explanation of this vaguely formulated claim.

3.3.1. Isometries and isomorphisms. Recall that in mathematics an *isomorphism* is a bijective (one-to-one and onto) map that preserves the essential structure of something. Similarly, an *isometry* is a map that preserves distances.

Definition 3.3.1. Given two metric spaces (X, d_X) and (Y, d_Y).

- A map $\varphi : X \to Y$ is called an **embedding** and it is an **isometric embedding** of X into Y if for all $x_1, x_2 \in X$ we have
$$d_X(x_1, x_2) = d_Y\left(\varphi(x_1), \varphi(x_2)\right).$$
- If $\varphi : X \to Y$ is surjective and isometric, then we say that (X, d_X) and (Y, d_Y) are **isometric**.

The function φ is called an **isometry**.

Note that we did not require an isometry to be injective, because it is a consequence of the isometry condition.

Lemma 3.15. Suppose φ is an isometry between (X, d_X) and (Y, d_Y). Then φ is injective.

PROOF. For two points $x_1 \neq x_2$ in X we have $0 \neq d_X(x_1, x_2) = d_Y\left(\varphi(x_1), \varphi(x_2)\right)$ and thus φ is injective. □

Example 3.3.2. The spaces $C[0,1]$ and $C[a,b]$ (for $a < b$) of continuous, real-valued functions endowed with the supremum norm $\|\cdot\|_\infty$ are isometric. An isometry between these spaces is given by

$$\varphi : C[0,1] \to C[a,b], \quad f(\cdot) \to f\left(\frac{\cdot - a}{b-a}\right).$$

To show that this is an isometry, we calculate that:

$$\begin{aligned} d_{C[0,1]}(f,g) &= \max_{x\in[0,1]} |f(x) - g(x)| \\ &= \max_{x\in[a,b]} \left| f\left(\frac{x-a}{b-a}\right) - g\left(\frac{x-a}{b-a}\right)\right| = d_{C[a,b]}(\varphi(f), \varphi(g)). \end{aligned}$$

We leave it to the reader to check the φ is a surjection. Note that the metrics on $C[0,1]$ and $C[a,b]$ are those induced by the supremum norm.

Definition 3.3.3. A **vector space isomorphism** is a bijective linear map $T : X \to Y$ between vector spaces X and Y (over the same field $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$). If such a function exists, the vector spaces X and Y are **isomorphic**, and we write

$$X \cong Y.$$

Recall from Section 2.1 that $T : X \to Y$ is a linear map if

$$T(x+y) = T(x) + T(y) \quad \text{and} \quad T(\lambda x) = \lambda T(x), \quad \text{for all} \quad x, y \in X,\ \lambda \in \mathbb{F}.$$

Example 3.3.4. Regarded as a real vector space (i.e. where $\mathbb{F} = \mathbb{R}$), the space $\mathbb{C}^n$ is isomorphic to Euclidean space $\mathbb{R}^{2n}$ via the isomorphism

$$z = (x_1 + iy_1, \ldots, x_n + iy_n) \to (x, y) = (x_1, \ldots, x_n, y_1, \ldots, y_n).$$

Example 3.3.5. Recall from Example 2.1.2 that the set of polynomials with real-valued coefficients of degree at most n, denoted $\mathcal{P}_n$, is a vector space. It is isomorphic to Euclidean space $\mathbb{R}^{n+1}$, because

$$T : \mathcal{P}_n \to \mathbb{R}^{n+1}, \quad a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 \to (a_0, a_1, \ldots, a_n) \tag{3.5}$$

is an isomorphism.

Exercise 3.3.6. *Show that the map $T : \mathcal{P}_n \to \mathbb{R}^{n+1}$ in (3.5) is both linear and bijective.*

Definition 3.3.7. We say that two normed spaces $(X, \|.\|_X)$ and $(Y, \|.\|_Y)$ are called **isometrically isomorphic** if the isomorphism T between the vector spaces X and Y is also an isometry.

For instance, the spaces of real-valued continuous functions $C[0,1]$ and $C[a,b]$ are isometric; returning to example 3.3.2, it is not difficult to check that the provided isometry is also a vector space isomorphism.

Example 3.3.8. i) The vector space of polynomials of degree at most one, $\mathcal{P}_1(\mathbb{R})$, is embedded in Euclidean space $\mathbb{R}^3$, since

$$\mathcal{P}_1 \cong \mathbb{R}^2 \subset \mathbb{R}^3.$$

ii) The identity operator provides an embedding of the vector space of continuously differentiable functions on an interval I into the vector space of continuous functions on the same interval:

$$C^1(I,\mathbb{R}) \hookrightarrow C(I,\mathbb{R}).$$

3.3.2. Dense subsets and separability.

Definition 3.3.9. A subset $M \subset X$ of a metric space (X,d) is **dense** in X if its closure is the whole space:

$$M \text{ is dense in } X \quad \Leftrightarrow \quad \overline{M} = X.$$

Hence, the subspace (M,d) is **densely embedded** in (X,d).

By going back to the definition of closure in a metric space, one can see that $M \subset X$ is dense if and only if for every $x \in X$ and every $\epsilon > 0$ there exists some $m \in M$ such that $d(x,m) < \epsilon$. In a sense, one can think that a dense subset M contains "almost all of X".

Example 3.3.10.
(1) $\mathbb{Q}$ is dense in $\mathbb{R}$: for any $\lambda \in \mathbb{R}$ and $\epsilon > 0$, there is a rational number $q \in \mathbb{Q}$ such that $d(q,\lambda) < \epsilon$. We will not prove this fact in this course.
(2) $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$ and $\mathbb{Q}^n + i\,\mathbb{Q}^n$ is dense in $\mathbb{C}^n$.
(3) The space c_f of sequences with finitely many non-zero entries is dense in ℓ^p for every $1 \leq p < \infty$ (Exercise).
(4) c_f is dense in the normed space $(c_0, \|.\|_\infty)$, where c_0 denotes the space of real-valued sequences converging to zero.
The argument is based on identifying for a given $x \in c_0$ an approximant y to x in c_f.
Since x is in c_0, i.e. there exists an $N \in \mathbb{N}$ such that $|x_n| < \varepsilon$ for all $n \geq N$.
We set $y := (x_1, ..., x_N, 0, 0, ...)$, then $y \in c_f$ and we have $\|x - y\|_\infty < \varepsilon$.

Another very important example of a dense subset is given by the following result.

Theorem 3.16 (Stone-Weierstrass)**.** Let $[a,b]$ be a bounded interval in $\mathbb{R}$. The space $\mathcal{P}$ of polynomials is dense in $(C[a,b], \|\cdot\|_\infty)$. In other words, for any $f \in C[a,b]$ and $\varepsilon > 0$, there exists a polynomial $p \in \mathcal{P}$ such that $\|f - p\|_\infty < \varepsilon$.

There are many versions of the Stone-Weierstrass theorem, and the one stated above is perhaps the most classical. Another variant established by Weierstrass is that for continuous periodic functions: let us denote by $\mathcal{T}$ the space of all *trigonometric polynomials*, that is all functions $t_n(x)$ of the form

$$t_n(x) = c_{-n}e^{-2\pi inx} + \cdots + c_{-1}e^{-2\pi ix} + c_0 + c_1e^{2\pi ix} + \cdots + c_ne^{2\pi inx}, \quad n \in \mathbb{N}.$$

Theorem 3.17 (Weierstrass)**.** Suppose f is a continuous, 1-periodic function. Then for every $\varepsilon > 0$ there exists a trigonometric polynomial t such that

$\|f - t\|_\infty < \epsilon$. In other words, $\mathcal{T}$ is dense in the space of all 1-periodic continuous functions (with respect to the supremum norm).

The proof of the Stone-Weierstrass theorem is outside the scope of this course.

Definition 3.3.11. A metric space is said to be **separable** if it contains a countable dense set:

$$X \text{ separable} \quad \Leftrightarrow \quad \exists\, (x_n)_{n\in\mathbb{N}} \subset X \text{ such that } \overline{(x_n)_{n\in\mathbb{N}}} = X.$$

Example 3.3.12. Since $\mathbb{Q}$ is countable, and $\overline{\mathbb{Q}} = \mathbb{R}$ (with respect to the norm $\|\cdot\|_1 = |\cdot|$), it follows that $(\mathbb{R}, |\cdot|)$ is separable. Using this, one can show that all the spaces $\mathbb{R}^n$, $\mathbb{C}^n$, $\ell^p(\mathbb{R})$ and $\ell^p(\mathbb{C})$ with $1 \leq p < \infty$ are separable.

However, as we will see from the following proposition, ℓ^∞ is *not* separable. In this respect, the space ℓ^∞ is "bigger" than any other ℓ^p.

Proposition 3.18. The space $(\ell^\infty, \|\cdot\|_\infty)$ is not separable.

Proof. For an index set $I \subseteq \mathbb{N}$ we define vectors $e_I \in \ell^\infty$ by $(e_I)_n = 1$ for $n \in I$ and $(e_I)_n = 0$ for $n \notin I$. Note that $\|e_I - e_J\|_\infty = 1$ whenever $I \neq J$. We define a collection of disjoint open balls by $\mathcal{B} = \{B_{1/2}(e_I) : I \subseteq \mathbb{N}\}$.
Recall from Chapter 1.3 that there exist uncountably many binary sequences of zeros and ones. Hence $\mathcal{B}$ is uncountable. Now let M be *any* dense set in ℓ^∞. Then because M is dense in ℓ^∞, each of the elements in $\mathcal{B}$ must contain an element of M. Hence, M is uncountable and this shows ℓ^∞ does not contain a countable dense subset. Consequently, ℓ^∞ is not separable. □

3.3.3. The completion theorem. With density of subsets and embeddings defined, we are finally equipped to state precisely what is meant by saying that any metric (or normed) space can be "made complete".

Theorem 3.19. Every metric (normed) space is densely and isometrically embedded in a complete metric (normed) space.

For the case of a normed space $(X, \|\cdot\|_X)$, Theorem 3.19 states that there exists a Banach space $(\tilde{X}, \|\cdot\|_{\tilde{X}})$, and an injective linear mapping $T : X \to \tilde{X}$ such that $T(X)$ is dense in $\tilde{X}$, and

$$\|x - y\|_X = \|T(x) - T(y)\|_{\tilde{X}} \quad \text{for all } x, y \in X.$$

We call $(\tilde{X}, \|\cdot\|_{\tilde{X}})$ the **completion** of $(X, \|\cdot\|_X)$.

Example 3.3.13.
(1) The completion of the metric space $(\mathbb{Q}, |\cdot|)$ is the complete metric space $(\mathbb{R}, |\cdot|)$
(2) If we complete the normed space of smooth functions $C^\infty([0,1], \mathbb{R})$ with respect to the supremum norm $\|\cdot\|_\infty$, we get the Banach space $(C[0,1], \|\cdot\|_\infty)$.

Proposition 3.20. For $1 \leq p < \infty$, the normed space $(C[a,b], \|\cdot\|_p)$ has a completion which we denote by $L^p[a,b]$.

In this course we will not define precisely what the space $L^p[a,b]$ is. However, you should think of this as the space of all (measurable) functions such that the integral

$$\int_a^b |f(x)|^p \, dx$$

exists and is finite. The space $L^p[a,b]$ is equipped with the norm

$$\|f\|_{L^p[a,b]} = \left(\int_a^b |f(x)|^p \right)^{1/p}.$$

In particular, it contains all functions f where $|f|^p$ is Riemann integrable. The case $p = 2$ is of special interest; $L^2[a,b]$ is called *the space of square-integrable functions*. A deep result in analysis states that $L^2[a,b] \cong \ell^2$.

3.4. Banach's Fixed Point Theorem

The *Banach fixed point theorem* or *contraction theorem* concerns certain mappings (so-called *contractions*) of a complete metric space into itself. It states conditions sufficient for the existence and uniqueness of a *fixed point*, which we will see is a point that is mapped onto itself. The theorem also gives an iterative process by which we can obtain approximations to the fixed point and error bounds.

Definition 3.4.1. A **fixed point** of a mapping $T : X \to X$ of a set X into itself is an $x \in X$ which is mapped onto itself, that is

$$Tx = x.$$

Example 3.4.2.

i) A translation $x \to x + a$ in $\mathbb{R}$ has no fixed points.
ii) A rotation of the plane has a single fixed point, namely the center of rotation.
iii) The mapping $x \to x^2$ on $\mathbb{R}$ has two fixed points; 0 and 1.
iv) The projection $(x_1, x_2) \to (x_1, 0)$ on $\mathbb{R}^2$ has infinitely many fixed points; all points of the form $(x, 0)$.

Banach's fixed point theorem is an existence and uniqueness theorem for fixed points of certain mappings. As we will see from the proof, it also provides us with a constructive procedure for getting better and better approximations of the fixed point. This procedure is called *iteration*; we start by choosing an arbitrary x_0 in a given set, and calculate recursively a sequence $x_0, x_1, x_2 \ldots$ by letting

$$x_{n+1} = Tx_n, \quad n = 0, 1, 2 \ldots$$

Such iteration procedures appear in nearly every branch of applied mathematics, and Banach's fixed point theorem is often what guarantees convergence of the scheme and uniqueness of the solution.

Definition 3.4.3. Let (X, d) be a metric space. A mapping $T : X \to X$ is called a **contraction** on X if there exists a positive constant $K < 1$ such that

$$d\,(T(x), T(y)) \leq K d(x, y) \quad \text{for all } x, y \in X. \tag{3.6}$$

Geometrically, this means that the images $T(x)$ and $T(y)$ are closer together than the points x and y. Note in particular that if $(X, \|\cdot\|)$ is a normed space, then T is a contraction on X if there exists a positive constant $K < 1$ such that

$$\|T(x) - T(y)\| \leq K\|x - y\| \quad \text{for all } x, y \in X.$$

Theorem 3.21 (Banach's fixed point theorem)**.** Let (X, d) be a complete metric space and let $T : X \to X$ be a contraction on X. Then T has a unique fixed point $x \in X$ (such that $T(x) = x$).

Proof. Let us choose any $x_0 \in X$, and define the sequence (x_n), where

$$x_{n+1} = T(x_n), \quad n = 0, 1, 2, \ldots \tag{3.7}$$

Our proof strategy will be to 1) show that this sequence is Cauchy; 2) show that its limit is a fixed point in X; and 3) show that the fixed point is unique.

Step 1: By (3.6) and (3.7), we have that

$$\begin{aligned} d(x_{m+1}, x_m) &= d(T(x_m), T(x_{m-1})) \\ &\leq K d(x_m, x_{m-1}) \\ &= K d(T(x_{m-1}), T(x_{m-2})) \\ &\leq K^2 d(x_{m-1}, x_{m-2}) \\ &\ldots \leq K^m d(x_1, x_0). \end{aligned}$$

Hence by the triangle inequality we get (for $n \geq m$) that

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x_{m+1}) + d(x_{m+1}, x_{m+2}) + \cdots + d(x_{n-1}, x_n) \\ &\leq (K^m + K^{m+1} + \cdots + K^{n-1}) d(x_1, x_0) = K^m \frac{1 - K^{n-m}}{1 - K} d(x_0, x_1), \end{aligned}$$

where in the last equality we have used the summation formula for a geometric series. Since $0 < K < 1$, we have $1 - K^{n-m} < 1$, and consequently

$$d(x_m, x_n) \leq \frac{K^m}{1 - K} d(x_1, x_0). \tag{3.8}$$

Since $0 < K < 1$ and $d(x_0, x_1)$ are fixed, it is clear that that we can make $d(x_m, x_n)$ as small as we please by choosing m sufficiently large (and $n > m$). This proves that (x_n) is Cauchy. Finally, since (X, d) is complete, there exists an $x \in X$ such that $x_n \to x$.

Step 2: To show that x is a fixed point, we consider the distance $d(x, T(x))$. From the triangle inequality and (3.6), we get

$$\begin{aligned} d(x, T(x)) &\leq d(x, x_m) + d(x_m, T(x)) \\ &= d(x, x_m) + d(T(x_{m-1}), T(x)) \\ &\leq d(x, x_m) + K d(x_{m-1}, x), \end{aligned}$$

and because $x_n \to x$ it is clear that we can make this distance as small as we please by choosing m sufficiently large. We conclude that

$$d(x, T(x)) = 0 \quad \Rightarrow \quad T(x) = x,$$

so $x \in X$ is a fixed point of T.

Step 3: Suppose there are two fixed points $x = T(x)$ and $\tilde{x} = T(\tilde{x})$. Then from (3.6) it follows that

$$d(x, \tilde{x}) = d(T(x), T(\tilde{x})) \leq K d(x, \tilde{x}),$$

which implies $d(x, \tilde{x}) = 0$ since $0 < K < 1$. Hence $x = \tilde{x}$, and the fixed point x of T is unique. □

Note that for Banach's fixed point theorem to hold, it is crucial that T is a contraction; it is not sufficient that (3.6) holds for $K = 1$, i.e. that

$$d(T(x), T(y)) \leq d(x, y) \quad \text{for all } x, y \in X.$$

To see this, observe that the maps $T_1, T_2 : \mathbb{R} \to \mathbb{R}$ given by $T_1(x) = x + 1$ and $T_2(x) = x$ both satisfy (3.6) with $K = 1$. The map T_1 has no fixed points, whereas T_2 has infinitely many.

Corollary 3.22 (Iterations and error bounds)**.** The iterative sequence (3.7) with arbitrary $x_0 \in X$ converges (under the assumptions in Banach's fixed point theorem) to the unique fixed point x of T. Error estimates are the **a prior estimate**

$$d(x_m, x) \leq \frac{K^m}{1 - K} d(x_0, x_1), \tag{3.9}$$

and the **posterior estimate**

$$d(x_m, x) \leq \frac{K}{1 - K} d(x_{m-1}, x_m). \tag{3.10}$$

The prior error bound (3.9) can be used at the beginning of a calculation for estimating the number of steps necessary for obtaining a given accuracy. The posterior bound (3.10) can be used at intermediate stages to check whether we are possibly converging faster than suggested by (3.9). We see that if two successive iterations x_m and $x_{m+1} = T(x_m)$ are nearly equal, then this guarantees that we are very close to the true fixed point x.

Proof of Corollary 3.22. The first statement is obvious from the proof of Banach's fixed point theorem. The prior bound (3.9) follows from (3.8) by letting $n \to \infty$. Finally let us establish (3.10). Since x is a fixed point and T is a contraction, we have

$$\begin{aligned} d(x_m, x) &= d(T(x_{m-1}), T(x)) \\ &\leq K d(x_{m-1}, x) \\ &\leq K \left(d(x_{m-1}, x_m) + d(x_m, x) \right), \end{aligned}$$

where in the last step we have used the triangle inequality. Rearranging terms, we arrive at (3.10). □

A classical application of Banach's fixed point theorem is Newton's method for finding roots of equations. Starting with a differentiable function f and an initial guess x_0 for a root of f, Newton's method suggests

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, 2, \dots \tag{3.11}$$

as a sequence of successively better approximations to the true root of f. We look at a specific example.

Example 3.4.4. Consider the equation $f(x) = x^2 - 3$, which we know has two roots, and let us apply Banach's fixed point theorem to determine when we can expect the scheme (3.11) to converge to $x = \sqrt{3}$. Setting

$$T(x) := x - \frac{f(x)}{f'(x)} = x - \frac{x^2 - 3}{2x} = \frac{1}{2}(x + \frac{3}{x}),$$

we see that T is a map from the closed set $[\sqrt{3}, \infty)$ into itself. Moreover, a point $x \in [\sqrt{3}, \infty)$ is a fixed point of T if and only if $f(x) = 0$. Finally, we observe that

$$\begin{aligned} d(T(x), T(y)) &= |T(x) - T(y)| \\ &= \frac{1}{2}\left|\left(x + \frac{3}{x}\right) - \left(y + \frac{3}{y}\right)\right| \\ &= \frac{1}{2}|x - y| \cdot \left|1 - \frac{3}{xy}\right| \\ &\leq \frac{1}{2}|x - y| = \frac{1}{2}d(x, y), \end{aligned}$$

for all $x, y \in [\sqrt{3}, \infty)$. Hence, T is a contraction on the complete space $\big([\sqrt{3}, \infty), |\cdot|\big)$, and by Banach's fixed point theorem we conclude that the scheme (3.11) converges to the root $x = \sqrt{3}$ for any starting point $x_0 \in [\sqrt{3}, \infty)$.

In fact, the scheme will converge to $x = \sqrt{3}$ for any starting point $x_0 \in (0, \infty)$; one can check that for any $x_0 \in (0, \sqrt{3})$, we have

$$x_1 = T(x_0) = \frac{1}{2}(x_0 + \frac{3}{x_0}) > \sqrt{3},$$

and we may therefore use Banach's fixed point theorem with the "new" starting point x_1.

3.5. Applications of Banach's fixed point theorem

The most interesting applications of Banach's fixed point theorem arise in connection with function spaces. The theorem then yields existence and uniqueness results for differential and integral equations, as we will now see.

3.5.1. Applications to integral equations. In this section we consider integral equations of the form

$$f(x) = \lambda \int_a^b k(x, y) f(y)\, dy + g(x), \tag{3.12}$$

where $f : [a, b] \to \mathbb{R}$ is an unknown function, $k : [a, b] \times [a, b] \to \mathbb{R}$ is a given function (called the **kernel**) and λ is a parameter. Such integral equations can be considered in various function spaces. In this section we consider (3.12) only on $(C[a, b], \|\cdot\|_\infty)$.

Recall that this is a complete normed space. We assume that $g \in C[a,b]$, and that the kernel k is continuous on the square $[a,b] \times [a,b]$.

Equation (3.12) can be restated as $T(f) = f$, where

$$T(f)(x) = g(x) + \lambda \int_a^b k(x,y) f(y)\, dy. \tag{3.13}$$

Since g and k are both continuous, this defines an operator $T : C[a,b] \to C[a,b]$. Let us now determine for which values of λ the map T is a contraction. Note first that since k is continuous, it must also be bounded

$$|k(x,y)| \le c \quad \text{for all } (x,y) \in [a,b] \times [a,b]. \tag{3.14}$$

We have

$$\begin{aligned} d(T(f_1), T(f_2)) &= \|T(f_1) - T(f_2)\|_\infty \\ &= |\lambda| \max_{x \in [a,b]} \left| \int_a^b k(x,y) \left(f_1(y) - f_2(y)\right)\, dy \right| \\ &\le |\lambda| \max_{x \in [a,b]} \int_a^b |k(x,y)||f_1(y) - f_2(y)|\, dy \\ &\le c|\lambda| \max_{x \in [a,b]} |f_1(x) - f_2(x)| \int_a^b dy \\ &= c|\lambda|(b-a) d(f_1, f_2). \end{aligned}$$

Recall that T is a contraction if

$$d(T(f_1), T(f_2)) \le K d(f_1, f_2) \quad \text{for all } f_1, f_2 \in C[a,b]$$

for some constant $0 < K < 1$, and we see that this is indeed the case if

$$|\lambda| < \frac{1}{c(b-a)}. \tag{3.15}$$

Banach's fixed point theorem now gives:

Theorem 3.23. Suppose k and g in (3.12) are continuous on $[a,b] \times [a,b]$ and $[a,b]$, respectively, and assume that the parameter λ satisfies (3.15), with c defined in (3.14). Then the integral equation (3.12) has a unique solution $f \in C[a,b]$. This solution is the limit of the iterative sequence $(f_0, f_1, f_2, \ldots)$, where f_0 is any continuous function on $[a,b]$, and

$$f_{n+1}(x) = g(x) + \lambda \int_a^b k(x,y) f_n(y)\, dy, \quad n = 0, 1, 2, \ldots$$

3.5.2. Applications to differential equations. Let us consider the *initial value problem*

$$x'(t) = \frac{dx}{dt} = f(t,x), \quad x(t_0) = x_0, \tag{3.16}$$

where $f : A \subset \mathbb{R}^2 \to \mathbb{R}$ is a given function and $x(t)$ is an unknown function which we wish to determine. In this subsection we will use Banach's fixed point theorem to prove the famous *Picard-Lindelöf Theorem*, which guarantees the uniqueness and existence of a solution to (3.16).

Theorem 3.24 (Picard-Lindelöf)**.** Let f be continuous on a rectangle

$$R = \{(t,x) : |t-t_0| \le a,\ |x-x_0| \le b\},$$

and thus bounded on R, say $|f(x,t)| \le c$. Suppose that f satisfies a *Lipschitz condition* on R with respect to its second argument, meaning there exists a constant k such that

$$|f(t,x) - f(t,y)| \le k|x-y| \quad \text{for all } (t,x),(t,y) \in R.$$

Then the initial value problem (3.16) has a unique solution which exists on an interval $[t_0-\beta, t_0+\beta]$, where

$$\beta < \min\left\{a, \frac{b}{c}, \frac{1}{k}\right\}. \tag{3.17}$$

Proof. *Step 1: Equivalent formulation as an integral equation:* We observe first that if a function $x \in C^1[t_0-a, t_0+a]$ solves (3.16), then necessarily

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds \tag{3.18}$$

by integration. On the other hand, if $x \in C[t_0-a, t_0+a]$ fulfils (3.18), then x is a continuously differentiable solution to (3.16) (this follows from the Fundamental Theorem of Calculus). Thus, the initial value problem (3.16) for $x \in C^1[t_0-a, t_0+a]$ is equivalent to (3.18) for $x \in C[t_0-a, t_0+a]$.

Step 2: Constructing an operator T on a complete space to which we can apply Banach's fixed point theorem: For $J = [t_0-\beta, t_0+\beta]$ and $y \in C(J)$, define the operator

$$T(y)(t) := x_0 + \int_{t_0}^{t} f(s, y(s))\, ds, \quad t \in J.$$

Consider the set

$$X := \left\{ y \in C(J) \,:\, y(t_0) = x_0,\ \sup_{t\in J} |x_0 - y(t)| \le c\beta \right\}.$$

This is a closed subspace of $C(J)$ (endowed with the norm $\|\cdot\|_\infty$), so $(X, \|\cdot\|_\infty)$ is complete.

Step 3: Observe that $T : X \to X$: For $y \in X$, we need to show that $T(y) \in X$. Observe that $T(y)(t_0) = x_0$. Moreover, we have

$$|x_0 - T(y)(t)| = \left| \int_{t_0}^{t} f(s, y(s))\, ds \right| \le |t-t_0| \cdot \max_{t\in J} |f(t, y(t))| \le c\beta,$$

so $T(y) \in X$.

Step 4: Showing T is a contraction: Fix $y_1, y_2 \in X$. We have

$$\begin{aligned} |T(y_1)(t) - T(y_2)(t)| &= \left| \int_{t_0}^{t} f(s, y_1(s)) - f(s, y_2(s))\, ds \right| \\ &\le |t-t_0| \cdot \max_{s\in J} k|y_1(s) - y_2(s)| \\ &\le k\beta d(y_1, y_2). \end{aligned}$$

The right hand side above is independent of t, so taking the maximum over $t \in J$ on both sides, we get

$$d(T(y_1), T(y_2)) \leq k\beta d(y_1, y_2).$$

Recalling (3.17), we see that $k\beta < 1$, so T is a contraction on X.

Step 5: Conclusion: Banach's fixed point theorem implies that T has a unique fixed point $x \in X$ such that

$$x(t) = T(x)(t) = x_0 + \int_t^{t_0} f(s, x(s))\, ds.$$

It thus follows from Step 1 that (3.16) has a unique, continuous solution $x(t)$ on the interval $[t_0 - \beta, t_0 + \beta]$. □

In addition to existence and uniqueness of a solution, Banach's fixed point theorem provides us with an iterative procedure for finding the solution.

Corollary 3.25 (Picard iteration)**.** Under the assumptions of the Picard-Lindelöf Theorem, the sequence given by

$$x_0(t) = x_0, \quad x_{n+1}(t) = T(x_n)(t) = x_0 + \int_{t_0}^{t} f(s, x_n(s))\, ds, \quad n = 0, 1, 2, \ldots,$$

converges uniformly to the unique solution $x(t)$ on $J = [t_0 - \beta, t_0 + \beta]$.

Note, however, that the practical usefulness of Picard iteration is rather limited, due to the integrations involved. This is illustrated by the following example.

Example 3.5.1. The first Picard iteration for the initial value problem

$$x'(t) = \sqrt{x} + x^3, \quad x(1) = 2,$$

is given by

$$x_1(t) = 2 + \int_1^t \left(\sqrt{2} + 2^3\right) ds = 2 + (\sqrt{2} + 8)(t - 1).$$

The second is

$$\begin{aligned} x_2(t) &= 2 + \int_1^t \left(\sqrt{x_1(s)} + (x_1(s))^3\right) ds \\ &= 2 + \int_1^t \left(\sqrt{2 + (\sqrt{2} + 8)(s - 1)} + (2 + (\sqrt{2} + 8)(s - 1))^3\right) ds. \end{aligned}$$

We see that even the second integral in Picard iteration looks quite uninviting. The next iterations $x_3, x_4, \ldots$ will involve even worse integrals, illustrating that Picard iteration is often of limited use in practice.

CHAPTER 4

Bounded linear operators between normed spaces

In this section we focus on bounded linear operators between normed spaces. We define the operator norm, and see that the vector space of bounded linear operators between two normed spaces is itself a normed space when endowed with the operator norm.

4.1. Revisiting linear operators

Recall from Section 2.1 that a transformation, or operator, $T : V \to W$ between two vector spaces V and W over the same field $\mathbb{F}$ is *linear* if it respects linear structure, meaning that

$$T(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 T(v_1) + \alpha_2 T(v_2) \quad \text{for all } v_1, v_2 \in V, \alpha_1, \alpha_2 \in \mathbb{F}.$$

When we work with a linear operator $T : V \to W$, we often write Tv instead of $T(v)$.

Example 4.1.1.

(1) For any vector space V, the *identity operator* $I_V : V \to V$ defined by $I_V v = v$ for all $v \in V$, is a linear operator.

(2) For any two vector spaces V and W, the *zero operator* $0 : V \to W$ defined by $0v = 0$ for all $v \in V$, is a linear operator.

(3) *Differentiation:* Let X be the vector space of all polynomials on $[a, b]$ Then differentiation defines a linear operator T on X by setting

$$Tx(t) = x'(t).$$

(4) *Integration:* A linear operator T from $C[a, b]$ into itself can be defined by

$$Tx(t) = \int_a^t x(\tau)\, d\tau, \quad t \in [a, b].$$

(5) Another linear operator from $C[a, b]$ into itself can be defined by multiplication by t:

$$Tx(t) = tx(t).$$

(6) A *real matrix* $A = (\alpha_{jk})$ with r rows and n columns defines an operator $T : \mathbb{R}^n \to \mathbb{R}^r$ by means of matrix multiplication

$$y = Ax,$$

where $y \in \mathbb{R}^r$ and $x \in \mathbb{R}^n$.

For a linear operator $T : V \to W$, we denote by $\ker(T)$ its *kernel*

$$\ker(T) = \{v \in V \mid Tv = 0\} \subseteq V,$$

and by $\operatorname{ran}(T) \subseteq W$ its range. It is easily checked in the examples above that the kernels and ranges of the given operators are vector subspaces, and we recall from Section 2.1 that this is true in general.

Lemma 4.1. Let $T : V \to W$ be a linear operator between vector spaces V and W over the same field $\mathbb{F}$. Then

- The kernel $\ker(T)$ is a vector subspace of V.
- The range $\operatorname{ran}(T)$ is a vector subspace of W.

4.2. Bounded and continuous linear operators

4.2.1. Continuous operators. We are familiar with what it means for a real- or complex-valued function to be continuous in a point or on an interval. This definition has a natural extension to operators between metric (or normed) spaces.

Definition 4.2.1. Let $X = (X, d)$ and $Y = (Y, \tilde{d})$ be metric spaces. An operator $T : X \to Y$ is said to be *continuous at a point* $x_0 \in X$ if for every $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\tilde{d}(Tx, Tx_0) < \varepsilon \quad \text{for all } x \text{ satisfying} \quad d(x, x_0) < \delta.$$

T is said to be *continuous* if it is continuous at every point of X.

In particular, an operator $T : X \to Y$ between *normed* spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ is continuous at a point $x_0 \in X$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\|x - x_0\|_X < \delta \quad \Rightarrow \quad \|Tx - Tx_0\|_Y < \varepsilon.$$

Definition 4.2.2. An operator $T : X \to Y$ between normed spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ is called *Lipschitz continuous* if there exists a constant $L > 0$ such that

$$\|Tx - Tx'\|_Y \leq L\|x - x'\|_X \quad \text{for all } x, x' \in X.$$

The constant L is referred to as the *Lipschitz constant*.

Note in particular that a Lipschitz continuous operator from a space X to itself with Lipschitz constant $L < 1$ defines a *contraction* on X.

Exercise 4.2.3. *Show that a Lipschitz continuous operator between normed spaces is necessarily continuous.*

The following characterization of continuity will be useful as we move forward. Note that it applies to all operators between metric spaces, not just linear operators.

Theorem 4.2. An operator $T : X \to Y$ between metric spaces (X, d) and $(Y, \tilde{d})$ is continuous at a point $x_0 \in X$ if and only if for any sequence $(x_n)_n$ in X converging to x_0 we have that

$$x_n \to x_0 \quad \text{implies} \quad Tx_n \to Tx_0.$$

Proof. Assume T is continuous at x_0. Then for a given $\varepsilon > 0$ there exists $\delta > 0$ such that

$$d(x, x_0) < \delta \quad \text{implies} \quad \tilde{d}(Tx, Tx_0) < \varepsilon.$$

Let $x_n \to x_0$. Then there exists an N such that for all $n > N$ we have

$$d(x_n, x_0) < \delta.$$

Hence, for all $n > N$, we have

$$\tilde{d}(Tx_n, Tx_0) < \varepsilon.$$

By definition, this means that $Tx_n \to Tx_0$.

Conversely, say

$$x_n \to x_0 \quad \text{implies} \quad Tx_n \to Tx_0,$$

and suppose that T is *not* continuous at x_0. Then there is an $\varepsilon > 0$ such that for every $\delta > 0$ there is an $x \neq x_0$ satisfying

$$d(x, x_0) < \delta \quad \text{but} \quad \tilde{d}(Tx, Tx_0) \geq \varepsilon.$$

Then in particular, for $\delta = \frac{1}{n}$ there is an x_n satisfying

$$d(x_n, x_0) < \frac{1}{n} \quad \text{but} \quad \tilde{d}(Tx_n, Tx_0) \geq \varepsilon.$$

We clearly have $x_n \to x_0$, but (Tx_n) does not converge to Tx_0. This contradicts $Tx_n \to Tx_0$, so T must be continuous at x_0. □

Finally, we make the following simple observation:

Proposition 4.3. Given a normed space $(X, \|\cdot\|)$, the norm is a continuous mapping $x \to \|x\|$ of $(X, \|\cdot\|)$ into $\mathbb{R}$.

Proof. Exercise. □

4.2.2. Bounded linear operators. Continuity of an operator is naturally defined in any metric space. When we now turn our attention to *bounded* operators, we need the richer structure of a normed space.

Definition 4.2.4. Let X and Y be normed spaces. A linear operator $T : X \to Y$ is said to be *bounded* if there exists a real number c such that

$$\|Tx\| \leq c\|x\| \quad \text{for all } x \in X. \tag{4.1}$$

In equation (4.1), the norm on the left hand side is the norm in Y, whereas the norm on the right hand side is the norm in X. For simplicity we denote both by $\|\cdot\|$. Notice that a bounded linear operator maps bounded sets in X to bounded sets in Y. This motivates the name "bounded operator".

A natural question now is: what is the smallest possible c such that (4.1) holds for all nonzero $x \in X$? Dividing by $\|x\|$, we get

$$\frac{\|Tx\|}{\|x\|} \leq c \text{ for all non-zero } x \in X.$$

So the smallest possible c is actually the *least* upper bound of the left hand side, which is the supremum by definition. This quantity is denoted by $\|T\|$; we have

$$\|T\| = \sup \left\{ \frac{\|Tx\|_Y}{\|x\|_X} \, : \, x \neq 0 \right\}, \tag{4.2}$$

and $\|T\|$ is called the *norm* of the operator T. In the (relatively uninteresting) special case $X = \{0\}$, we define $\|T\| = 0$.

Notice that inserting $c = \|T\|$ in (4.1), we get

$$\|Tx\| \leq \|T\|\|x\|.$$

We will apply this formula frequently.

So far we have not justified the use of the word "norm" in this setting. This is done in the following lemma.

Lemma 4.4. Let T be a bounded linear operator. Then

(1) An alternative formula for the norm (4.1) of T is

$$\|T\| = \sup_{\|x\|=1} \|Tx\|.$$

(2) The norm defined by (4.2) satisfies the norm axioms stated in Definition 2.2.1.

PROOF. (1) For $x \neq 0$, we write $\|x\|_X = a$ and set $y = (1/a)x$. Then $\|y\|_X = 1$, and since T is linear we have

$$\begin{aligned}\|T\| &= \sup_{x\in X, x\neq 0} \frac{\|Tx\|_Y}{\|x\|_X} = \sup_{x\in X, x\neq 0} \frac{1}{a}\|Tx\|_Y \\ &= \sup_{x\in X, x\neq 0} \left\|T\left(\tfrac{1}{a}x\right)\right\|_Y = \sup_{y\in X, \|y\|_X=1} \|Ty\|_Y.\end{aligned}$$

(2) Exercise.

□

Let us compute the operator norm of a composition of operators.

Lemma 4.5. Let $(X, \|.\|_X), (Y, \|.\|_Y)$ and $(Z, \|.\|_Z)$ be normed spaces and $S : X \to Y$ and $T : Y \to Z$ linear bounded mappings. Then $\|TS\| \leq \|T\|\|S\|$. In particular, if $T : X \to X$ is a bounded linear operator, then $\|T^n\| \leq \|T\|^n$ for any $n \in \mathbb{N}$.

PROOF. $\|TS\| = \sup_{x\neq 0} \frac{\|T(Sx)\|_Z}{\|x\|_X} \leq \|T\| \sup_{x\neq 0} \frac{\|Sx\|_Y}{\|x\|_X} = \|T\|\|S\|$. □

Example 4.2.5. (1) The identity operator $I : X \to X$ on a normed space $X \neq \{0\}$ is bounded and has norm $\|I\| = 1$.

(2) The zero operator $0 : X \to Y$ between normed spaces is bounded and has norm $\|0\| = 0$.

(3) *Differentiation operator:* Let X be the normed space of all polynomials on $J = [0, 1]$ endowed with the norm $\|x\| = \max_{t\in J} |x(t)|$ (sup-norm). We can define the differentiation operator T on X by

$$Tx(t) = x'(t).$$

This operator is linear, but **not** bounded. Indeed let $x_n(t) = t^n$, with $n \in \mathbb{N}$. Then $\|x_n\| = 1$ and

$$Tx_n(t) = x_n'(t) = nt^{n-1}.$$

It follows that $\|Tx_n\| = n$ and $\|Tx_n\|/\|x_n\| = n$. Since $n \in \mathbb{N}$ is arbitrary, this shows that there is no fixed number c such that $\|Tx_n\|/\|x_n\| \leq c$. We thus conclude that T is not bounded.

(4) *Integral operators:* We can define an integral operator $T : C[0,1] \to C[0,1]$ by

$$y(t) = (Tx)(t) = \int_0^1 k(t,\tau)x(\tau)\,d\tau.$$

Recall from our discussion of integral equations that k is called the *kernel* of T (not to be confused with $\ker(T)$, which is a very different concept) and is assumed to be continuous on the closed square $[0,1]\times[0,1]$. This operator is linear. It is also bounded. To see this, observe first that the continuity of k guarantees that it is bounded, say $|k(t,\tau)| \le k_0$ for all $(t,\tau)\in[0,1]\times[0,1]$. Hence, we have

$$\begin{aligned}\|y\| = \|Tx\| &= \max_{0\le t\le 1}\left|\int_0^1 k(t,\tau)x(\tau)\,d\tau\right| \\ &\le \max_{0\le t\le 1}\int_0^1 |k(t,\tau)|\cdot|x(\tau)|\,d\tau \\ &\le k_0 \max_{0\le t\le 1}\int_0^1 x(\tau)\,d\tau = k_0\|x\|.\end{aligned}$$

Hence, T is bounded and $\|T\| \le k_0$. Note, however, that we can not conclude that $\|T\| = k_0$ (this is in general incorrect).

(5) *Shift operators:* On a given sequence space, say ℓ^∞, we can define the *left* and *right shift* operators by

$$Lx = (x_2, x_3, x_4, \ldots),$$

and

$$Rx = (0, x_1, x_2, x_3, \ldots),$$

where $x = (x_n) \in \ell^\infty$. These are linear operators from ℓ^∞ to itself. They are also bounded; observe that

$$\|Lx\| = \sup_{2\le j\le\infty}|x_j| \le \sup_{1\le j\le\infty}|x_j| = \|x\|,$$

and thus L is a bounded operator with $\|L\| \le 1$. Similarly, we have

$$\|Rx\| = \sup_{1\le j\le\infty}|x_j| = \|x\|,$$

so R is a bounded operator with $\|R\| = 1$.

(6) *Matrices:* Recall that a real matrix $A = (\alpha_{jk})$ with r rows and n columns defines a linear operator $T : \mathbb{R}^n \to \mathbb{R}^r$ by means of matrix multiplication

$$y = Ax, \tag{4.3}$$

where $y = (\eta_j) \in \mathbb{R}^r$ and $x = (\xi_j) \in \mathbb{R}^n$. In terms of components, (4.3) reads

$$\eta_j = \sum_{k=1}^n \alpha_{jk}\xi_k \quad \text{for each } j = 1, 2, \ldots, r.$$

Let us now show that T is a bounded operator. We endow $\mathbb{R}^r$ and $\mathbb{R}^n$ with the $\|\cdot\|_2$-norm, which we recall is the norm associated with the

classical inner product $\langle \cdot, \cdot \rangle$ on these spaces. Using the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\|Tx\|_2^2 &= \sum_{j=1}^{r} \eta_j^2 = \sum_{j=1}^{r} \left(\sum_{k=1}^{n} \alpha_{jk} \xi_k \right)^2 \\ &\leq \sum_{j=1}^{r} \left[\left(\sum_{k=1}^{n} \alpha_{jk}^2 \right)^{1/2} \left(\sum_{k=1}^{n} \xi_k^2 \right)^{1/2} \right]^2 \\ &= \|x\|_2^2 \sum_{j=1}^{r} \sum_{k=1}^{n} \alpha_{jk}^2.\end{aligned}$$

Thus, we have

$$\|Tx\|_2^2 \leq c^2 \|x\|_2^2 \quad \text{with } c^2 = \sum_{j=1}^{r} \sum_{k=1}^{n} \alpha_{jk}^2,$$

meaning that T is bounded and $\|T\| \leq c$.

Operator norms of linear mappings between finite-dimensional vector spaces are an important class of examples. We restrict our discussion to linear mappings $T : \mathbb{R}^n \to \mathbb{R}^n$. Suppose $(\mathbb{R}^n, \|.\|_\alpha)$ and $(\mathbb{R}^n, \|.\|_\beta)$ be the space of n-tuples with norms $\|.\|_\alpha$ and $\|.\|_\beta$ are norms for the domain and co-domain, resp. Then the operator norm of T is given by

$$\|T\|_{(\alpha,\beta)} = \sup_{\|x\|_\alpha = 1} \|Tx\|_\beta.$$

Note that the supremum is attained since $\{x \in \mathbb{R}^n | \, \|x\|_\alpha = 1\}$ is a closed bounded set and thus the continuous function $\|, \|_\beta$ attains its maximum.
Let A be the matrix representing T with respect to the standard basis of $\mathbb{R}^n$. Then we have that

- $\|T\|_{(1,1)} = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|$ is the maximal absolute column sum of A. (Exercise)
- $\|T\|_{(\infty,\infty)} = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$ is the maximal absolute row sum of A.

$$\begin{aligned}\|Tx\|_\infty &= \max_{1 \leq i \leq n} |\sum_{j=1}^{n} a_{ij} x_j| \\ &\leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}||x_j| \\ &\leq \Big(\max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| \Big) \max_{1 \leq j \leq n} |x_j| \\ &= \Big(\max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| \Big) \|x\|_\infty.\end{aligned}$$

Claim: $\|T\|_{(\infty,\infty)} = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$.
We construct a vector $\widehat{x} \in \mathbb{R}^n$ such that $\|T\widehat{x}\|_\beta = \|T\|_{(\infty,\infty)} \|\widehat{x}\|_\infty$. Let $\underline{i}$ be the row the maximal absolute row sum of A. Then $\widehat{x} := (\text{sign}(a_{i1}), ..., \text{sign}(a_{in}))^T$, where sign(x) denotes the sign of $x \in \mathbb{R}$, is the desired n-tuple.

- $\|T\|_{(2,2)}$ is the largest eigenvalue of A^*A. (Proof in the linear algebra chapter)

4.2.3. Bounded and continuous linear operators. Let us now show the remarkable property of linear operators that boundedness and continuity are equivalent concepts.

Theorem 4.6 (Continuity and boundedness)**.** Let $T : X \to Y$ be a linear operator between normed spaces X and Y. Then:

i) T is continuous if and only if T is bounded.
ii) If T is continuous at a single point, then T is continuous (everywhere).

PROOF. i) For $T = 0$ the statement is trivial, so let $T \neq 0$ and thus $\|T\| \neq 0$. Assume first that T is bounded. Consider any $x_0 \in X$ and any fixed $\varepsilon > 0$. Then, since T is linear, for every $x \in X$ with

$$\|x - x_0\| < \delta = \frac{\varepsilon}{\|T\|},$$

we have

$$\|Tx - Tx_0\| = \|T(x - x_0)\| \leq \|T\|\|x - x_0\| < \|T\|\delta = \varepsilon.$$

This shows that T is continuous at x_0, which was chosen arbitrarily, so T is continuous.

Conversely, assume that T is continuous, and consider an arbitrary point $x_0 \in X$. Then given any $\varepsilon > 0$ we can find $\delta > 0$ such that

$$\|Tx - Tx_0\| \leq \varepsilon \quad \text{whenever} \quad \|x - x_0\| \leq \delta.$$

Now take any $y \in X \setminus \{0\}$, and set

$$x = x_0 + \frac{\delta}{\|y\|}y. \quad \text{Then} \quad x - x_0 = \frac{\delta}{\|y\|}y,$$

and we have $\|x - x_0\| = \delta$. By the linearity of T we get

$$\|Tx - Tx_0\| = \|T(x - x_0)\| = \left\|T\left(\frac{\delta}{\|y\|}y\right)\right\| = \frac{\delta}{\|y\|}\|Ty\|,$$

and from the continuity of T it thus follows that

$$\frac{\delta}{\|y\|}\|Ty\| \leq \varepsilon \quad \Rightarrow \quad \|Ty\| \leq \frac{\varepsilon}{\delta}\|y\|.$$

Thus, T is bounded and $\|T\| \leq \varepsilon/\delta$.

ii) Continuity of T at a point implies boundedness of T by the second part of the proof of (i), which in turn implies continuity of T by (i).

□

Now recall that the kernel, or null space, of a linear operator T is the subspce

$$\ker(T) = \{x \in X : Tx = 0\}.$$

Corollary 4.7 (Closed kernels)**.** Let T be a bounded linear operator between normed spaces X and Y. Then:

i) $x_n \to x$ (in X) implies $Tx_n \to Tx$ (in Y).
ii) The kernel $\ker(T)$ is closed.

PROOF. i) Follows directly from the assumption that T is bounded, as

$$\|Tx_n - Tx\| = \|T(x_n - x)\| \leq \|T\|\|x - x_0\| \to 0.$$

One might also use that T is continuous from the above Theorem, and conclude using Theorem 4.2.

ii) We want to show that $\overline{\ker(T)} = \ker(T)$. Choose an arbitrary $x \in \overline{\ker(T)}$. Then there exists a sequence $(x_n) \in \ker(T)$ such that $x_n \to x$. By (i), it follows that $Tx_n \to Tx$. But $Tx_n = 0$ for all n, and thus $Tx = 0$, so we have $x \in \ker(T)$. Since $x \in \overline{\ker(T)}$ was arbitrary, $\ker(T)$ is closed.

□

Note, however, that the range of a bounded linear operator is in general not closed.

Example 4.2.6. (1) Let T be the multiplication operator $Tx = (\frac{x_n}{n})$ on ℓ^∞. Then T is a bounded linear operator. However, the range of T is not closed:

The sequence $x^{(n)} = (1, \sqrt{2}, ..., \sqrt{n}, 0, 0, ...) \in \ell^\infty$ is mapped to the sequence $y^{(n)} = (1, 1/\sqrt{2}, ..., 1/\sqrt{n}, 0, 0, 0, ...)$. Hence $y^{(n)} \in T(\ell^\infty)$ for every n. It is easily verified that $y^{(n)} \to y = (1, 1/\sqrt{2}, ..., 1/\sqrt{n}, \ldots) \in \ell^\infty$. However, $y \notin \text{ran}(T)$; The only sequence which can satisfy $T(x) = y$ is $x = (1, \sqrt{2}, ..., \sqrt{n}, ...)$, but this is not an element of ℓ^∞. Hence $\text{ran}(T)$ is not closed.

(2) We define the linear operator $V : (C[0,1], \|.\|_\infty) \to (C[0,1], \|.\|_\infty)$ by

$$Vf(x) := \int_0^x f(y)dy.$$

This is known as the *Volterra integral operator*. It is clearly bounded, as

$$\|Vf\|_\infty \leq \sup_{x \in [0,1]} \int_0^x |f(y)|dy \leq \int_0^1 |f(y)|dy \leq \|f\|_\infty.$$

The range of V is the set of continuously differentiable functions on $[0,1]$ that vanish at $x = 0$. This is a subspace of $C[0,1]$ which is not closed. (Exercise: prove this!)

Finally, let us close this section with a result on bounded linear extensions.

Definition 4.2.7. Let $M \subset X$ be a proper subset of X, and let $T : M \to Y$ be an operator. An *extension* of T to X is an operator

$$\tilde{T} : X \to Y \quad \text{such that} \quad \tilde{T}(m) = T(m) \text{ for all } m \in M.$$

A given operator T can have many extensions. Of practical interest are usually those which preserve some basic property, such as linearity or boundedness. The following important result is typical in this respect. It concerns the extension of a bounded linear operator T from a dense subset of a space to the entire space such that the extended operator is again bounded and linear, and even has the

same norm. In particular, this includes extensions from a normed space X to its completion.

Theorem 4.8. Let X be a normed space and Y be a Banach space. Suppose M is a dense subspace of X and $T : M \to Y$ is a bounded linear operator. Then T has an extension

$$\tilde{T} : X \to Y,$$

where $\tilde{T}$ is a bounded linear operator of norm

$$\|\tilde{T}\| = \|T\|.$$

Proof. Consider any $x \in X$. Since $M \subset X$ is dense, there exists a sequence $(x_n) \in M$ such that $x_n \to x$. Since T is linear and bounded, we have

$$\|Tx_n - Tx_m\|_Y = \|T(x_n - x_m)\| \leq \|T\|\|x_n - x_m\|_X.$$

This shows that the sequence (Tx_n) in Y is Cauchy, since (x_n) is convergent in X. By assumption, Y is complete (it is a Banach space), so (Tx_n) converges in Y, say

$$Tx_n \to y \in Y.$$

We now define $\tilde{T}$ by

$$\tilde{T}x = y.$$

Let us first show that this definition is independent of the particular choice of sequence converging to x. Suppose that $x_n \to x$ and $z_n \to x$. Then necessarily

$$\|x_n - z_n\|_X \leq \|x_n - x\|_X + \|x - z_n\|_X,$$

and it follows that $x_n - z_n \to 0$ as $n \to \infty$. As M is a subspace of X, we have $0 \in M$, and it thus follows from Corollary 4.7 that $T(x_n - z_n) = Tx_n - Tz_n \to T(0) = 0$ as $n \to \infty$. In other words, the definition of $\tilde{T}$ does not depend on the approximation sequence.

The map $\tilde{T}$ is linear because T is linear, and it is easily seen that $\tilde{T} = T$ on the subspace M. Finally, let us see that $\|\tilde{T}\| = \|T\|$. It is clear that $\|\tilde{T}\| \geq \|T\|$, because the norm is defined as a supremum, and thus cannot decrease for an extension. For the opposite inequality, we use

$$\|Tx_n\|_Y \leq \|T\|\|x_n\|_X,$$

and let $n \to \infty$. Then $Tx_n \to y = \tilde{T}x$. Since $x \to \|x\|$ defines a continuous mapping (recall Proposition 4.3), we thus obtain

$$\|\tilde{T}x\|_Y \leq \|T\|\|x\|_X.$$

Hence $\tilde{T}$ is bounded and $\|\tilde{T}\| \leq \|T\|$. We thus get $\|\tilde{T}\| = \|T\|$. □

4.3. Normed spaces of operators. Dual space

Let us now take any two normed spaces X and Y (either both real or both complex), and consider the set

$$B(X, Y)$$

consisting of all bounded linear operators from X into Y. In this section, we aim to show that $B(X,Y)$ can itself be made into a normed space.[1]

Note first that $B(X,Y)$ becomes a vector space if we define the sum T_1+T_2 of two operators $T_1, T_2 \in B(X,Y)$ by

$$(T_1+T_2)x = T_1x + T_2x$$

and the product αT of $T \in B(X,Y)$ and a scalar α by

$$(\alpha T)(x) = \alpha Tx.$$

Recalling Lemma 4.4 (2), we immediately get:

Theorem 4.9. The vector space $B(X,Y)$ of all bounded linear operators from a normed space X to a normed space Y is itself a normed space with norm defined by

$$\|T\| = \sup_{x\in X, x\neq 0} \frac{\|Tx\|}{\|x\|} = \sup_{x\in X, \|x\|=1} \|Tx\|.$$

When is the normed space $B(X,Y)$ a Banach space? This is a central question which is answered in the next theorem. Remarkably, the answer does not pose any conditions on the set X.

Theorem 4.10. If Y is a Banach space, then $B(X,Y)$ is a Banach space.

The curious reader might ask whether this implication goes both ways. The answer is yes if X is nontrivial; in this case, if $B(X,Y)$ is a Banach space, then so is Y. However, in this course we focus on the implication in Theorem 4.10.

Proof of Theorem 4.10. Let Y be a Banach space. We consider an arbitrary Cauchy sequence (T_n) in $B(X,Y)$, and aim to show that (T_n) converges to an operator $T \in B(X,Y)$. Fix $\varepsilon > 0$. Since (T_n) is Cauchy, we can find $N \in \mathbb{N}$ such that

$$\|T_n - T_m\| < \varepsilon \quad \text{for } m,n > N.$$

Thus for all $x \in X$ and all $m,n > N$, we have

$$\|T_nx - T_mx\| = \|(T_n - T_m)x\| \leq \|T_n - T_m\|\|x\| < \varepsilon\|x\|. \tag{4.4}$$

For each fixed $x \in X$ we can make the right hand side in this inequality as small as we wish by choosing n, m sufficiently large, so the sequence (T_nx) is Cauchy in Y. Since Y is complete, (T_nx) converges, say $T_nx \to y \in Y$. Clearly, the limit y depends on $x \in X$, so this defines an operator $T : X \to Y$, where $Tx = y$. The operator is linear, since

$$\begin{aligned} T(\alpha x + \beta z) &= \lim_{n\to\infty} T_n(\alpha x + \beta z) \\ &= \lim_{n\to\infty} (\alpha T_n x + \beta T_n z) \\ &= \alpha \lim_{n\to\infty} T_n x + \beta \lim_{n\to\infty} T_n z \\ &= \alpha Tx + \beta Tz. \end{aligned}$$

[1]The "B" in $B(X,Y)$ suggests "bounded". In section 2.1, we referred to this same set of operators as $\mathcal{L}(X,Y)$, where $\mathcal{L}$ suggests "linear". Both notations commonly appear in literature. We will stick to $B(X,Y)$ from now on.

Now let us see that T is bounded, and that $T_n \to T$, meaning $\|T_n - T\| \to 0$. Returning to (4.4), we let $m \to \infty$, and using the continuity of the norm we get

$$\begin{aligned} \|T_n x - Tx\| &= \|T_n x - \lim_{m\to\infty} T_m x\| \\ &= \lim_{m\to\infty} \|T_n x - T_m x\| \le \varepsilon \|x\| \quad \text{for } n > N. \end{aligned} \tag{4.5}$$

This shows that $(T_n - T)$ is a bounded linear operator when $n > N$. Since T_n is bounded, it follows that $T = T_n - (T_n - T)$ is also bounded, that is $T \in B(X, Y)$. Finally, if in (4.5) we take the supremum over all x with $\|x\| = 1$, we get

$$\|T_n - T\| = \sup_{\|x\|=1} \frac{\|T_n x - Tx\|}{\|x\|} \le \varepsilon, \quad n > N.$$

Hence, we have $\|T_n - T\| \to 0$. □

Let us now look at a very important special case of $B(X, Y)$, namely that where Y is the field $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ over which X is defined.

Definition 4.3.1. A *linear functional* f is a linear operator whose domain is a vector space X and where the range of f lies in the scalar field of X; thus,

$$f : X \to \mathbb{F},$$

where $\mathbb{F} = \mathbb{R}$ if X is a real vector space and $\mathbb{F} = \mathbb{C}$ if X is a complex vector space.

Definition 4.3.2. A *bounded linear functional* f is a bounded linear operator

$$f : X \to \mathbb{F},$$

where $\mathbb{F} = \mathbb{R}$ if X is a real vector space and $\mathbb{F} = \mathbb{C}$ if X is a complex vector space. Thus, there exists a real number c such that

$$|f(x)| \le c\|x\| \quad \text{for all } x \in X.$$

Moreover, the *norm* of f is

$$\|f\| = \sup_{x\in X, x\neq 0} \frac{|f(x)|}{\|x\|} = \sup_{x\in X, \|x\|=1} |f(x)|.$$

The results we have established for bounded linear operators in general also hold for the special case of bounded linear functionals. For instance, a linear functional is bounded if and only if it is continuous. Similarly, we have the inequality

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

for all bounded linear functionals on a normed space X.

Example 4.3.3. i) *Definite integral:* Consider the space $C[a, b]$ of real-valued functions on $[a, b]$, endowed with the sup-norm. Then f defined by

$$f(x) = \int_a^b x(t)\, dt, \quad x \in C[a, b],$$

is a linear functional on $C[a,b]$. Let us see that f is also bounded, with norm $\|f\| = b-a$. We have that

$$|f(x)| = \left| \int_a^b x(t)\,dt \right| \leq (b-a) \max_{0\leq t\leq 1} |x(t)| = (b-a)\|x\|,$$

and taking the supremum over all x with $\|x\| = 1$, we get $\|f\| \leq b-a$. To see that, in fact, $\|f\| = b-a$, we choose the particular function $x_0(t) = 1$; we then have $\|x_0\| = 1$ and

$$\|f\| \geq \frac{|f(x_0)|}{\|x_0\|} = |f(x_0)| = \int_a^b dt = b-a.$$

ii) *Point evaluation:* Another practically important functional on $C[a,b]$ (with the sup-norm) is obtained if we choose a fixed $t_0 \in [a,b]$, and set

$$f(x) = x(t_0), \quad x \in C[a,b].$$

This is a bounded linear functional on $C[a,b]$ with norm $\|f\| = 1$.

Exercise: Show this.

iii) *On ℓ^2:* We can obtain a linear functional f on the Hilbert space ℓ^2 by choosing a fixed sequence $a = (a_j)_{j\in\mathbb{N}} \in \ell^2$, and setting

$$f(x) = \sum_{j=1}^{\infty} \xi_j a_j, \quad x = (\xi_j)_{j\in\mathbb{N}} \in \ell^2.$$

This series converges absolutely and f is bounded, as the Cauchy-Schwarz inequality gives

$$\begin{aligned} |f(x)| = \left|\sum \xi_j a_j\right| &\leq \sum |\xi_j a_j| \\ &\leq \left(\sum |\xi_j|^2\right)^{1/2} \left(\sum |a_j|^2\right)^{1/2} = \|x\|\|a\|, \end{aligned}$$

where the summation everywhere is over j from 1 to ∞.

The space of bounded linear functionals $B(X,\mathbb{F})$ on a normed space $(X, \|\cdot\|)$ has been given a special name.

Definition 4.3.4 (Dual space)**.** Let X be a normed space. Then the set of all bounded linear functionals on X constitutes a normed space with norm defined by

$$\|f\| = \sup_{x\in X, x\neq 0} \frac{|f(x)|}{\|x\|} = \sup_{x\in X, \|x\|=1} |f(x)|.$$

It is called the *dual space* of X and is denoted by X'.

Since a linear functional on X maps X into $\mathbb{R}$ or $\mathbb{C}$ (depending on whether the vector space X is real or complex), and since $\mathbb{R}$ and $\mathbb{C}$ are both complete (with the $|\cdot|$ norm), we see that $X' = B(X,\mathbb{F})$ with the complete space $\mathbb{F}$. It thus follows immediately from Theorem 4.10 that:

Theorem 4.11. The dual space X' of a normed space X is a Banach space (whether or not X is).

CHAPTER 5

Best approximation and projection theorem

The focus of this chapter is Hilbert space theory, more specifically the projection theorem and its ramifications. We will see that the projection theorem indicates that Hilbert spaces are, in some sense, infinite-dimensional Euclidean spaces. In contrast, the structure of Banach spaces can be rich and full of strange phenomena.

5.1. Best approximations and the projection theorem

In a metric space X, the *distance* δ from an element $x \in X$ to a nonempty subset $M \subset X$ is defined as

$$\delta = \inf_{m \in M} d(x, m).$$

In a normed space, this becomes

$$\delta = \inf_{m \in M} \|x - m\|.$$

In theory as well as applications, it can be important to determine whether there exists a $y \in M$ such that

$$\delta = \|x - y\|, \tag{5.1}$$

that is, whether there exists a $y \in M$ which is closest to the given point $x \in X$. Moreover, if such an element exists, it is of interest whether or not this point is unique. This is an *existence and uniqueness problem.*

Even in very simple spaces such as $\mathbb{R}^2$ it is not difficult to construct examples of sets M and points x where either (5.1) is satisfied for no $y \in M$, or (5.1) is satisfied for infinitely many $y \in M$ (*Exercise: Try to construct examples of both cases*). However, our first result shows that if we restrict our attention to proper closed subspaces of a Hilbert space, then both existence and uniqueness is guaranteed.

Theorem 5.1 (Best Approximation Theorem)**.** Suppose M is a closed subspace of a Hilbert space X. Then for any $x \in X$ there exists a unique element $z \in M$ such that

$$\|x - z\| = \inf_{m \in M} \|x - m\|.$$

Remark 5.1.1. Note that in general, Theorem 5.1 is not true in Banach spaces. Consider the Banach space ℓ^∞ and its closed subspace c_0 (the space of sequences converging to 0). For the point $x = (1, 1, 1, \ldots) \in \ell^\infty$, we have

$$\inf_{y \in c_0} \|x - y\|_\infty = 1,$$

and this infimum is indeed attained, for instance by the sequence $z = (0, 0, 0, \ldots) \in c_0$. However, z is by no means unique. We also have

$$\|x - z'\|_\infty = \inf_{y \in c_0} \|x - y\|_\infty = 1$$

for $z' = (1, 0, 0, 0, \ldots)$ or $z' = (1, 1, 0, 0, 0, \ldots)$.

PROOF OF THEOREM 5.1. *Existence:* Denote by $\delta^2 = \inf_{m \in M} \|x - m\|^2$. By the definition of an infimum there exists a sequence (m_k) in M such that for each $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ such that

$$\|x - m_k\|^2 \leq \delta^2 + \varepsilon \quad \text{for all } k \geq N.$$

We show now that the sequence (m_k) is Cauchy. Applying the parallelogram identity to $x - m_k$ and $x - m_l$, we get

$$\begin{aligned} 2\left(\|x - m_k\|^2 + \|x - m_l\|^2\right) &= \|2x - m_k - m_l\|^2 + \|m_k - m_l\|^2 \\ &= 4\left\|x - \frac{m_k + m_l}{2}\right\|^2 + \|m_k - m_l\|^2. \end{aligned}$$

Since M is a vector subspace, we have $(m_k + m_l)/2 \in M$, and it follows that $\|x - (m_l + m_k)/2\| \geq \delta$. We thus get

$$\|m_k - m_l\|^2 \leq 2\left(\|x - m_k\|^2 + \|x - m_l\|^2\right) - 4\delta^2$$

Finally, using that $\|x - m_k\|^2 \leq \delta^2 + \varepsilon/4$ for all sufficiently large k, we get

$$\|m_k - m_l\|^2 \leq 2\big(\delta^2 + \frac{\varepsilon}{4} + \delta^2 + \frac{\varepsilon}{4}\big) - 4\delta^2 = \varepsilon,$$

for all sufficiently large l, k. Hence, (m_k) is a Cauchy sequence. Since M is closed, we recall from Theorem 3.14 that $(M, \|\cdot\|)$ is complete, so m_k converges to some element $z \in M$. Since the norm is continuous we have $\|x - z\| = \lim_{k\to\infty} \|x - m_k\| = \delta$.

Uniqueness: We assume that $z \in M$ and $z_0 \in M$ both satisfy

$$\|x - z\| = \delta \quad \text{and} \quad \|x - z_0\| = \delta,$$

and show that then $z = z_0$. By the parallelogram identity, we have

$$\begin{aligned} \|z - z_0\|^2 &= \|(z - x) - (z_0 - x)\|^2 \\ &= 2\|z - x\|^2 + 2\|z_0 - x\|^2 - \|(z - x) + (z_0 - x)\|^2 \\ &= 2\delta^2 + 2\delta^2 - 4\left\|\frac{z + z_0}{2} - x\right\|^2. \end{aligned}$$

Again, we have that $(z - z_0)/2 \in M$, so $\|(z - z_0)/2 - x\| \geq \delta$. Inserting this above, we get

$$\|z - z_0\| \leq 0.$$

Clearly, $\|z - z_0\| \geq 0$, so we must have equality, and $z = z_0$. □

Let us now see that the familiar idea of elementary geometry that the unique point in a given subspace M closest to a given point x is found by "dropping a perpendicular from x to M" generalizes to the Hilbert space setting.

Lemma 5.2. Suppose M is a closed subspace of a Hilbert space X, and let z be the unique point in M at minimal distance from x, that is

$$\|x - z\| = \inf_{m \in M} \|x - m\|.$$

Then $x - z$ is orthogonal to M.

PROOF. Suppose z minimizes $\|x - m\|$, but $x - z$ is *not* orthogonal to M, meaning there exists an element $m \in M$ with norm $\|m\| = 1$ such that $\langle x - z, m\rangle = d \neq 0$. We will see that this leads to a contradiction.

Consider the element $y = z + dm$, which is clearly an element in M since this is a vector subspace. We have

$$\begin{aligned}
\|x - y\|^2 &= \|x - z - dm\|^2 \\
&= \langle x - z - dm, x - z - dm\rangle \\
&= \langle x - z, x - z\rangle - \langle x - z, dm\rangle - \langle dm, x - z\rangle + \langle dm, dm\rangle \\
&= \|x - z\|^2 - \overline{d}d - d\overline{d} + d\overline{d}\|m\| \\
&= \|x - z\|^2 - |d|^2.
\end{aligned}$$

This indicates that $\|x - y\| \leq \|x - z\|$, contradicting the fact that z minimizes $\|x - m\|$. We conclude that $\langle x - z, m\rangle = 0$ for all $m \in M$, and thus $x - z$ is orthogonal to M. □

Example 5.1.2. Let X be the space of square-integrable real-valued functions $L^2[0, 1]$ and the subspace $\mathcal{P}_1$ is the space of polynomials of degree at most 1, $M = \{a_1 x + a_0 : a_0, a_1 \in \mathbb{R}\}$. We look for the best approximation of $f(x) = x^2$ from M. Hence we want to find the polynomial $p \in \mathcal{P}_1$ such that

$$\|f - p\|_2 = \Big(\int_0^1 |x^2 - p(x)|^2\, dx\Big)^{1/2} = \inf_{a_0, a_1 \in \mathbb{R}} \Big(\int_0^1 |x^2 - a_0 - a_1 x|^2\, dx\Big)^{1/2}.$$

We determine p via the orthogonality of the error vector/residual $f - p$ onto $\mathcal{P}_1$. Since $\mathcal{P}_1$ is spanned by $b_0(x) = 1$ and $b_1(x) = x$ this amounts to $f - p \perp b_0$ and $f - p \perp b_1$, i.e.

$$\begin{aligned}
\int_0^1 (x^2 - a_0 - a_1 x) 1\, dx &= 0 \\
\int_0^1 (x^2 - a_0 - a_1 x) x\, dx &= 0
\end{aligned}$$

which gives the following system of equations:

$$\begin{aligned}
\frac{1}{3} - a_0 - \frac{a_1}{2} &= 0 \\
\frac{1}{4} - \frac{a_0}{2} - \frac{a_1}{3} &= 0,
\end{aligned}$$

which has the unique solution $a_0 = -1/6$ and $a_1 = 1$, i.e. the best approximation is attained by $p(x) = x - 1/6$.

Our current goal is to establish the projection theorem, which states that any Hilbert space can be represented as a particularly simple direct sum if we make use of orthogonality. The concept *direct sum* is introduced below, and makes sense for any vector space.

Definition 5.1.3. A vector space X is said to be the direct sum of two subspaces Y and Z of X, written

$$X = Y \oplus Z,$$

if each $x \in X$ has a unique representation

$$x = y + z, \quad y \in Y, z \in Z.$$

Let us pause for a moment and compare the direct sum and the sum of two subspaces Y and Z of X. The sum $Y + Z = \{y + z | \, y \in Y \text{ and } z \in Z\}$ and we say that X is the sum of Y and Z if $X = Y + Z$, i.e. every $x \in X$ may be decomposed in the form $x = y + z$ for some vectors $y \in Y$ and $z \in Z$.

Example 5.1.4. Let $Y = \{(x_1, x_2, 0) : x_i \in \mathbb{R} \text{ for } i = 1, 2\}$ and $Z = \{(0, x_2, x_3) : x_i \in \mathbb{R} \text{ for } i = 2, 3\}$ be two subspaces of $\mathbb{R}^3$. We have that $X + Y = \mathbb{R}^3$. However, this is not a direct sum since vectors of the form $(0, x, 0)$ have non-unique decompositions, e.g. $(0, x, 0) = (0, 0, 0) + (0, x, 0)$ for $(0, 0, 0) \in Y$, $(0, x, 0) \in Z$ and $(0, x, 0) = (0, x, 0) + (0, 0, 0)$ for $(0, x, 0) \in Y$, $(0, 0, 0) \in Z$.

Observe that in the example $Y \cap Z \neq \{0\}$ and the next result shows that this is the reason for not giving a direct sum decomposition.

Lemma 5.3. Let Y and Z be subspaces of X. Then $X = Y \oplus Z$ if and only if the following two conditions hold:

(i) $X = Y + Z$
(ii) $Y \cap Z = \{0\}$

Proof. ($\Rightarrow$) Suppose $X = Y \oplus Y$. Then Condition (i) holds by the definition of direct sum. Let x be a vector in $\in Y \cap Z$. Then $-x \in Y \cap Z$ since $Y \cap Z$ is a vector space. Hence we have that $0 = x + (-x)$ but we also have $0 = 0 + 0$. Since $X = Y \oplus Z$ the decomposition $0 = 0 + 0$ is unique and thus $x = 0$. Equivalently, $Y \cap Z = \{0\}$.
($\Leftarrow$) Suppose Condition (i) and (ii) hold. Then we want to show that vectors in X have a unique decomposition. Assume that 0 has a non-trivial decomposition $0 = y + z$ for $y \in Y$ and $z \in Z$. Hence we have that $y = -z$ and thus $y \in Z$, i.e. $y \in Y \cap Z$. By Condition (ii) we have $y = 0$ and thus $z = 0$. □

Example 5.1.5.

(1) Take for $X = C(\mathbb{R})$, the space of real-valued continuous functions on $\mathbb{R}$. The subspace Y is the space of all even functions, $Y = \{f \in C(R) : f(x) = f(-x) \text{ for all } x \in \mathbb{R}\}$ and Z is the space of all odd functions, $Z = \{f \in C(R) : f(x) = -f(-x) \text{ for all } x \in \mathbb{R}\}$. Since $Y \cap Z = \{0\}$ we have $C(\mathbb{R}) = Y \oplus Z$ and we have the unique decomposition $f(x) = \frac{f(x)+f(-x)}{2} + \frac{f(x)-f(-x)}{2}$.

(2) Let X be the space of real $n \times n$ matrices $M_n(\mathbb{R})$. We define the subspaces $Y = \{A \in M_n(\mathbb{R}) : A^T = A\}$ and $Z = \{A \in M_n(\mathbb{R}) : A^T = -A\}$ of symmetric and skew-symmetric matrices. Then $M_n(R) = Y \oplus Z$ since $Y \cap Z = \{0\}$, and we have the unique decomposition: $A = \frac{A+A^T}{2} + \frac{A-A^T}{2}$.

Inspired by the notation used in $\mathbb{R}^n$, we will write $x \perp y$ when x, y are orthogonal elements of some Hilbert space. Hence $x \perp y$ means that $\langle x, y \rangle = 0$.

Definition 5.1.6 (Orthogonal complement)**.** Let M be a closed subspace of a Hilbert space X. Then the *orthogonal complement* $M^\perp$ of M is the set of all vectors orthogonal to M, that is

$$M^\perp = \{x \in X \,:\, x \perp y \text{ for any } y \in M\}\,.$$

The orthogonal complement $M^\perp$ is a vector subspace, since for all $x, y \in M^\perp$ and scalars α, β, we have

$$\langle \alpha x + \beta y, m\rangle = \alpha \langle x, m\rangle + \beta \langle y, m\rangle = 0 \quad \text{ for all } m \in M,$$

and thus $\alpha x + \beta y \in M^\perp$. It is also easy to see that $M \cap M^\perp = \{0\}$, since any $x \in M \cap M^\perp$ must satisfy $\|x\|^2 = \langle x, x\rangle = 0$, hence $x = 0$. Moreover, it is always closed:

Lemma 5.4. Let M be a subspace of an inner product space $(X, \langle\cdot,\cdot\rangle)$. Then $M^\perp$ is closed.

Proof. *Exercise. Hint: Use that for a fixed $y \in X$ the inner product is continuous in the first entry $x \mapsto \langle x, y\rangle$ is continuous from X to $\mathbb{F}$.* □

We are now equipped to state the projection theorem. An explanation of why this name is fitting is given after the proof.

Theorem 5.5 (Projection Theorem)**.** Let M be any closed subspace of a Hilbert space X. Then

$$X = M \oplus M^\perp.$$

In other words, every $x \in X$ has a unique representation

$$x = y + z, \quad y \in M, z \in M^\perp. \tag{5.2}$$

Proof. By Theorem 5.1 there exists a best approximation $y \in M$ of $x \in X$, and by Lemma 5.2 we have $z = x - y \in M^\perp$. It is thus clear that

$$x = y + (x - y) = y + z, \quad y \in M, z \in M^\perp.$$

To prove uniqueness, we assume that

$$x = y + z = y_1 + z_1,$$

where $y, y_1 \in M$ and $z, z_1 \in M^\perp$. Then $y - y_1 = z_1 - z$. Since $y - y_1 \in M$ and $z_1 - z \in M^\perp$, we see that $y - y_1 \in M \cap M^\perp = \{0\}$. This implies $y = y_1$, and likewise $z = z_1$. □

The element y in (5.2) is called the *orthogonal projection* (or just *projection*) of x on M. In fact, we see that (5.2) defines a mapping

$$P : X \to M \quad \text{ by } \quad x \mapsto y = Px.$$

The map P is called the **projection** of X onto M. It is clear that P is a bounded linear operator which maps X onto M, M onto itself (meaning $P|_M$ is the identity map), and $M^\perp$ onto $\{0\}$. Moreover, P is *idempotent*, meaning

$$P^2 = P,$$

or, for every $x \in X$, we have $P^2 x = P(Px) = Px$.

Let us now look at some consequences of the projection theorem.

Lemma 5.6. If M is a closed subspace of a Hilbert space X, then
$$M = M^{\perp\perp}.$$

PROOF. In general, we have $M \subseteq M^{\perp\perp}$, because
$$x \in M \quad \Rightarrow \quad x \perp M^{\perp} \quad \Rightarrow \quad x \in (M^{\perp})^{\perp}.$$
Let us now see that for closed subspaces, we also have $M^{\perp\perp} \subseteq M$.

Let $x \in M^{\perp\perp}$. Then by the projection theorem, we have
$$x = y + z, \quad y \in M, z \in M^{\perp}.$$
Note that $y \in M \subseteq M^{\perp\perp}$. Since $M^{\perp\perp}$ is a vector space, and $x \in M^{\perp\perp}$ by assumption, we also have $z = x - y \in M^{\perp\perp}$, so $z \perp M^{\perp}$. Combined with the fact that $z \in M^{\perp}$, this means $z \perp z$, or equivalently $\langle z, z\rangle = 0$, so $z = 0$. We thus have $x = y \in M$. This shows $M^{\perp\perp} \subseteq M$, as x was arbitrary. □

The projection theorem also provides a characterization of sets in Hilbert spaces whose span is dense, as follows. Recall that the span of a subset M is the set of all linear combinations of vectors in M.

Lemma 5.7. For any subset $M \neq \emptyset$ of a Hilbert space X, the span of M is dense in X if and only if $M^{\perp} = \{0\}$.

PROOF. Let $x \in M^{\perp}$, and assume $V = \text{span}M$ is dense in X. We will show that $x = 0$.

Since $\overline{V} = X$, we have $x \in \overline{V}$, so there exists a sequence (x_n) in V such that $x_n \to x$. Since $x \in M^{\perp}$ and $M^{\perp} \perp V$, we have $\langle x_n, x\rangle = 0$ for all n. On the other hand, using the linearity of the inner product and Cauchy-Schwarz inequality, we get
$$|\,\langle x_n, x\rangle - \langle x, x\rangle\,| = |\,\langle x_n - x, x\rangle\,| \leq \|x_n - x\|\|x\|,$$
so if $x_n \to x$ then $\langle x_n, x\rangle \to \langle x, x\rangle$. It follows that $\langle x, x\rangle = \|x\|^2 = 0$, so $x = 0$.

For the other direction, suppose that $M^{\perp} = \{0\}$. It then follows that $V^{\perp} = \{0\}$, since
$$x \perp V \quad \Rightarrow \quad x \perp M \quad \Rightarrow \quad x \in M^{\perp} \quad \Rightarrow \quad x = 0.$$
Now observe that $\overline{V}^{\perp} \subseteq V^{\perp} = \{0\}$, and applying the projection theorem with the closed subspace $\overline{V} \subseteq X$, we get
$$X = \overline{V} \oplus \overline{V}^{\perp} = \overline{V} \oplus \{0\} = \overline{V}.$$
This shows that V is dense in X. □

Finally, let us look at a "practical" application of the projection theorem.

Example 5.1.7. Recall that $L^2[-1, 1]$, the space of square integrable functions on the interval $[-1, 1]$, is the completion of the normed space $(C[-1, 1], \|\cdot\|_2)$. When endowed with the inner product
$$\langle f, g\rangle = \int_{-1}^{1} f(x)\overline{g(x)}\,dx,$$
this is a Hilbert space.

Let $M_e \subset L^2[-1,1]$ be the subspace of even functions, that is

$$M_e = \left\{ f \in L^2[-1,1] \,:\, f(-x) = f(x) \right\}.$$

This is a closed subspace (*exercise: show this*). Thus, according to the projection theorem, we can write any function $f \in L^2[-1,1]$ uniquely as a sum

$$f = f_1 + f_2, \quad f_1 \in M_e, \quad f_2 \in M_e^\perp.$$

Let us now see that $M_e^\perp$ is in fact the closed subspace of odd functions, that is

$$M_e^\perp = M_o = \left\{ f \in L^2[-1,1] \,:\, f(x) = -f(-x) \right\}.$$

If $f \in M_e$ and $g \in M_o$, then

$$\langle f, g \rangle = \int_{-1}^{1} f(x)\overline{g(x)}\, dx = 0,$$

since the integrand $f\overline{g}$ is an odd function and the interval of integration is symmetric about the origin. This shows that $M_o \subseteq M_e^\perp$. To see that $M_e^\perp \subseteq M_o$, we observe that any function $f \in L^2[-1,1]$ can indeed be written as a combination of an odd and an even function, by simply putting

$$f(x) = \frac{f(x) + f(-x)}{2} + \frac{f(x) - f(-x)}{2} = f_e(x) + f_o(x). \tag{5.3}$$

Thus, for any $f \in M_e^\perp$, we have

$$0 = \langle f, f_e \rangle = \langle f_e + f_o, f_e \rangle = \langle f_e, f_e \rangle = \|f_e\|^2 = 0.$$

This implies $f_e = 0$ and $f = f_o \in M_o$, so $M_e^\perp \subseteq M_o$.

Returning to the projection theorem, we see that the representation (5.3) of f as a sum of an odd function f_o and an even function f_e is in fact unique. Moreover, it is clear from the argument above that the associated projection $P : L^2[-1,1] \to M_e$ is given by

$$Pf(x) = \frac{f(x) + f(-x)}{2}.$$

5.2. Riesz' representation theorem

In this section we will state and prove Riesz' representation theorem. This result characterizes the dual space of bounded linear functionals on a Hilbert space.

Theorem 5.8 (Riesz' representation theorem)**.** Let X be a Hilbert space. For each $z \in X$ define $\varphi_z(x) = \langle x, z \rangle$. Then $\varphi_z \in X'$ is a bounded linear functional on X.

On the other hand, every $\varphi \in X'$ is given by an inner product

$$\varphi(x) = \varphi_z(x) = \langle x, z \rangle$$

for some unique $z \in X$, and $\|\varphi\| = \|z\|$.

In other words, a Hilbert space is its own dual. Note that the first part of the statement is easily verified using the Cauchy-Schwarz inequality. The final assertion is the subtle part of the theorem, and we prove only this.

PROOF. *Existence:* Let $M = \ker \varphi$. This is a closed linear subspace of X. If $M = X$, then $\varphi = 0$ (the zero operator), and

$$\varphi(x) = \varphi_0(x) = \langle x, 0 \rangle .$$

Now assume $M \neq X$, so M is a proper closed subspace of X. Then by the Projection Theorem there exists a non-zero element $z_0 \in M^\perp$, $z_0 \neq 0$. Since $z_0 \perp \ker \varphi$, we have $\varphi(z_0) \neq 0$, and by the linearity of φ we see that

$$x - \frac{\varphi(x)}{\varphi(z_0)} z_0 \in \ker \varphi \quad \text{for all } x \in X.$$

As $z_0 \perp \ker \varphi$ it follows that

$$\begin{aligned} \left\langle x - \frac{\varphi(x)}{\varphi(z_0)} z_0, z_0 \right\rangle = 0 \quad &\Rightarrow \quad \langle x, z_0 \rangle = \frac{\varphi(x)}{\varphi(z_0)} \|z_0\|^2 \\ &\Rightarrow \quad \varphi(x) = \frac{\varphi(z_0)}{\|z_0\|^2} \langle x, z_0 \rangle = \left\langle x, \frac{\overline{\varphi(z_0)}}{\|z_0\|^2} z_0 \right\rangle . \end{aligned}$$

Thus, we have for any $x \in X$ that

$$\varphi(x) = \langle x, z \rangle \quad \text{for} \quad z = \frac{\overline{\varphi(z_0)}}{\|z_0\|^2} z_0.$$

Uniqueness: Suppose there exist two elements $z, w \in X$ such that

$$\varphi(x) = \langle x, z \rangle = \langle x, w \rangle \quad \text{for all } x \in X.$$

Then

$$\langle x, z - w \rangle = \langle x, z \rangle - \langle x, w \rangle = 0 \quad \text{for all } x \in X.$$

In particular, this holds for $x = z - w$, and it follows that

$$\|z - w\|^2 = 0 \quad \Rightarrow \quad z = w.$$

Equality of norms: We have

$$\|\varphi\| = \sup_{\|x=1\|} |\varphi(x)| = \sup_{\|x\|=1} |\langle x, z \rangle| \leq \sup_{\|x\|=1} \|x\| \|z\| = \|z\|,$$

where we have used Cauchy-Schwarz for the final inequality. On the other hand, we have

$$\|z\|^2 = \langle z, z \rangle = |\varphi(z)| \leq \|\varphi\| \|z\| \quad \Rightarrow \quad \|\varphi\| \geq \|z\|.$$

We thus get $\|\varphi\| = \|z\|$. □

Remark 5.2.1. In the proof of uniqueness above, we observed that if $\langle x, z \rangle = \langle x, w \rangle$ for all $x \in X$, then $z = w$. As this is a fact we will use repeatedly in the following subsection, we write it out as a separate result:

Proposition 5.9. Let X be a Hilbert space, and let $z, w \in X$. If

$$\langle x, z \rangle = \langle x, w \rangle \quad \text{for all } x \in X,$$

then $z = w$.

Let us now apply Riesz' representation theorem to some familiar Hilbert spaces.

Example 5.2.2. i) Every bounded linear functional φ on $\mathbb{C}^n$ is realized by a dot product

$$\varphi(x) = \langle x, y\rangle = x \cdot \overline{y}, \quad \text{for some fixed } y \in \mathbb{C}^n.$$

ii) Every bounded linear functional φ on $L^2[a,b]$ is realized by an inner product

$$\varphi(f) = \int_a^b f(x)\overline{g(x)}\,dx \quad \text{for some fixed } g \in L^2[a,b].$$

Moreover, observe that by the Cauchy-Schwarz inequality, we have

$$|\varphi(f)| = |\langle f, g\rangle| = \left|\int_a^b f(x)\overline{g(x)}\,dx\right| \le \left(\int_a^b |f(x)|^2\right)^{1/2}\left(\int_a^b |g(x)|^2\right)^{1/2} = \|f\|\|g\|,$$

with equality for $f = \lambda g$, so $\|\varphi\| = \|g\|$.

iii) Every bounded linear functional φ on ℓ^2 is realized by an inner product

$$\varphi(x) = \langle x, a\rangle = \sum_{j=1}^{\infty} x_j\overline{a_j}, \quad \text{for some fixed } a = (a_j)_{j\in N} \in \ell^2.$$

5.3. Adjoint operators

Consider a Hilbert space X over a field $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. In this section we introduce *adjoint operators*, which provide us with an alternative description of bounded linear operators on X. We will see that the existence of so-called adjoints is guaranteed by Riesz' representation theorem.

Theorem 5.10 (Adjoint operator)**.** Let $T \in B(X,X)$ be a bounded linear operator on a Hilbert space X. There exists a unique operator $T^* \in B(X,X)$ such that

$$\langle Tx, y\rangle = \langle x, T^*y\rangle \quad \text{for all } x, y \in X.$$

The operator T^* is called the *adjoint* of T.

Proof. *Existence:* Fix $y \in X$, and define the map

$$\varphi(x) = \langle Tx, y\rangle, \quad x \in X.$$

This is a bounded linear functional on X, as it is easily seen to be linear and

$$|\varphi(x)| = |\langle Tx, y\rangle| \le \|Tx\|\|y\| \le \|T\|\|y\|\|x\|.$$

By Riesz' representation theorem it follows that there exists a unique element $y^* \in X$ such that

$$\varphi(x) = \langle Tx, y\rangle = \langle x, y^*\rangle \quad \text{for all } x \in X.$$

We thus define $T^*y := y^*$, so by definition T^* satisfies $\langle Tx, y\rangle = \langle x, T^*y\rangle$. It remains to show that T^* is linear, bounded and unique.

Linearity of T^*: We have

$$\begin{aligned}\langle x, T^*(\alpha y_1 + \beta y_2)\rangle &= \langle Tx, \alpha y_1 + \beta y_2\rangle \\ &= \overline{\alpha}\langle Tx, y_1\rangle + \overline{\beta}\langle Tx, y_2\rangle \\ &= \overline{\alpha}\langle x, T^* y_1\rangle + \overline{\beta}\langle x, T^* y_2\rangle \\ &= \langle x, \alpha T^* y_1\rangle + \langle \beta T^* y_2\rangle \quad \text{for all } x \in X,\end{aligned}$$

and it follows by Proposition 5.9 that

$$T^*(\alpha y_1 + \beta y_2) = \alpha T^* y_1 + \beta T^* y_2.$$

Boundedness of T^*: By the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\|T^* y\|^2 = \langle T^* y, T^* y\rangle &= \langle TT^* y, y\rangle \\ &\leq \|TT^* y\|\|y\| \\ &\leq \|T\|\|T^* y\|\|y\|.\end{aligned}$$

If $\|T^* y\| > 0$, we divide by $\|T^* y\|$ on both sides in the inequality and obtain

$$\|T^* y\| \leq \|T\|\|y\|.$$

This inequality is clearly also satisfied when $\|T^* y\| = 0$, so T^* is a bounded operator. Moreover, we have attained the additional information that

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

Uniqueness: Suppose there exists another operator $S \in B(X, X)$ such that

$$\langle x, Sy\rangle = \langle x, T^* y\rangle \quad \text{for all } x, y \in X.$$

Then necessarily, for each $y \in X$, we have

$$\langle x, Sy - T^* y\rangle = 0 \quad \text{for all } x \in X,$$

so by Proposition 5.9 we get that $Sy = T^* y$ for every $y \in X$, meaning $S = T^*$. □

We list and prove some general properties of adjoints which are frequently used in applying these operators.

Proposition 5.11. Let X be a Hilbert space, $S : X \to X$ and $T : X \to X$ be bounded linear operators and $\alpha, \beta \in \mathbb{F}$ any two scalars. We then have:

i) $(\alpha S + \beta T)^* = \overline{\alpha} S^* + \overline{\beta} T^*$

ii) $(ST)^* = T^* S^*$

iii) $(T^*)^* = T$

iv) $\|T^*\| = \|T\|$

v) $\|TT^*\| = \|T^* T\| = \|T\|^2$

Proof. *i*) and *ii*): *Exercise.*

iii) Fix any $y \in X$. We have

$$\begin{aligned}\langle x, T^{**}y\rangle &= \langle T^*x, y\rangle = \overline{\langle y, T^*x\rangle} \\ &= \overline{\langle Ty, x\rangle} = \langle x, Ty\rangle\end{aligned}$$

for all $x \in X$. It thus follows from Proposition 5.9 that $T^{**} = T$.

iv) In the proof of the existence of the adjoint, we established that $\|T^*\| \leq \|T\|$. For the opposite inequality, simply observe that by *iii*), we have

$$\|T\| = \|T^{**}\| \leq \|T^*\|,$$

and thus $\|T^*\| = \|T\|$.

v) For any $x \in X$, we have

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

and accordingly

$$\|T^*T\| \leq \|T^*\|\|T\| = \|T\|^2.$$

On the other hand, using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\|Tx\|^2 = \langle Tx, Tx\rangle &= \langle T^*Tx, x\rangle \\ &\leq \|T^*Tx\|\|x\| \leq \|T^*T\|\|x\|^2,\end{aligned}$$

and it follows that $\|T\| \leq \|T^*T\|^{1/2}$, or equivalently $\|T\|^2 \leq \|T^*T\|$. We conclude that

$$\|T^*T\| = \|T\|^2.$$

Finally, replacing T by T^* in the equality above, and recalling that $T^{**} = T$, we also get

$$\|TT^*\| = \|T^*\|^2 = \|T\|^2.$$

□

Example 5.3.1. i) *Left and right shift operators:* Consider the right shift operator R on ℓ^2, given by

$$Rx = (0, x_1, x_2, x_3, \ldots), \quad x = (x_j)_{j\in\mathbb{N}} \in \ell^2.$$

Its adjoint is the left shift operator L, given by

$$Lx = (x_2, x_3, x_4, \ldots).$$

To see this, observe that

$$\begin{aligned}\langle Rx, y\rangle &= \langle (0, x_1, x_2, x_3, \ldots), (y_1, y_2, y_3, \ldots)\rangle \\ &= x_1\overline{y_2} + x_2\overline{y_3} + x_3\overline{y_4} + \ldots \\ &= \langle (x_1, x_2, x_3, \ldots), (y_2, y_3, y_4, \ldots)\rangle = \langle x, Ly\rangle\end{aligned}$$

for any $x, y \in \ell^2$. Thus, the operator R^* satisfying $\langle Rx, y\rangle = \langle x, R^*y\rangle$ for all $x, y \in \ell^2$ is $R^* = L$.

ii) *Multiplication operator on ℓ^2:* Consider the multiplication operator $T_a : \ell^2 \to \ell^2$ given by

$$T_a x = (a_j x_j)_{j\in\mathbb{N}}, \quad x = (x_j)_{j\in\mathbb{N}} \in \ell^2,$$

for some fixed $a \in \ell^\infty$. The adjoint of T_a is the multiplication operator for the conjugate sequence $\overline{a}$, that is

$$T_a^* = T_{\overline{a}}.$$

Exercise: Confirm this.

iii) *Multiplication operator on $L^2[0,1]$:* Consider the multiplication operator $T_a : L^2[0,1] \to L^2[0,1]$ given by

$$T_a f = af, \quad f \in L^2[0,1],$$

for some fixed function $a \in C[0,1]$. Its adjoint is the multiplication operator given by the conjugate function $\overline{a}$, that is $T_a^* = T_{\overline{a}}$. To see this, observe that

$$\langle T_a f, g\rangle = \int_0^1 a(t)f(t)\overline{g(t)}\,dt = \int_0^1 f(t)\overline{\overline{a(t)}g(t)}\,dt = \langle f, \overline{a}g\rangle = \langle f, T_{\overline{a}}g\rangle\,.$$

iv) *Matrices:* Consider $\mathbb{C}^n$ with the standard inner product

$$\langle x, y\rangle = x_1\overline{y_1} + \ldots + x_n\overline{y_n} = x^\top \overline{y},$$

and let $T : \mathbb{C}^n \to \mathbb{C}^n$ be the linear map given by matrix multiplication

$$Tx = Ax, \quad x \in \mathbb{C}^n,$$

for some fixed, $n \times n$ matrix A. Then the adjoint T^* of T is given by

$$T^*x = \overline{A}^\top x, \quad x \in \mathbb{C}^n.$$

To see this, observe that

$$\begin{aligned}
\langle Tx, y\rangle &= \left\langle (\sum_{j=1}^n a_{ij}x_j)_i, (y_i)_i \right\rangle \\
&= \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_j\overline{y_i} \\
&= \sum_{j=1}^n x_j \overline{\sum_{i=1}^n \overline{a_{ij}}y_i} \\
&= \sum_{j=1}^n x_j \overline{\sum_{i=1}^n \overline{a_{ji}}y_i} = \left\langle x, \overline{A}^\top y\right\rangle.
\end{aligned}$$

v) *Integral operators on $L^2[0,1]$:* For any $k \in C([0,1]\times[0,1])$ we have that the integral operator $T_k f(x) = \int_0^1 k(x,y)f*(y)dy$ is a bounded operator on $L^2[0,1]$. Hence T_k^* exists and equals $T_k^* f(x) = \int_0^1 \overline{k(y,x)}f(y)dy$. *Exercise: Confirm this assertion.*

Certain classes of bounded linear operators of great practical importance can be defined by the use of adjoint operators as follows.

Definition 5.3.2. A bounded linear operator $T : X \to X$ on a Hilbert space X is said to be

i) *normal* if $T^*T = TT^*$.

ii) *unitary* if T is bijective and $T^* = T^{-1}$. We then have
$$T^*T = TT^* = I.$$

iii) *self-adjoint* or *Hermitian* if $T = T^*$.

Example 5.3.3. i) *Multiplication operator on ℓ^2:* Recall the multiplication operator T_a on ℓ^2, defined for some fixed $a \in \ell^\infty$ by
$$T_a x = (a_j x_j)_{j\in\mathbb{N}}, \quad x \in \ell^2.$$
This is a normal operator, since it follows from $T_a^* = T_{\overline{a}}$ that
$$T_a^* T_a = T_a T_a^* = T_{|a|^2}.$$
We see that it is a unitary operator if and only if
$$|a| = (|a_1|, |a_2|, |a_3|, \ldots) = (1, 1, 1, \ldots).$$
For instance, T_a is unitary if
$$a = (1, i, -1, -i, \ldots) = (i^k)_{k=0}^\infty.$$
Moreover, we see that T_a is self-adjoint if and only if a is real-valued, since
$$T_a^* = T_{\overline{a}} = T_a$$
only in this case.

ii) *Shift operator on ℓ^2:* The right shift operator R on ℓ^2 is *not* normal. To see this, observe that
$$R^*Rx = LRx = L(0, x_1, x_2, x_3, \ldots) = (x_1, x_2, x_3, \ldots) = Ix,$$
but
$$RR^*x = RLx = R(x_2, x_3, x_4, \ldots) = (0, x_2, x_3, x_4, \ldots) \neq Ix.$$

Example 5.3.4. Consider $\mathbb{R}^n$ with the standard inner product
$$\langle x, y\rangle = x_1y_1 + \ldots + x_ny_n = x^\top y,$$
and let $T : \mathbb{R}^n \to \mathbb{R}^n$ be the linear map given by matrix multiplication
$$Tx = Ax, \quad x \in \mathbb{R}^n,$$
for some real-valued fixed, $n \times n$ matrix A. Then following Example 5.3.1iv), the adjoint T^* of T is given by
$$T^*x = A^\top x, \quad x \in \mathbb{R}^n.$$
Consequently, we see that the matrix A is

- Symmetric, meaning $A^T = A$, if T is self-adjoint.
- Invertible and orthogonal, meaning $A^T = A^{-1}$, if T is unitary.

We list certain properties of unitary operators.

Lemma 5.12. Let S and T be two unitary operators on a Hilbert space X. We then have:

i) S is isometric; thus $\|Sx\| = \|x\|$ for all $x \in X$.

ii) $\|S\| = 1$, provided $X \neq \{0\}$.

iii) The composition operators ST and TS are unitary.

iv) The identity operator I is unitary.

PROOF. i) We observe that

$$\|Sx\|^2 = \langle Sx, Sx\rangle = \langle x, S^*Sx\rangle = \langle x, Ix\rangle = \|x\|^2.$$

ii) This follows immediately from i).

iii) We have

$$(ST)^*(ST) = T^*S^*ST = T^*IT = T^*T = I,$$

and by an equivalent calculation one can verify that $(ST)(ST)^* = I$.

iv) It is clear that $I^* = I$ (i.e. the identity operator is also self-adjoint), since

$$\langle Ix, y\rangle = \langle x, y\rangle = \langle x, Iy\rangle\,, \quad \text{for all } x, y \in X.$$

It immediately follows that $I^*I = II^* = I$.

□

We close our discussion of adjoint operators with certain useful relations between the kernel and range of an operator and its adjoint.

Proposition 5.13. Let T be a bounded linear operator on a Hilbert space X. We then have

i) $\overline{\text{ran}(T)} = \ker(T^*)^\perp$;

ii) $\ker(T) = \text{ran}(T^*)^\perp$.

Equivalently, we have

$$\overline{\text{ran}(T^*)} = \ker(T)^\perp \quad \text{and} \quad \ker(T^*) = \text{ran}(T)^\perp,$$

and consequently

$$X = \ker T \oplus \overline{\text{ran}(T^*)} = \ker T^* \oplus \overline{\text{ran}(T)}.$$

PROOF. i) *Showing* $\overline{\text{ran}(T)} \subseteq \ker(T^*)^\perp$*:*

Let $y \in \text{ran}(T)$. Then $y = Tx$ for some $x \in X$, and for any $z \in \ker(T^*)$, we get

$$\langle y, z\rangle = \langle Tx, z\rangle = \langle x, T^*z\rangle = \langle x, 0\rangle = 0.$$

This shows that $y \in \ker(T^*)^\perp$, and thus $\text{ran}(T) \subseteq \ker(T^*)^\perp$. Finally, since $\ker(T^*)^\perp$ is closed, we must have $\overline{\text{ran}(T)} \subseteq \ker(T^*)^\perp$.

Showing $\ker(T^*)^\perp \subseteq \overline{\text{ran}(T)}$*:* Let $x \in \overline{\text{ran}(T)}^\perp$. Then necessarily $x \in \text{ran}(T)^\perp$, meaning

$$0 = \langle Ty, x\rangle = \langle y, T^*x\rangle\,, \quad \text{for all } y \in X.$$

It follows that $T^*x = 0$, so $x \in \ker(T^*)$. This shows $\overline{\text{ran}(T)}^\perp \subseteq \ker(T^*)$. Taking orthogonal complements, we get

$$\ker(T^*)^\perp \subseteq \overline{\text{ran}(T)}^{\perp\perp} = \overline{\text{ran}(T)}.$$

ii) *Exercise.*

□

Example 5.3.5. Let R and L be the right and left shift operator, respectively.

$$\ker R = \{0\} = \text{ran}R = \{(0, x_1, x_2, ...) : (x_i) \in \ell^2\}$$
$$\ker L = \{x_1, 0, 0, ...\} = \text{ran}L = \ell^2.$$

Note that $\ker R^\perp = \overline{\text{ran}L}$ and $\ker L^\perp = \overline{\text{ran}R}$ and $R^* = L$ and $L^* = R$.

Corollary 5.14. Let T be a bounded linear operator on a Hilbert space X. Then $\ker(T^*) = \{0\}$ if and only if $\text{ran}(T)$ is dense in X.

PROOF. This is immediate from Proposition 5.13, as we have

$$X = \ker(T^*) \oplus \overline{\text{ran}(T)}.$$

□

This corollary allows one to check if the range of an operator is dense in the space X by determining the adjoint operator and its kernel. This can be a very useful strategy in practice, as it is often more difficult to determine the range of an operator than its kernel.

We mention another consequence for the solvability of linear systems.

Proposition 5.15. Suppose that T is a bounded linear operator on Hilbert space $(X, \langle ., . \rangle)$ with closed range. Then $Tx = b$ has a solution for x if and only if b is orthogonal to $\ker(T^*)$.

Note there is an elementary condition that implies that an operator has closed range: Any $T \in \mathcal{B}(X)$ for a Banach space X with the following property: If there exists a constant $c > 0$ such that $\|Tx\| \geq c\|x\|$ for all $x \in X$, then T has closed range.

The result indicates that the solutions of $Tx = b$ are closely related to the structure of the adjoint linear system $T^*x = b$.

Proposition 5.16. Suppose $T \in \mathcal{B}(X)$ has closed range. If $T^*x = b$ has a unique solution, then $Tx = b$ has a solution for any $b \in X$.

PROOF. By assumption $T^*x = b$ has a unique solution, i.e. $\ker T^* = \{0\}$. Hence for any $b \in (\ker T^*)^\perp = X$ the equation $Tx = b$ has a solution. □

CHAPTER 6

Series and bases in normed spaces

In this chapter we investigate series and bases in normed spaces. In particular, we focus on Schauder bases for Banach spaces, and orthonormal bases for separable Hilbert spaces.

6.1. Linear dependence, bases and dimension

Let X be a vector space over a field $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, and let $S \subset X$ be any subset. Recall from Section 2.1 that a *linear combination* of vectors $v_1, \ldots, v_n \in S$ is a finite sum

$$a_1 v_1 + \cdots a_n v_n,$$

where $a_1, \ldots, a_n$ are scalars in $\mathbb{F}$, and the *span* of S is the set of all linear combinations of vectors in S:

$$\operatorname{span}(S) := \left\{ \sum_{j=1}^{N} a_j x_j \,:\, x_j \in S, a_j \in \mathbb{F} \right\}.$$

Definition 6.1.1. A set of vectors $x_1, x_2, \ldots$ is called *linearly independent* if

$$\sum_{j=1}^{n} a_j x_j = 0 \quad \Rightarrow \quad a_1 = a_2 \ldots = a_n = 0,$$

for all $n \in \mathbb{N}$ and $a_j \in \mathbb{F}$.

Contrarily, the set of vectors is called *linearly dependent* if one of the vectors is a linear combination of certain others, meaning

$$\sum_{j=1}^{n} a_j x_j = 0 \quad \text{for some } n \in \mathbb{N} \text{ and at least one } a_j \neq 0.$$

Example 6.1.2.

i) The vectors $x = (1, 0)$, $y = (2, 0)$ and $z = (1, 1)$ are linearly dependent i $\mathbb{R}^2$, since $y - 2x = 0$. However, both the sets $\{x, z\}$ and $\{y, z\}$ are linearly independent.

ii) $\{1, x, x^2, x^3, \ldots\}$ is linearly independent in the space of all polynomials $\mathcal{P}$.

iii) $\{1, \cos x, \sin x, \cos 2x, \sin 2x, \ldots\}$ is linearly independent in the space $C[a, b]$ of real-valued continuous functions on an interval $[a, b]$.

A linearly independent set which generates a vector space X is called a *Hamel basis*. This is the kind of basis you will most likely have seen in an earlier course on linear algebra.

Definition 6.1.3. We call a linearly independent set S of a vector space X a *Hamel basis* if S spans X, i.e. if any $x \in X$ has a unique and finite representation

$$x = a_1 x_1 + \cdots + a_n x_n, \quad x_j \in S, a_j \in \mathbb{F}.$$

Example 6.1.4.

i) The set $\{e_1, \dots e_n\} \subset \mathbb{R}^n$, where

$$e_j = (0, \dots, 0, 1, 0, \dots, 0),$$

and the 1 appears at the jth entry, is called the *standard basis* in $\mathbb{R}^n$. This is a Hamel basis.

ii) $\{1, x, x^2, \dots\}$ is a Hamel basis for $\mathcal{P}(\mathbb{R})$: every real polynomial can be uniquely expressed as a finite sum

$$p(x) = \sum_{j=0}^{N} a_j x^j, \quad a_j \in \mathbb{R}.$$

Definition 6.1.5. If the vector space X has a (Hamel) basis consisting of finitely many vectors, then X is said to be *finite-dimensional*. Otherwise, we call X *infinite-dimensional*.

It is a fact that all bases of a finite-dimensional vector space have the same number of elements. This unique number is called the **dimension** of the space.

Example 6.1.6.

i) $\mathbb{R}^n$ has dimension n.

ii) $\mathcal{P}_n(\mathbb{R})$ has dimension $n+1$. (Recall from Section 3.3 that $\mathcal{P}_n(\mathbb{R}) \cong \mathbb{R}^{n+1}$.)

iii) $\mathbb{C}^n$ has dimension n when considered as a *complex* vector space, but $2n$ when considered a *real* vector space.

iv) The spaces ℓ^p, $C[a,b]$ and $L^2[a,b]$ are all infinite-dimensional.

6.2. Schauder bases

The concept of a Hamel basis is very general, and any vector space has a Hamel basis. However, these bases are not particularly well suited for infinite-dimensional Banach spaces.

Proposition 6.1. Infinite-dimensional Banach spaces have only uncountable Hamel bases.

(We state this result without proof.)

In other words, given that X is an infinite-dimensional Banach space, there is no sequence $\{x_j\}_{j\in\mathbb{N}}$ of elements of X which can serve as a Hamel basis for X.

Definition 6.2.1. A countable set $\{x_1, x_2, x_3, \dots\}$ of a normed space $(X, \|\cdot\|)$ is called a Schauder basis if for every $x \in X$ there exists a unique sequence of

scalars $(a_j)_{j\in\mathbb{N}}$ such that

$$\|x - (a_1x_1 + \cdots a_nx_n)\| \to 0 \quad (\text{as } n \to \infty).$$

In this case we write x as the sum

$$x = \sum_{j=1}^{\infty} a_j x_j, \tag{6.1}$$

and this is called the *series expansion* of x with respect to $(x_j)_{j\in\mathbb{N}}$.

Example 6.2.2. i) Denote by $e_n \in \ell^p$ the sequence whose nth term is 1 and all other terms are zero. The set $\{e_n\}_{n\in\mathbb{N}}$ is a Schauder basis in ℓ^p for $1 \leq p < \infty$.

ii) The set of trigonometric functions $\{e^{2\pi ikx}\}_{k\in\mathbb{Z}}$ is a Schauder basis for $L^2[0,1]$.

A Schauder basis allows us to take infinite linear combinations for expressing elements of X, as the norm on X gives a way of defining this limiting procedure. Note that if a normed space has a Schauder basis, it is necessarily separable (Exercise.). As a consequence, the space ℓ^∞ cannot have a Schauder basis. Nevertheless, we want to be able to discuss infinite *series*, such as that given in (6.1), also here.

Definition 6.2.3. Let $(X, \|\cdot\|)$ be a normed space and $(x_j)_{j\in\mathbb{N}}$ a sequence of vectors in X. We can associate with (x_j) the sequence $(s_n)_{n\in\mathbb{N}}$ of *partial sums*

$$s_n = x_1 + x_2 + \cdots + x_n.$$

If (s_n) is convergent and $s_n \to s$ for some $s \in X$, meaning

$$\|s_n - s\| \to 0 \quad (\text{as } n \to \infty),$$

then the *(infinite) series* $\sum_{j=1}^{\infty} x_j$ is said to *converge* to s, and we write

$$s = \sum_{j=1}^{\infty} x_j.$$

6.3. Orthonormal systems and the closest point property

Recall from Section 2.3 that a set of vectors $\{x_1, x_2, \ldots\}$ in an inner product space $(X, \langle\cdot,\cdot\rangle)$ is said to be *orthogonal* if

$$\langle x_j, x_k\rangle = 0 \quad \text{for all } j \neq k.$$

Moreover, if $\|x_j\| = 1$ for each $j \in \mathbb{N}$, then the set is called *orthonormal*.

Example 6.3.1. i) In the space $\mathbb{R}^3$, the three unit vectors $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ form an orthonormal set.

ii) In the space ℓ^2, the sequence $(e_n)_{n\in\mathbb{N}}$ introduced in Example 6.2.2i) is an orthonormal set.

iii) The set of exponential functions $\{e^{2\pi ikx}\}_{k\in\mathbb{N}}$ is an orthonormal set in $L^2[0,1]$.

Lemma 6.2. An orthonormal set is linearly independent.

PROOF. Exercise. □

A great advantage of orthonormal sequences over arbitrary linearly independent sequences is the following. If we know that a given x can be represented as a linear combination of some elements of an orthonormal sequence, then the orthonormality makes the determination of coefficients very easy. Suppose $\{e_1, e_2, \ldots\}$ is an orthonormal sequence in an inner product space X and say $x \in \text{span}\{e_1, \ldots e_n\}$ for some fixed n. Then by the definition of span we have

$$x = \sum_{j=1}^{n} a_j e_j,$$

and by taking the inner product of this sum with a fixed e_k, we obtain

$$\langle x, e_k \rangle = \left\langle \sum a_j e_j, e_k \right\rangle = \sum a_j \langle e_j, e_k \rangle = a_k.$$

Thus, we get

$$x = \sum_{j=1}^{n} \langle x, e_j \rangle e_j.$$

More generally, if we consider any $x \in X$ (not necessarily in $Y_n = \text{span}\{e_1, \ldots, e_n\}$), we can define $y \in Y_n$ by

$$y = \sum_{j=1}^{n} \langle x, e_j \rangle e_j,$$

and then define z by setting $x = y + z$ (or equivalently $z = x - y$). We will now show that $z \perp y$. By orthonormality it follows that

$$\|y\|^2 = \left\langle \sum \langle x, e_j \rangle e_j, \sum \langle x, e_k \rangle e_k \right\rangle = \sum |\langle x, e_j \rangle|^2,$$

and thus

$$\begin{aligned} \langle z, y \rangle &= \langle x - y, y \rangle = \langle x, y \rangle - \langle y, y \rangle \\ &= \left\langle x, \sum \langle x, e_j \rangle e_j \right\rangle - \|y\|^2 \\ &= \sum \langle x, e_j \rangle \overline{\langle x, e_j \rangle} - \sum |\langle x, e_j \rangle|^2 = 0. \end{aligned}$$

This shows $z \perp y$, and it follows from the Pythagorean relation that

$$\|x\|^2 = \|y\|^2 + \|z\|^2,$$

or equivalently

$$\|z\|^2 = \|x\|^2 - \|y\|^2 = \|x\|^2 - \sum_{j=1}^{n} |\langle x, e_j \rangle|^2. \tag{6.2}$$

Since $\|z\| \geq 0$, we get for every $n = 1, 2, \ldots$

$$\sum_{j=1}^{n} |\langle x, e_j \rangle|^2 \leq \|x\|^2.$$

These sums have non-negative terms, so they form a monotone, increasing sequence bounded above by $\|x\|^2$. Thus, it must converge, and we get:

Theorem 6.3 (Bessel's inequality)**.** Let $(e_j)_{j\in\mathbb{N}}$ be an orthonormal sequence in an inner product space X. Then for every $x \in X$, we have

$$\sum_{j=1}^{\infty} |\langle x, e_j\rangle|^2 \leq \|x\|^2.$$

The inner products $\langle x, e_j\rangle$ are called the **Fourier coefficients** of x with respect to the orthonormal sequence (e_j).

In light of Lemma 5.2 and what we have just seen, it is tempting to suggest that $y = \sum_{j=1}^{n} \langle x, e_j\rangle e_j$ is a best approximation of $x \in X$ in the subspace $\operatorname{span}\{e_1, \dots e_n\}$. Let us now see that this is indeed the case.

Theorem 6.4. An orthonormal sequence (e_j) in an inner product space X satisfies

$$\left\| x - \sum_{j=1}^{n} a_j e_j \right\| \geq \left\| x - \sum_{j=1}^{n} \langle x, e_j\rangle e_j \right\|,$$

for any $x \in X$, any $n \in \mathbb{N}$ and any scalars $a_1, \dots a_n \in \mathbb{F}$. Equality holds if and only if $a_j = \langle x, e_j\rangle$ for each $j = 1, \dots, n$.

Proof. We have

$$\begin{aligned}
\left\| x - \sum_{j=1}^{n} a_j e_j \right\|^2 &= \|x\|^2 - 2\Re\left\langle x, \sum_{j=1}^{n} a_j e_j \right\rangle + \left\| \sum_{j=1}^{n} a_j e_j \right\|^2 \\
&= \|x\|^2 - 2\Re \sum_{j=1}^{n} \overline{a_j} \langle x, e_j\rangle + \sum_{j=1}^{n} |a_j|^2 \\
&= \|x\|^2 + \sum_{j=1}^{n} |\langle x, e_j\rangle - a_j|^2 - \sum_{j=1}^{n} |\langle x, e_j\rangle|^2 \\
&\geq \|x\|^2 - \sum_{j=1}^{n} |\langle x, e_j\rangle|^2 = \left\| x - \sum_{j=1}^{n} \langle x, e_j\rangle e_j \right\|^2,
\end{aligned}$$

where the last equality follows from (6.2), and we see that equality holds throughout if and only if $a_j = \langle x, e_j\rangle$ for every j. □

Corollary 6.5. If $\{e_1, \dots, e_n\}$ is an orthonormal system in an inner product space X, then $y = \sum_{j=1}^{n} \langle x, e_j\rangle e_j$ is the unique closest point to x in $\operatorname{span}\{e_1, \dots, e_n\}$, with $d = \|x - y\|$ given by

$$d^2 = \|x\|^2 - \sum_{j=1}^{n} |\langle x, e_j\rangle|^2.$$

We have seen that orthonormal sequences are very convenient to work with. The remaining practical problem is how to obtain an orthonormal sequence if an arbitrary linearly independent sequence is given. We have the following result,

which we will prove using a constructive procedure known as the *Gram-Schmidt orthogonalization algorithm*. Hence the proof is an important part of the next result, as it shows *how* we obtain the orthogonal sequence.

Proposition 6.6. Let X be an infinite-dimensional inner product space. Then X contains a countable orthonormal set.

PROOF. By assumption there exists a linearly independent subset $\{x_1, x_2, \ldots\}$ in X. We will show that there exists an orthonormal sequence $\{e_1, e_2, \ldots\}$ such that

$$\operatorname{span}\{x_1, \ldots, x_n\} = \operatorname{span}\{e_1, \ldots, e_n\}$$

for every fixed $n \in \mathbb{N}$.

First step: Set $e_1 := x_1/\|x_1\|$. Then $\operatorname{span}(x_1) = \operatorname{span}(e_1)$ and $\|e_1\| = 1$.

Induction step: Suppose that for some $n \geq 2$ we have constructed an orthonormal set $E_{n-1} = \{e_1, \ldots, e_{n-1}\}$ such that

$$\operatorname{span}(E_{n-1}) = \operatorname{span}\{x_1, \ldots, x_{n-1}\}.$$

We now project x_n onto E_{n-1}, and set

$$\tilde{e_n} := x_n - \sum_{j=1}^{n-1} \langle x_n, e_j \rangle \, e_j.$$

We have that $\tilde{e_n}$ must be contained in the span of the linearly independent set $\{E_{n-1}, x_n\}$, so $\tilde{e_n}$ must be nonzero (since the coefficient in front of x_n is non-zero). By construction, we have

$$\langle \tilde{e_n}, e_k \rangle = 0 \quad \text{for } k = 1, \ldots, n-1.$$

Thus, the normalized vector $e_n := \tilde{e_n}/\|\tilde{e_n}\|$ can be added to E_{n-1} to obtain the orthonormal set $E_n = \{e_1, \ldots, e_n\}$. Finally note that

$$\operatorname{span}(E_n) = \operatorname{span}\{E_{n-1}, x_n\} = \operatorname{span}\{x_1, \ldots, x_n\}.$$

□

6.4. Orthonormal bases and the Fourier series theorem

We have now seen that any infinite-dimensional inner product space contains a countable orthonormal sequence. Given such a sequence (e_n), we now raise the question of when a series of the form

$$\sum_{j=1}^{\infty} a_j e_j, \quad a_j \in \mathbb{F},$$

converges in the inner product space. Recall that we defined convergence of a series in Definition 6.2.3. We restrict our discussion to complete inner product spaces X (i.e. Hilbert spaces).

Theorem 6.7. Let $\{e_1, e_2, \ldots\}$ be an orthonormal sequence in a Hilbert space X. Then the series $\sum_{j=1}^{\infty} a_j e_j$ converges if and only if $a = (a_j)_{j \in \mathbb{N}} \in \ell^2$. In

this case, we have

$$\left\| \sum_{j=1}^{\infty} a_j e_j \right\| = \|a\|_{\ell^2}. \tag{6.3}$$

PROOF. Let s_n be the partial sum

$$s_n = a_1 e_1 + \cdots + a_n e_n,$$

and let

$$\sigma_n = |a_1|^2 + \cdots + |a_n|^2.$$

Then, by orthonormality, for any m and $n > m$ we have

$$\begin{aligned} \|s_n - s_m\|^2 &= \|a_{m+1} e_{m+1} + \cdots + a_n e_n\|^2 \\ &= |a_{m+1}|^2 + \cdots + |a_n|^2 = |\sigma_n - \sigma_m|. \end{aligned} \tag{6.4}$$

Hence (s_n) is Cauchy in X if and only if (σ_n) is Cauchy in $\mathbb{R}$. Since both spaces are complete, the sequences either both converge or both diverge. In the case of convergence, we put $m = 0$ and let $n \to \infty$ in (6.4) to obtain (6.3). □

The truly interesting orthonormal sets in inner product spaces and Hilbert spaces are those which consist of "sufficiently many" elements so that every element in the space can be represented (or sufficiently well approximated) by the use of these elements. In this respect, the following notions are relevant.

Definition 6.4.1. An orthonormal sequence $\{e_1, e_2, \ldots\}$ is *maximal* (or *total*) in an inner product space X if

$$\overline{\operatorname{span}\{e_1, e_2 \ldots\}} = X,$$

or, equivalently, if $\overline{\operatorname{span}\{e_1, e_2 \ldots\}}^{\perp} = \{0\}$.

Thus we have that $\{e_1, e_2, \ldots\}$ is total in X if and only if for any $x \in X$ we have $\langle x, e_n \rangle = 0$ for all $n \in \mathbb{N}$ implies that $x = 0$.

Definition 6.4.2. An orthonormal sequence $\{e_1, e_2, \ldots\}$ in a Hilbert space X is called an *orthonormal basis* of X if

$$x = \sum_{j=1}^{\infty} \langle x, e_j \rangle \, e_j$$

holds for any $x \in X$.

Example 6.4.3. i) In $\mathbb{R}^n$, $\mathbb{C}^n$ and ℓ^2, the canonical basis $\{e_j\}_{j \in \mathbb{N}}$ is also an orthonormal basis.

ii) The set $\{1/\sqrt{2}, \cos x, \sin x, \cos 2x, \sin 2x, \ldots\}$ is an orthonormal basis for real-valued functions in $L^2[-\pi, \pi]$ if we equip it with the inner product

$$\langle f, g \rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) g(x) \, dx.$$

iii) $\{e^{2\pi ikx}\}_{k\in\mathbb{Z}}$ is an orthonormal basis for $L^2[0,1]$ (equipped with the usual inner product).

It is precisely when the orthonormal sequence $\{e_1, e_2, \dots\}$ is maximal that it provides an orthonormal basis in the Hilbert space X.

Theorem 6.8 (The Fourier series theorem)**.** Let $S = (e_j)_{j\in\mathbb{N}}$ be an orthonormal sequence in a Hilbert space X. The following are equivalent:

i) S is maximal, meaning $\overline{\text{span}S} = X$.

ii) S is an orthonormal basis for X.

iii) $\sum_{j=1}^{\infty} |\langle x, e_j\rangle|^2 = \|x\|^2$ for all $x \in X$.

Remark. The last equality in Theorem 6.8 is known as **Parseval's identity**.

Proof of Theorem 6.8. *i) ⇒ ii):* Let $x \in X$. By Bessel's inequality, we know that

$$\sum_{j=1}^{\infty} |\langle x, e_j\rangle|^2 \leq \|x\|^2 < \infty,$$

so Theorem 6.7 ensures that the series $\sum_{j=1}^{\infty} \langle x, e_j\rangle e_j$ converges. We need to show that $x = \sum_{j=1}^{\infty} \langle x, e_j\rangle e_j$. For any $k \in \mathbb{N}$, we find that

$$\left\langle x - \sum_{j=1}^{\infty} \langle x, e_j\rangle e_j, e_k \right\rangle = \langle x, e_k\rangle - \sum_{j=1}^{\infty} \langle x, e_j\rangle \langle e_j, e_k\rangle = \langle x, e_k\rangle - \langle x, e_k\rangle = 0.$$

This shows $x - \sum_{j=1}^{\infty} \langle x, e_j\rangle e_j \in \overline{\text{span}(S)}^{\perp} = \{0\}$, so $x = \sum_{j=1}^{\infty} \langle x, e_j\rangle e_j$.

ii) ⇒ iii): If S is an orthonormal basis, then $x = \sum \langle x, e_j\rangle e_j$, and

$$\|x\|^2 = \left\langle \sum_{j\in\mathbb{N}} \langle x, e_j\rangle e_j, \sum_{j\in\mathbb{N}} \langle x, e_j\rangle e_j \right\rangle = \sum_{j\in\mathbb{N}} |\langle x, e_j\rangle|^2$$

iii) ⇒ i): Suppose $x \in \overline{\text{span}S}^{\perp}$. Then $\|x\|^2 = \sum_{j\in\mathbb{N}} |\langle x, e_j\rangle|^2 = 0$, and it follows that $x = 0$. This shows that $\overline{\text{span}S}^{\perp} = \{0\}$, and thus S is maximal. □

We observe the following:

Theorem 6.9. Any separable Hilbert space X has a (countable) orthonormal basis.

We will not prove this rigorously, but point out that it can be proven inductively by starting with a countable, dense set in X, reducing this set until it is also linearly independent, and finally applying the Gram-Schmidt algorithm to obtain an orthonormal set. In light of Theorem 6.9, it follows from Theorem 6.8 that the elements of any separable Hilbert space are uniquely determined by their Fourier coefficients. In other words, any separable Hilbert space "looks like" ℓ^2.

Theorem 6.10 (Riesz-Fischer)**.** Every infinite-dimensional separable Hilbert space X is isometrically isomorphic to ℓ^2.

PROOF. As any infinite-dimensional separable Hilbert space X has an orthonormal basis (e_j), we can express every $x \in X$ uniquely by $x = \sum_{j=1}^{\infty} \langle x, e_j \rangle e_j$. We may therefore define a map $T : X \to \ell^2$ by

$$Tx = (\langle x, e_j \rangle)_{j \in \mathbb{N}}.$$

By Parseval's identity we have $\|Tx\| = \|x\|$ for any $x \in X$. Finally, T is surjective (this follows from Theorem 6.7). This shows that X and ℓ^2 are isometrically isomorphic,

$$X \cong \ell^2.$$

□

Corollary 6.11. Any finite-dimensional Hilbert space is isomorphic to $\mathbb{F}^n$.

6.5. Equivalent norms

We briefly make a detour back to normed spaces, and discuss the concept of *equivalent norms*.

Definition 6.5.1. Let X be a vector space and let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two norms on X. These are called *equivalent* if there exist (positive) constants C_1 and C_2 such that

$$C_1\|x\|_a \leq \|x\|_b \leq C_2\|x\|_a \quad \text{for all } x \in X.$$

We denote by $B_r^a(x) = \{y \in X : \|x-y\|_a < r\}$ and $B_r^b(x) = \{y \in X : \|x-y\|_b < r\}$ the open balls of radius r and center $x \in X$ with respect to the norms $\|.\|_a$ and $\|.\|_b$.

Proposition 6.12. Let $\|.\|_a$ and $\|.\|_b$ be two norms on a vector space X. Then the following statements are equivalent:

(1) $\|.\|_a$ and $\|.\|_b$ are equivalent norms.
(2) There exists some $r > 0$ such that $B_{1/r}^a(0) \subseteq B_1^b(0) \subseteq B_r^a(0)$.

PROOF. (1) $\Leftarrow$ (2):
Suppose that $\|.\|_a$ and $\|.\|_b$ are equivalent norms. Then there exists an $r > 0$ such that

$$\frac{1}{r}\|x\|_a \leq \|x\|_b \leq r\|x\|_a \qquad \text{for all } x \in X.$$

Then for x with $\|x\|_b < 1$ we have $\frac{1}{r}\|x\|_a \leq \|x\|_b < 1$ and thus we have $\|x\|_a < r$, i.e. $B_1^b(0) \subseteq B_r^a(0)$.
Now we assume $x \in X$ and $\|x\|_a < 1/r$. Then we get that $\|rx\|_a < 1$. Since the norms are equivalent we have $\frac{1}{r}\|rx\|_b \leq \|rx\|_a < 1$ and thus we have $\|x\|_b < 1$, i.e. $B_{1/r}^a(0) \subseteq B_1^b(0)$.

(2) $\Leftarrow$ (1):
Suppose $B_{1/r}^a(0) \subseteq B_1^b(0) \subseteq B_r^a(0)$ holds for some $r > 0$. Then for any $x \in X$ we have that $\frac{x}{2\|x\|_b}$ is in $B_1^b(0)$ and consequently in $B_r^a(0)$, i.e. $\|\frac{x}{2\|x\|_b}\|_a < r$. Hence

we have

$$\|x\|_b \leq 2r\|x\|_a.$$

The other inclusion follows by the same reasoning. □

Two equivalent norms on a vector space X necessarily give the same classes of convergent and Cauchy sequences.

Lemma 6.13. Suppose $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent norms on a vector space X.

i) A sequence (x_n) in X converges to $x \in X$ with respect to the norm $\|\cdot\|_a$ if and only if it converges to $x \in X$ with respect to the norm $\|\cdot\|_b$.

ii) A sequence (x_n) in X is Cauchy with respect to the norm $\|\cdot\|_a$ if and only if it is Cauchy with respect to the norm $\|\cdot\|_b$.

Proof. Exercise. □

An important consequence is the following:

Proposition 6.14. Suppose $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent norms on a vector space X. Then $(X, \|\cdot\|_a)$ is a Banach space if and only if $(X, \|\cdot\|_b)$ is a Banach space.

One may raise the question of whether *all* norms on a given vector space are equivalent. If the vector space in question is infinite-dimensional, then the answer is no.

Example 6.5.2. i) Let s be the vector space of real-valued sequences. Then the $\|\cdot\|_1$-norm and the $\|\cdot\|_\infty$-norm are not equivalent on s. To see this, fix some $N \in \mathbb{N}$, and consider the sequence $x = (1, \ldots, 1, 0, 0, \cdots)$ with N non-zero entries. Then $\|x\|_1 = N$ and $\|x\|_\infty = 1$. Hence we have

$$N\|x\|_\infty = \|x\|_1,$$

and since we can do this for every $N \in \mathbb{N}$ it is not possible to find a constant $C > 0$ such that

$$\|x\|_1 \leq C\|x\|_\infty \quad \text{for all } x \in s.$$

ii) Consider the space of continuous functions $C[0,1]$, and endow it with the two norms $\|\cdot\|_2$ and $\|\cdot\|_\infty$. We have seen that $(C[0,1], \|\cdot\|_\infty)$ is a Banach space, whereas $(C[0,1], \|\cdot\|_2)$ is not. Thus, by Proposition 6.14 these two norms cannot be equivalent.

Here is a general result on non-equivalent norms.

Lemma 6.15. Suppose $\|.\|_a$ and $\|.\|_b$ are two norms on a vector space X. Then $\|.\|_a$ and $\|.\|_b$ are not equivalent if there exists a sequence (x_n) in X such that $\|x_n\|_b = 1$ for all $n \in \mathbb{N}$ but $\|x_n\|_a = n$ for all $n \in \mathbb{N}$.

Proof. Exercise. □

However, if the vector space in question is *finite-dimensional*, then the answer to the aforementioned question is in fact yes.

Theorem 6.16. On a finite-dimensional vector space X, all norms are equivalent. For instance, all norms are equivalent on $\mathbb{R}^n$.

We state this result without proof.

CHAPTER 7

Topics in linear algebra

In Chapter 4, we discussed bounded linear operators between normed spaces X and Y. In this chapter, we focus on the case when X and Y are finite-dimensional. We first establish the fundamental fact that $B(X,Y) \cong \mathcal{M}_{m\times n}(\mathbb{C})$ if X and Y are complex vector spaces of dimensions n and m, respectively. We then go on to discuss spectral theory for linear operators between finite-dimensional vector spaces, and finally consider certain useful matrix decompositions.

7.1. Linear transformations between finite-dimensional spaces

We learned in the previous chapter that any finite-dimensional vector space X of dimension n has a set of n linearly independent spanning vectors $\{x_1, \ldots, x_n\}$. We call this set a basis for X, and any other basis must necessarily have the same number of spanning vectors. As a consequence of the Riesz-Fischer theorem we noted that an n-dimensional vector space is isomorphic to $\mathbb{F}^n$. Now let $T : X \to Y$ be a linear operator between finite-dimensional vector spaces X and Y. We make the useful observation that *T is determined by its action on any basis of X.*

Lemma 7.1. Let X be a finite-dimensional vector space with basis $\{b_1, \ldots, b_n\}$. For any vectors $y_1, \ldots, y_n \in Y$ there exists precisely one linear transformation $T : X \to Y$ such that

$$Tb_j = y_j, \quad j = 1, \ldots, n.$$

Proof. Any $x \in X$ has a unique representation $x = \sum_{j=1}^n x_j b_j$. Hence we have

$$Tx = T(\sum_{j=1}^n x_j b_j) = \sum_{j=1}^n x_j T b_j.$$

Thus T is uniquely determined by the vectors $Tb_1, ..., Tb_n$ in Y. □

Example 7.1.1. Let $T : \mathbb{C}^n \to \mathbb{C}^m$ be the linear map given by matrix multiplication

$$Tx = Ax, \quad A \in \mathcal{M}_{m\times n}(\mathbb{C}).$$

Then the columns A_j of the matrix A are determined by the action on the standard basis $\{e_j\}_{j=1}^n$:

$$Ae_j = A_j, \quad j = 1, \ldots, n.$$

Note that A_j plays the role of y_j in the above lemma.

Example 7.1.2. The differential operator $\frac{d}{dx}$ is a linear operator on $\mathcal{P}_n(\mathbb{R})$. Since $\mathcal{P}_2(\mathbb{R}) \cong \mathbb{R}^3$ via the vector space isomorphism

$$\sum_{j=0}^{2} a_j x^j \to (a_0, a_1, a_2),$$

we see that

$$\frac{d}{dx} : \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ 2a_2 \\ 0 \end{bmatrix}$$

expresses the derivation

$$\frac{d}{dx}(a_0 + a_1 x + a_2 x^2) = a_1 + 2a_2 x + 0x^2.$$

Next we discuss the link between matrices and linear transformations. On the one hand a $m \times n$ matrix A defines a linear transformation from $\mathbb{C}^n$ to $\mathbb{C}^m$ by $Tx = Ax$. On the other hand any linear transformation on finite-dimensional vector spaces can by represented in matrix form relative to a choice of bases.

We present the details for this assertion. Let $\mathcal{B} = \{x_1, ..., x_n\}$ be a basis of X and $\mathcal{C} = \{y_1, ..., y_m\}$ be a basis of Y. Suppose T is a linear transformation $T : X \to Y$ Then

$$x = \sum_{i=1}^{n} \alpha_i x_i$$

yields

$$T(x) = \sum_{i=1}^{n} \alpha_i T(x_i)$$

and thus

$$[T(x)]_{\mathcal{C}} = \sum_{i=1}^{n} \alpha_i [T(x_i)]_{\mathcal{C}}.$$

We define a $m \times n$ matrix A which has as its j-th column $[[T(x_j)]_{\mathcal{C}}]$. Then we have

$$[Tx]_{\mathcal{C}} = A[x]_{\mathcal{B}}.$$

The matrix A represents T with respect to the bases $\mathcal{B}$ and $\mathcal{C}$. Sometimes, we denote this A sometimes by $[T]_{\mathcal{B}}^{\mathcal{C}}$.

We address now the relation between the matrix representation of T depending on the change of bases. Suppose we have two bases $\mathcal{B} = \{x_1, ..., x_n\}$ and $\mathcal{R} = \{y_1, ..., y_n\}$ for X. Let $x = \sum_{j=1}^{n} \alpha_i x_i$. Then

$$[x]_{\mathcal{R}} = \sum_{j=1}^{n} \alpha_i [x_i]_{\mathcal{R}}.$$

Define the $n \times n$ matrix P with j-th column $[x_j]_{\mathcal{R}}$, and we call P the *change of bases matrix*:

$$[x]_{\mathcal{R}} = P[x]_{\mathcal{B}}$$

and by the invertibility of P we also have

$$[x]_{\mathcal{B}} = P^{-1}[x]_{\mathcal{R}}.$$

Let now $\mathcal{C}$ and $\mathcal{S}$ be two bases for Y. Then a linear transformation $T : X \to Y$ has two matrix representations:

$$A = [T]_{\mathcal{B}}^{\mathcal{C}} \text{ and } B = [T]_{\mathcal{R}}^{\mathcal{S}}.$$

In other words we have

$$[Tx]_{\mathcal{C}} = A[x]_{\mathcal{B}} \quad , [Tx]_{\mathcal{S}} = B[x]_{\mathcal{R}}$$

for any $x \in X$. Let P be the change of bases matrix of size $n \times n$ such that $[x]_{\mathcal{R}} = P[x]_{\mathcal{B}}$ for any $x \in X$ and let Q be the invertible $m \times m$ matrix such that $[y]_{\mathcal{S}} = Q[y]_{\mathcal{C}}$.
Hence we get that

$$[Tx]_{\mathcal{S}} = BP[x]_{\mathcal{B}}$$

and

$$[y]_{\mathcal{S}} = [Tx]_{\mathcal{S}} = Q[Tx]_{\mathcal{C}} = QA[x]_{\mathcal{B}}$$

for any $x \in X$. Hence we get that

$$B = QAP^{-1} \text{ and } A = Q^{-1}BP.$$

In the case $X = Y$ we have $P = Q$ and we set $S = Q^{-1}$ to get $B = S^{-1}AS$. Then the matrices A and B represent the same linear transformation T on V with respect to different bases.

Remark 7.1.3. If X and Y are both finite-dimensional normed spaces, then any linear transformation $T : X \to Y$ is automatically bounded. We therefore use $B(X, Y)$ to denote the linear transformations from X to Y when X and Y are finite-dimensional. Note that the preceding discussion yields that $B(X, Y) \cong \mathcal{M}_{m\times n}(\mathbb{C})$.

Recall from Section 2.1 that the kernel of a linear operator $T : X \to Y$,

$$\ker(T) = \{x \in X : Tx = 0\},$$

is a vector subspace of X, whereas the range of T,

$$\text{ran}(T) = \{y \in Y : Tx = y \text{ for some } x \in X\},$$

is a vector subspace of Y. When X and Y are finite-dimensional, and T is represented by a matrix $A \in \mathcal{M}_{m\times n}(\mathbb{C})$, then these subspaces are equivalently given by the so-called *null space* and *column space* of the matrix A:

- **Null space of A:** The kernel of T represented by A is clearly equal to the null space of A. We have

$$\begin{aligned} x \in \ker(T) \quad &\Leftrightarrow \quad Ax = 0 \quad \Leftrightarrow \quad \sum_{j=1}^{n} a_{ij}x_j = 0 \quad \forall i = 1, \dots m \\ &\Leftrightarrow (x_1, \dots, x_n) \perp (\overline{a_{i1}}, \dots, \overline{a_{in}}) \quad \forall i = 1, \dots m. \end{aligned}$$

  Note that the final line above tells us that the kernel of T (or null space of A) is the space of vectors $x \in \mathbb{C}^n$ orthogonal to the conjugated row vectors of A. We call the dimension of this subspace the **nullity** of T.

- **Column space of A:** The column space of A is the range of T. Since
$$Tx = Ax = A_1x_1 + \cdots + A_nx_n,$$
where $A_j = (a_{1j}, \ldots, a_{mj})^\top$ is the jth column vector of A, we have that
$$\operatorname{ran}(T) = \{Ax : x \in \mathbb{C}^n\} = \operatorname{span}\{A_1, \ldots, A_n\}.$$
This is precisely the column space of A. We call the dimension of this subspace the **rank** of T.
- **Row space of A:** The row space of A is the space spanned by the row vectors of A. Note that
$$\text{row space of } A = \text{column space of } A^\top,$$
where $A^\top$ is the transpose of A. The following result follows almost immediately.

Proposition 7.2. Let $A \in \mathcal{M}_{m\times n}(\mathbb{C})$. Then
$$\ker(A) \perp \operatorname{ran}(\overline{A}^\top).$$
In words, the kernel of A is orthogonal to the range of $\overline{A}^\top$.

Proof. We have just seen that the kernel, or null space, of A is orthogonal to the row space of $\overline{A}$. This is in turn equal to the column space, or range, of $\overline{A}^\top$. □

Finally let us state the *rank-nullity theorem* and see some important consequences.

Theorem 7.3. Let $T \in B(\mathbb{C}^n, \mathbb{C}^m)$. Then
$$\dim(\ker(T)) + \dim(\operatorname{ran}(T)) = n.$$

Proof. Pick a basis $\{e_1, \ldots, e_k\}$ for $\ker T$. If $k = n$ and $\ker(T) = \mathbb{C}^n$, we are done, since then $\operatorname{ran}(T) = \{0\}$, and
$$\dim(\ker(T)) + \dim(\operatorname{ran}(T)) = n + 0 = n.$$

Now assume $k < n$, and extend $\{e_1, \ldots, e_k\}$ to a basis $\{e_1, \ldots, e_k, f_1, \ldots, f_l\}$ for $\mathbb{C}^n$. This can be done in the following way: pick $f_1 \notin \operatorname{span}\{e_1, \ldots, e_k\}$. Then $\{e_1, \ldots, e_k, f_1\}$ is linearly independent. If this set of vectors spans all of $\mathbb{C}^n$, we stop. If not, we pick $f_2 \notin \operatorname{span}\{e_1, \ldots, e_k, f_1\}$. This process will necessarily stop when $k + l = n$ (because any linearly independent set of vectors spanning $\mathbb{C}^n$ has precisely n elements).

To finish the proof, we prove that $Tf = \{Tf_1, \ldots, Tf_l\}$ is a basis for $\operatorname{ran}(T)$. We observe first that Tf is linearly independent:
$$\begin{aligned}\sum_{j=1}^{l} a_j T f_j = T\left(\sum_{j=1}^{l} a_j f_j\right) = 0 \quad &\Leftrightarrow \quad \sum_{j=1}^{l} a_j f_j \in \ker T \\ &\Leftrightarrow \quad a_j = 0 \text{ for } j = 1, 2, \ldots, l.\end{aligned}$$
The last implication follows from the fact that by construction, no nonzero linear combination of vectors f_j lies in $\ker(T)$. Now let us see that Tf spans $\operatorname{ran}(T)$. By

the linearity of T we have

$$\begin{aligned}\operatorname{ran}(T) = \{Tx \,:\, x \in \mathbb{C}^n\} &= \left\{ T\Big(\sum_{j=1}^{k} a_j e_j + \sum_{j=1}^{l} b_j f_j\Big) \,:\, a_j, b_j \in \mathbb{C} \right\} \\ &= \left\{ T\Big(\sum_{j=1}^{k} a_j e_j\Big) + T\Big(\sum_{j=1}^{l} b_j f_j\Big) \,:\, a_j, b_j \in \mathbb{C} \right\} \\ &= \left\{ \sum_{j=1}^{l} b_j T f_j \,:\, b_j \in \mathbb{C} \right\}.\end{aligned}$$

Hence $\{Tf_1, \ldots, Tf_l\}$ is a basis for $\operatorname{ran}(T)$, and

$$\dim(\ker(T)) + \dim(\operatorname{ran}(T)) = k + l = n.$$

□

An immediate consequence of the rank-nullity theorem is that a linear map $T : \mathbb{C}^n \to \mathbb{C}^n$ is injective if and only if it is surjective.

Corollary 7.4. Let $T \in B(\mathbb{C}^n, \mathbb{C}^n)$. Then the following are equivalent.

i) T is injective ($\ker(T) = \{0\}$).

ii) T is surjective ($\operatorname{ran}(T) = \mathbb{C}^n$).

iii) T is invertible.

iv) The matrix representation A of T (in any given basis) is invertible.

v) For any $b \in \mathbb{C}^n$, the system $Ax = b$ has a unique solution.

We close this section with a geometric version of the rank-nullity theorem.

Corollary 7.5. Let $A \in \mathcal{M}_{m\times n}(\mathbb{C})$. Then

$$\mathbb{C}^n = \ker(A) \oplus \operatorname{ran}(\overline{A}^{\top}).$$

Replacing A by $\overline{A}^{\top}$ in the Corollary above, we immediately also have

$$\mathbb{C}^m = \ker(\overline{A}^{\top}) \oplus \operatorname{ran}(A).$$

In light of the fact that the adoint T^* of the operator T represented by the matrix A is given by

$$T^*x = \overline{A}^{\top} x,$$

Corollary 7.5 is an immediate consequence of Proposition 5.13. Nevertheless, let us also deduce this using the rank-nullity theorem. We will need the following preliminary result on dimension of subspaces.

Lemma 7.6. Let M, N be subspaces of a finite-dimensional vector space X. Then

$$\dim(M + N) + \dim(M \cap N) = \dim(M) + \dim(N).$$

In particular, if $M \perp N$, then

$$\dim(M+N) = \dim(M) + \dim(N).$$

PROOF OF COROLLARY 7.5. We have already seen that $\ker(A) \perp \operatorname{ran}(\overline{A}^\top)$ in $\mathbb{C}^n$, so by the Lemma above we have

$$\dim(\ker(A)) + \dim(\operatorname{ran}(\overline{A}^\top)) = \dim(\ker(A) + \operatorname{ran}(\overline{A}^\top)) = l, \quad l \leq n.$$

If we let $k = \dim(\ker(A))$, then by the rank-nullity theorem we have

$$\dim(\operatorname{ran}(\overline{A}^\top)) = l - k = l - (n - \dim(\operatorname{ran}(A))) \leq \dim(\operatorname{ran}(A)).$$

However, this argument is independent of the specific matrix A, and replacing A by $\overline{A}^\top$ above, we get

$$\dim(\operatorname{ran}(A)) \leq \dim(\operatorname{ran}(\overline{A}^\top)),$$

and thus

$$\dim(\operatorname{ran}(A)) = \dim(\operatorname{ran}(\overline{A}^\top)).$$

It follows that

$$\dim(\ker(A) + \operatorname{ran}(\overline{A}^\top)) = \dim(\ker(A)) + \dim(\operatorname{ran}(\overline{A}^\top)) = n,$$

where the last equality is the rank-nullity theorem since $\dim(\operatorname{ran}(A)) = \dim(\operatorname{ran}(\overline{A}^\top))$, and

$$\mathbb{C}^n = \ker(A) \oplus \operatorname{ran}(\overline{A}^\top).$$

□

7.2. Eigenvalues and eigenvectors

In the next section, we will discuss similarity transformations between matrices and establish Schur's triangulization lemma. This requires that we recall some properties of eigenvalues and eigenvectors.

Definition 7.2.1. Let $T : X \to X$ be a linear transformation (for example, $T : \mathbb{C}^n \to \mathbb{C}^n$ given by a matrix A). Then the scalar $\lambda \in \mathbb{C}$ is called an *eigenvalue* of T if there exists a nonzero vector $v \in X$ such that

$$Tv = \lambda v.$$

The vector v is called an *eigenvector* corresponding to the eigenvalue λ.

Definition 7.2.2. Let $T : X \to X$ be a linear transformation. The set $\sigma(T)$ of scalars satisfying

$$\sigma(T) = \{z \in \mathbb{C} : T - zI \text{ is not invertible}\}$$

is called the *spectrum* of T.

Proposition 7.7. For a linear transformation represented by $A \in \mathcal{M}_{n\times n}(\mathbb{C})$,

$$\sigma(A) = \{\lambda \in \mathbb{C} : \det(A - \lambda I) = 0\}$$

consists of the roots $(\lambda_1, \ldots, \lambda_n)$ of the *characteristic polynomial* $p_A(\lambda) = \det(A - \lambda I)$; these are precisely the eigenvalues of A.

PROOF. Exercise. □

We recall the following notions related to eigenvalues of a matrix $A \in \mathcal{M}_{n\times n}(\mathbb{C})$:

- The multiplicity of a root λ of $p_A(\lambda)$ is the **algebraic multiplicity** of the eigenvalue λ.
- The eigenvectors corresponding to an eigenvalue λ span a subspace of $\mathbb{C}^n$,
$$\ker(A - \lambda I),$$
called the **eigenspace** of λ. The dimension of this space is the **geometric multiplicity** of λ.

In other words, x is an eigenvector of T if and only if $x \in \ker T - \lambda I$. For finite-dimensional vector spaces $\sigma(T)$ is the set of all eigenvalues counting multiplicities of T.

Theorem 7.8. Suppose T is a linear transformation on a finite-dimensional complex vector space. Then there exists an eigenvalue $\lambda \in \mathbb{C}$ for an eigenvector x of T.

PROOF. We assume that $\dim(X) = n$ and choose any non-zero vector x in X. Consider the following set of $n + 1$ vectors in X:
$$\{x, Tx, T^2x, ..., T^nx\}.$$
Since $n + 1$ vectors in an n-dimensional vector space X are linearly independent, there exists a non-trivial linear combination:
$$a_0x + a_1Tx + \cdots + a_nT^nx = (a_0I + a_1T + \cdots + a_nT^n)x = 0.$$
Note that not all $a_1, ..., a_n$ are zero. If they were all zero, then $a_0x = 0$ which would imply that $a_0 = 0$. Hence that the linear combination is trivial.

Let us denote by $p(z) = a_0 + a_1z + \cdots + a_nz^n$ the polynomial associated to the linear transformation T. Powers of numbers correspond to powers of T by the corresponding iterates of T and $T^0 = I$.

Then the non-trivial linear combination among the vectors turns into a polynomial equation in T:
$$p(T) = 0.$$
By the Fundamental Theorem of Algebra any polynomial can be written as a product of linear factors:
$$p(t) = c(t - \lambda_1)(t - \lambda_2)\cdots(t - \lambda_n),\ \lambda_i \in \mathbb{C}, c \neq 0.$$
Hence $p(T)$ has a factorization of the form:
$$p(T) = c(T - \lambda_1 I)(T - \lambda_2 I)\cdots(T - \lambda_m I).$$
Hence $p(T)$ is a product of linear mappings $T - \lambda_j I$ for $j = 1, ..., m$. We know that $p(T)x = 0$ for a non-zero $x \neq 0$, which implies that at least one of these linear mappings is not invertible. Thus it has to have a non-trivial kernel, let's

say $y \in \ker(T - \lambda_i I)$, which yields that y is an eigenvector for the eigenvalue λ_i. Consequently, we have shown the desired assertion. □

Definition 7.2.3. Suppose that the matrix $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ has n linearly independent eigenvectors. If these eigenvectors are the columns of a matrix S, then $S^{-1}AS$ is a diagonal matrix Λ with the eigenvalues of A on its diagonal:

$$S^{-1}AS = \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

This is called the *diagonalization* of A.

Note that the definition above is a simple consequence of the fact that if A has eigenvalues $\lambda_1, \ldots, \lambda_n$ with associated, and linearly independent, eigenvectors $v_1, \ldots, v_n$, then we may rewrite the set of equations

$$\begin{aligned} Av_1 &= \lambda_1 v_1 \\ &\vdots \\ Av_n &= \lambda_n v_n \end{aligned}$$

in matrix form $AS = S\Lambda$, where S is the matrix with column vectors $v_1, \ldots v_n$. Since the vectors v_j are linearly independent, the matrix S is invertible.

Remark 7.2.4.

i) If the eigenvectors $v_1, \ldots, v_k$ correspond to *different* eigenvalues $\lambda_1, \ldots, \lambda_k$, then they are automatically linearly independent, as you will prove in problem set 12. Therefore any $(n \times n)$ matrix with n distinct eigenvalues can be diagonalized.

ii) The diagonalization is not unique, as any eigenvector v_j can be multiplied by a constant and remains an eigenvector. Repeated eigenvalues leave even more freedom. For the trivial example $A = I$, any invertible S will do, since $S^{-1}IS = I$ is diagonal.

iii) Not all matrices possess n linearly independent eigenvectors, so not all matrices are diagonalizable. The standard example of a "defective" matrix is

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Exercise: Show that this matrix cannot be diagonalized.

Recall from Section 5.3 that a map $T \in B(\mathbb{C}^n, \mathbb{C}^n)$ is called

i) normal if $TT^* = T^*T$,

ii) unitary if $T^* = T^{-1}$, and

iii) self-adjoint or Hermitian if $T = T^*$.

Let $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ be the matrix representation of T. We have seen in Example 5.3.1iv) that T^* is then represented by the matrix $\overline{A}^\top$. Accordingly, we let $A^* = \overline{A}^\top$, and call the matrix A

i) normal if $AA^* = A^*A$,

ii) unitary if $A^* = A^{-1}$, and

iii) Hermitian if $A = A^*$.

We make certain observations on the eigenvalues and eigenvectors of Hermitian and unitary matrices.

Proposition 7.9. Let $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ be a Hermitian matrix. Then all eigenvalues of A are real, and any two eigenvectors corresponding to different eigenvalues are orthogonal.

Proof. Let λ be an eigenvalue of A, and v the corresponding eigenvector. Then

$$\langle Av, v\rangle = \langle v, A^*v\rangle = \langle v, Av\rangle,$$

and since the inner product is conjugate symmetric ($\langle x, y\rangle = \overline{\langle y, x\rangle}$), it follows that $\langle Av, v\rangle$ is real-valued. On the other hand, we have

$$\langle Av, v\rangle = \langle \lambda v, v\rangle = \lambda\|v\|^2,$$

and since both $\langle Av, v\rangle$ and $\|v\|^2$ are real, the eigenvalue λ must be real-valued.

Now let λ_1 and λ_2 be two distinct eigenvalues of A, with corresponding eigenvectors x and y:

$$Ax = \lambda_1 x \quad \text{and} \quad Ay = \lambda_2 y.$$

Then

$$\lambda_1 \langle x, y\rangle = \langle Ax, y\rangle = \langle x, A^*y\rangle = \langle x, Ay\rangle = \lambda_2 \langle x, y\rangle,$$

and it follows that we must have $\langle x, y\rangle = 0$, meaning $x \perp y$. □

Proposition 7.10. Let $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ be a unitary matrix. Then every eigenvalue of A has absolute value $|\lambda| = 1$. Moreover, eigenvectors corresponding to different eigenvalues are orthogonal.

Proof. Let λ be an eigenvalue of A and v the corresponding eigenvector. Then

$$\langle Av, Av\rangle = \langle v, A^*Av\rangle = \langle v, v\rangle = \|v\|^2.$$

On the other hand

$$\langle Av, Av\rangle = \langle \lambda v, \lambda v\rangle = |\lambda|^2\|v\|^2,$$

and it follows that $|\lambda| = 1$ since $v \neq 0$.

Now let λ_1 and λ_2 be two distinct eigenvalues of A, with corresponding eigenvectors x and y:

$$Ax = \lambda_1 x \quad \text{and} \quad Ay = \lambda_2 y.$$

Then

$$\langle x, y\rangle = \langle Ax, Ay\rangle = \lambda_1\overline{\lambda_2}\langle x, y\rangle,$$

which implies that either $\lambda_1\overline{\lambda_2} = 1$ or $\langle x, y\rangle = 0$. However, we know that $\lambda_1\overline{\lambda_1} = 1$, so the first condition cannot possibly hold. We conclude that $\langle x, y\rangle = 0$. □

7.3. Similarity transformations and Schur's triangulization lemma

We saw in the previous section that if a matrix $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ has n linearly independent eigenvectors, then it has a diagonalization $\Lambda = S^{-1}AS$, where the matrix S has the eigenvectors of A as its columns. Let us now look at *all* combinations $M^{-1}AM$ formed with an invertible matrix M on the right and its inverse on the left.

Definition 7.3.1. We say that the matrices A and B in $\mathcal{M}_{n\times n}(\mathbb{C})$ are *similar* if there exists an invertible matrix M such that

$$B = M^{-1}AM.$$

The matrix M provides a *similarity transformation* from A to B. If M can be chosen unitary, then we say that A and B are *unitarily equivalent*.

At first glance it might not be obvious why we would be interested in similarity transforms, but the general idea is that a matrix B similar to A shares many properties with A, yet B might have a much more useful form than A.

Example 7.3.2. Similarity transformations arise in systems of differential equations, when a "change of variables" $u = Mv$ introduces the new unknown v:

$$\frac{du}{dt} = Au \quad \text{becomes} \quad M\frac{dv}{dt} = AMv, \quad \text{or} \quad \frac{dv}{dt} = M^{-1}AMv.$$

The new matrix in the equation is $M^{-1}AM$. In the special case that M is the eigenvector matrix S, the system becomes completely uncoupled, because $\Lambda = S^{-1}AS$ is diagonal. This is a maximal simplification, but other M's can also be useful. We try to make $M^{-1}AM$ easier to work with than A.

Note also that the similar matrix $B = M^{-1}AM$ is closely connected to A if we go back to linear transformations. Recall the key idea: Every linear transformation is represented by a matrix. However, this matrix depends on the choice of basis. If we recall our observations on page 91, we see that if we change the basis from $e = \{e_1, \dots, e_n\}$ to Me, then we change the matrix from A to B.

We will try to shed light on the following two questions:

(1) What do similar matrices $M^{-1}AM$ have in common?

(2) By picking M in a clever way, can we ensure that $M^{-1}AM$ has a special form?

Our first observation is that similar matrices have the same eigenvalues.

Lemma 7.11. Suppose $B = M^{-1}AM$. Then A and B have the same eigenvalues.

PROOF. We consider the characteristic polynomial of B:

$$\begin{aligned} p_B(z) &= \det(M^{-1}AM - zI) = \det(M^{-1}AM - M^{-1}Mz) \\ &= \det(M^{-1})\det(AM - zM) = \det(M^{-1})\det(A - zI)\det(M) = p_A(z) \end{aligned}$$

It follows that A and B must have the same eigenvalues. □

Let us now focus on question (2) above. We restrict our attention to the case where $M = U$ is unitary (meaning $U^* = U^{-1}$, which necessarily implies that U has

orthonormal columns). Unless the eigenvectors of A are orthogonal, it is impossible for $U^{-1}AU$ to be diagonal. However, Schur's triangulization lemma states the very useful fact that $U^{-1}AU$ can always achieve a triangular form.

Theorem 7.12 (Schur's triangulization lemma)**.** For any $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ there exists a unitary matrix U such that

$$U^{-1}AU = U^*AU = T,$$

where T is an upper triangular matrix, and where the eigenvalues of A appear (with multiplicity) along the diagonal of T.

We recall that an upper triangular matrix is one with only zeros below its diagonal:

$$T = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix}$$

Proof of Theorem 7.12. We proceed by induction on $n \geq 1$. For $n = 1$ there is nothing to do. Suppose now that the result is true for matrices up to size $n-1$ ($n \geq 2$). Let $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ (counting multiplicities). Consider an eigenvector v_1 associated to λ_1, and assume that $\|v_1\| = 1$. We use it to form an orthonormal basis $(v_1, \ldots, v_n)$, and we let V be the unitary matrix with v_j as its columns. The matrix A is equivalent to the matrix of the linear map $x \to Ax$ relative to the basis V, i.e.

$$A = V \left[\begin{array}{c|ccc} \lambda_1 & * & \cdots & * \\ \hline 0 & & & \\ \vdots & & \tilde{A} & \\ 0 & & & \end{array}\right] V^{-1} =: V\tilde{T}V^{-1}, \tag{7.1}$$

The matrices A and $\tilde{T}$ are similar, so they have the same eigenvalues. We see that $p_A(z) = (\lambda_1 - z)p_{\tilde{A}}(z)$, so the eigenvalues of the matrix $\tilde{A}$ must be $\lambda_2, \ldots, \lambda_n$. By the induction hypothesis there exists an $(n-1)\times(n-1)$ unitary matrix $\tilde{W}$ such that

$$\tilde{A} = \tilde{W} \begin{bmatrix} \lambda_2 & * & \cdots & * \\ 0 & \ddots & & \vdots \\ \vdots & & & * \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} \tilde{W}^{-1}.$$

By a tedious calculation it is not difficult to check that if we let

$$W := \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \tilde{W} & \\ 0 & & & \end{array}\right],$$

then

$$W^{-1}\tilde{T}W = \begin{bmatrix} \lambda_1 & * & \cdots & * \\ 0 & \ddots & & \vdots \\ \vdots & & & * \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} =: T.$$

It follows that $\tilde{T} = WTW^{-1}$, and inserting this in equation (7.1), we get

$$A = VWTW^{-1}V^{-1} = (VW)T(VW)^{-1}.$$

Finally, we observe that W and V are both unitary, so VW is also unitary, and the matrix T is of the desired form. □

As a consequence of Schur's triangulization lemma quantifies how many square matrices are diagonalizable.

Proposition 7.13. The set of diagonalizable matrices $\mathcal{D}$ is dense in $\mathcal{M}_n(\mathbb{C})$ with respect to the Frobenius norm. More explicitly, given $A \in \mathcal{M}_n(\mathbb{C})$ and $\varepsilon > 0$. There exists a diagonalizable matrix $\tilde{A} \in \mathcal{M}_n(\mathbb{C})$ such that

$$\sum_{i,j=1}^{n} |a_{ij} - \tilde{a}_{ij}|^2 < \varepsilon.$$

Proof. We have the Schur form for A

$$A = U \begin{pmatrix} \lambda_1 & x & \cdots & x \\ 0 & \lambda_2 & \ddots & x \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \lambda_n \end{pmatrix} U^*,$$

for a unitary matrix and eigenvalues $\lambda_1, ..., \lambda_n$ counting multiplicities. Define small perturbations of these eigenvalues λ_j such that these new numbers $\tilde{\lambda}_1, ..., \tilde{\lambda}_n$ are all distinct. We add multiples of a number η to the λ_j's:

$$\tilde{\lambda}_j = \lambda_j + j\eta, \quad \eta > 0$$

and fixed at the end of the proof. Set $\tilde{A}$

$$U \begin{pmatrix} \tilde{\lambda}_1 & x & \cdots & x \\ 0 & \tilde{\lambda}_2 & \ddots & x \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \tilde{\lambda}_n \end{pmatrix} U^*,$$

where we only change the diagonal entries of the upper triangular matrix. Now $\tilde{A}$ is diagonalizable and we have

$$\operatorname{tr}((A - \tilde{A})^*(A - \tilde{A})) = \sum_{i,j=1}^{n} |a_{ij} - \tilde{a}_{ij}|^2$$

Since the diagonal matrix with entries $\lambda_1 - \tilde{\lambda}_1, ..., \lambda_n - \tilde{\lambda}_n$ is unitarily equivalent to $A - \tilde{A}$ we deduce that

$$\mathrm{tr}((A - \tilde{A})^*(A - \tilde{A})) = \sum_{j=1}^{n} |\lambda_j - \tilde{\lambda}_j|^2.$$

By the definition of $\tilde{la}_j$ this gives

$$\sum_{j=1}^{n} |\lambda_j - \tilde{\lambda}_j|^2 = \eta^2 \sum_{j=1}^{n} j^2 = \eta^2 C_n.$$

Consequently,

$$\sum_{j=1}^{n} |\lambda_j - \tilde{\lambda}_j|^2 \leq \varepsilon$$

for $\eta \leq 2(\varepsilon/C_n)^{1/2}$. □

A variation of this argument allows one to demonstrate a similar statement for the set of invertible matrices.

Proposition 7.14. The set of invertible matrices is dense in $\mathcal{M}_n(\mathbb{C})$ with respect to the Frobenius norm.

Proof. Exercise. □

7.4. The spectral theorem

The following theorem, known as the Spectral Theorem, tells us precisely which matrices can be diagonalized.

Theorem 7.15 (Spectral Theorem)**.** Let $A \in \mathcal{M}_{n\times n}(\mathbb{C})$. Then A is diagonalizable, meaning there exists a diagonal matrix Λ (with the eigenvalues of A on the diagonal) and a unitary matrix U such that

$$A = U\Lambda U^{-1} = U\Lambda U^*,$$

if and only if A is normal (meaning $AA^* = A^*A$).

Before proving Theorem 7.15, we establish the following preliminary result.

Lemma 7.16. An upper triangular matrix is normal if and only if it is diagonal.

Proof. ($\Rightarrow$) : Suppose T is an upper triangular matrix. Then the (n,n)-th entry of TT^* is $|t_{nn}|^2$, while the (n,n)-th entry of T^*T is $|t_{nn}|^2 + \sum_{i=1}^{n-1} |t_{in}|^2$. If T is normal, then these two entries have to be the same. Hence $t_{in} = 0$ for $i = 1, ..., n-1$. Repeating this argument for the entries $(n-1, n-1), ...(2,2), (1,1)$ gives that T is diagonal.
($\Leftarrow$) : If T is diagonal, then T is certainly normal. □

Proof of Theorem 7.15. By Schur's triangulization lemma, there exists a unitary matrix U and an upper triangular matrix T such that

$$U^*AU = U^{-1}AU = T.$$

We observe that the matrix T is normal if A is normal, since

$$TT^* = (U^*AU)(U^*AU)^* = U^*AUU^*A^*U = U^*AA^*U \\ = U^*A^*AU = U^*A^*UU^*AU = T^*T,$$

and similarly A is normal if T is normal. Finally, by Lemma 7.16, T is normal if and only if it is diagonal. We know from Schur's triangulization lemma that we must have

$$T = \Lambda,$$

where Λ is the matrix with the eigenvalues of A on its diagonal. Finally, we observe that it follows from

$$AU = U\Lambda$$

that the columns of U must be the (orthonormal) eigenvectors of A. □

Hence Hermitian matrices and unitary matrices are diagonalizable since these are normal matrices.

7.5. Singular value decomposition and applications

Let $A \in \mathcal{M}_{m\times n}(\mathbb{C})$. If $m \neq n$, it no longer makes sense to ask if A can be diagonalized. However, one can raise the question of whether there exist two *different* unitary matrices U and V such that

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

and where Σ is a diagonal (but rectangular) matrix. It turns out that the answer to this question is yes, and that the specific factorization, known as the *singular value decomposition*, is closely related to the diagonalization of the normal matrix AA^* (or similarly A^*A). Before we state the singular value decomposition in detail and prove its existence, let us briefly discuss positive definite matrices.

Definition 7.5.1. A self-adjoint matrix $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ is said to be *positive definite* if

$$\langle Ax, x\rangle > 0, \quad \text{for all nonzero } x \in \mathbb{C}^n.$$

Similarly, if A satisfies the weaker condition

$$\langle Ax, x\rangle \geq 0, \quad \text{for all nonzero } x \in \mathbb{C}^n,$$

the A is said to be *positive semi-definite*.

A useful test for positive definiteness (or semi-definiteness) is to consider the eigenvalues of the matrix in question.

Proposition 7.17. A self-adjoint matrix $A \in \mathcal{M}_{n\times n}(\mathbb{C})$ is positive definite if and only if all its eigenvalues are positive. Similarly, A is positive semi-definite if and only if all its eigenvalues are non-negative.

Proof. ($\Leftarrow$): Suppose A is positive definite. Then

$$\langle Ax, x\rangle > 0 \quad \text{for all nonzero } x \in \mathbb{C}^n.$$

In particular, this holds for any eigenvector of A. Let x be an eigenvector associated to the eigenvalue λ. We have

$$\langle Ax, x\rangle = \langle \lambda x, x\rangle = \lambda \|x\|^2 > 0,$$

and it follows that $\lambda > 0$.
($\Rightarrow$): By the Spectral Theorem, there exists a unitary matrix U such that

$$A = U^* \Lambda U,$$

and where Λ is a diagonal matrix with the positive eigenvalues of A on its diagonal. It follows that

$$\langle Ax, x\rangle = \langle U^*\Lambda U x, x\rangle = \langle \Lambda U x, Ux\rangle .$$

Now let $y := Ux \in \mathbb{C}^n$. We then have

$$\langle Ax, x\rangle = \langle \Lambda y, y\rangle = \lambda_1 |y_1|^2 + \cdots \lambda_n |y_n|^2,$$

which is greater than zero for all nonzero $y \in \mathbb{C}^n$. Finally note that $y = 0$ if and only if $x = 0$. □

An important pair of self-adjoint, positive semi-definite matrices is AA^* and A^*A for any given $A \in \mathcal{M}_{m\times n}(\mathbb{C})$. The following result follows almost immediately from the proposition above.

Corollary 7.18. Let $A \in \mathcal{M}_{m\times n}(\mathbb{C})$. Then the $(n \times n)$ matrix A^*A and the $(m \times m)$ matrix AA^* are self-adjoint with non-negative eigenvalues, and the positive eigenvalues of the two matrices coincide.

For the proof of Corollary 7.18, we need the following lemma.

Lemma 7.19. For any $A \in \mathcal{M}_{m\times n}(\mathbb{C})$ and $B \in \mathcal{M}_{n\times m}(\mathbb{C})$, the matrices AB and BA have the same non-zero eigenvalues.

Proof. Exercise. □

Proof of Corollary 7.18. It is clear that AA^* and A^*A are both self-adjoint. Moreover, we have that

$$\|Ax\|^2 = \langle Ax, Ax\rangle = \langle A^*Ax, x\rangle \geq 0,$$

so A^*A is clearly positive semi-definite. Running the same argument with $\|A^*x\|$ shows that also AA^* is positive semi-definite. By Proposition 7.17, the eigenvalues of both matrices are non-negative, and by the preceeding lemma it finally follows that the positive eigenvalues of the two matrices coincide. □

Let us now return to the so-called singular value decomposition of a matrix.

Definition 7.5.2. Let $A \in \mathcal{M}_{m\times n}(\mathbb{C})$ have rank r. Let $\sigma_1^2 \geq \cdots \geq \sigma_r^2$ be the positive eigenvalues of A^*A. The scalars $\sigma_1, \ldots, \sigma_r$ are called the *positive singular values* of A.

Since the matrix A^*A is of size $n \times n$, it has n eigenvalues. Those that are not positive are necessarily equal to zero, and accordingly the matrix A has $n - r$ singular values $\sigma_j = 0$, $j = r+1, \ldots, n$. As we have just established that AA^* and

A^*A have the same nonzero eigenvalues, one may choose either one for determining the positive singular values of A.

Theorem 7.20 (Singular Value Decomposition)**.** Suppose $A \in \mathcal{M}_{m\times n}(\mathbb{C})$ is of rank r, and let $\sigma_1 \geq \cdots \geq \sigma_r$ be the positive singular values of A. Let Σ be the $(m \times n)$ matrix defined by

$$\Sigma_{ij} = \begin{cases} \sigma_i & \text{if } i = j \leq r \\ 0 & \text{otherwise.} \end{cases}$$

Then there exists an $(m \times m)$ unitary matrix U and an $(n \times n)$ unitary matrix V such that

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

Through the proof of Theorem 7.20 below, we will see that the columns of V are the (orthonormal) eigenvectors of A^*A.

Proof. The matrix A^*A is self-adjoint with positive eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_r^2$ and $(n-r)$ eigenvalues equal to zero. Thus, by the Spectral Theorem, there exists an $(n \times n)$ unitary matrix V such that

$$V^*A^*AV = (AV)^*(AV) = D, \tag{7.2}$$

where $D = \Sigma^*\Sigma$ is the $(n \times n)$ diagonal matrix with

$$D_{ii} = \sigma_i^2, \quad i = 1, \ldots, r,$$

and zeros elsewhere. It is clear from (7.2) that the (i, j)th entry of V^*A^*AV is the inner product of columns i and j in AV. Thus, the columns $(AV)_j$ of AV are pairwise orthogonal. Moreover, for $1 \leq j \leq r$, the length of $(AV)_j$ is σ_j. Let U_r denote the $(m \times r)$ matrix with $(AV)_j/\sigma_j$ as its jth column. Complete U_r to an $(m \times m)$ unitary matrix U by finding an orthonormal basis for the orthogonal complement of (the column space of) U_r, and using these basis vectors as the last $(m-r)$ columns in U. We then have

$$AV = U\Sigma \quad \Leftrightarrow \quad A = U\Sigma V^*.$$

□

Remark 7.5.3. Since only the first r diagonal entries of Σ are nonzero, we see that the last $(m-r)$ columns of U, and likewise the last $(n-r)$ columns of V, are superfluous. As a consequence, we have that a given matrix A has an SVD where the diagonal matrix Σ is uniquely determined, but the unitary matrices U and V are *not*.

Example 7.5.4. Let us determine the singular value decomposition of

$$A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}.$$

The procedure for finding the SVD is as follows: We begin by determining the positive eigenvalues of A^*A (or similarly AA^*). We have

$$A^*A = \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} = \begin{bmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{bmatrix}$$

The positive eigenvalues of this matrix are $\sigma_1^2 = 25$ and $\sigma_2^2 = 9$. The last eigenvalue is $\sigma_3^2 = 0$. Since A^*A is self-adjoint (or Hermitian), the eigenvectors corresponding to σ_1^2, σ_2^2 and σ_3^2 are necessarily orthogonal. We find these eigenvectors, and choose them to have length 1:

$\underline{\sigma_1^2 = 25}$:

$$A^*A - 25I = \begin{bmatrix} 13-25 & 12 & 2 \\ 12 & 13-25 & -2 \\ 2 & -2 & 8-25 \end{bmatrix} \sim \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & -1 & -\frac{17}{2} \end{bmatrix},$$

and solving for $A^*A - 25I = 0$, we find that $v_1 = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ 0 \end{pmatrix}$ is a normalized eigenvector.

$\underline{\sigma_2^2 = 9}$:

$$A^*A - 9I = \begin{bmatrix} 13-9 & 12 & 2 \\ 12 & 13-9 & -2 \\ 2 & -2 & 8-9 \end{bmatrix} \sim \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & \frac{1}{4} \\ 1 & 0 & -\frac{1}{4} \end{bmatrix},$$

and solving for $A^*A - 9I = 0$, we find that $v_2 = \begin{pmatrix} \frac{\sqrt{2}}{6} \\ -\frac{\sqrt{2}}{6} \\ \frac{2\sqrt{2}}{3} \end{pmatrix}$ is a normalized eigenvector.

$\underline{\sigma_3^2 = 0}$:

$$A^*A = \begin{bmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{bmatrix} \sim \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -2 \\ 1 & 0 & 2 \end{bmatrix},$$

and solving for $A^*A = 0$, we find that $v_3 = \begin{pmatrix} \frac{2}{3} \\ -\frac{2}{3} \\ -\frac{1}{3} \end{pmatrix}$ is a normalized eigenvector.

We can now "build" all the matrices that enter into the SVD of the matrix A. We get

$$V = \big[v_1 | v_2 | v_3\big] = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{6} & \frac{2}{3} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} & -\frac{2}{3} \\ 0 & \frac{2\sqrt{2}}{3} & -\frac{1}{3} \end{bmatrix},$$

and

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}.$$

Finally, we find that

$$U = \big[U_1 | U_2\big] = \left[\tfrac{Av_1}{\|Av_1\|} \big| \tfrac{Av_2}{\|Av_2\|}\right] = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}.$$

With these choices of U, Σ and V, we have that $A = U\Sigma V^*$, or explicitly written out:

$$A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{6} & -\frac{\sqrt{2}}{6} & \frac{2\sqrt{2}}{3} \\ \frac{2}{3} & -\frac{2}{3} & -\frac{1}{3} \end{bmatrix}.$$

Let us now discuss some consequences and applications of the SVD Theorem.

Proposition 7.21. Let $A \in \mathcal{M}_{m\times n}(\mathbb{C})$ have positive singular values $\sigma_1 \geq \cdots \geq \sigma_r$. Then the operator norm of A (that is, the norm of the bounded linear operator associated with A) is

$$\|A\| = \sigma_1.$$

Proof. Let $A = U\Sigma V^*$ be the singular value decomposition of A, and let v_1 be the first column vector of V. The vector v_1 has length 1, and from the equation $AV = U\Sigma$ it is clear that $\|Av_1\| = \sigma_1$. It follows that

$$\|A\| = \sup_{\|x\|=1} \|Ax\| \geq \sigma_1.$$

Now let $x \in \mathbb{C}^n$ be any vector of length 1, and consider the equation $Ax = U\Sigma V^*x$. Since V^* is unitary, it represents an isometry, and it follows that $\|V^*x\| = 1$. Let us denote this vector by $y := V^*x$. Moreover, we note that Σy is the vector where the jth component of y is multiplied by σ_j. Thus, we have $\|\Sigma y\| \leq \sigma_1\|y\|$. Finally, since U is also unitary, we have

$$\|Ax\| = \|U\Sigma y\| = \|\Sigma y\| \leq \sigma_1\|y\| = \sigma_1,$$

and it follows that $\|A\| \leq \sigma_1$. We thus conclude that $\|A\| = \sigma_1$. □

Let us now see that the SVD of a matrix can be used to obtain so-called *polar decompositions*. A polar decomposition factors a square matrix in a manner analogous to the factoring of a complex number as the product of a complex number of length 1 and a nonnegative number ($z = |z|e^{2\pi i\varphi}$). In the case of matrices, the complex number of length 1 is replaced by a unitary matrix, and the nonnegative number is replaced by a positive semi-definite matrix.

Theorem 7.22 (Polar decomposition)**.** For any square matrix A, there exists a unitary matrix W and a positive semi-definite matrix P such that

$$A = WP.$$

Proof. By the singular value decomposition theorem, there exist unitary matrices U and V and a diagonal matrix Σ with nonnegative diagonal entries such that $A = U\Sigma V^*$. It follows that

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

where $W = UV^*$ and $P = V\Sigma V^*$. Since W is the product of unitary matrices, W is unitary. Moreover, since Σ is positive semi-definite, so is the matrix P. □

Example 7.5.5. To find the polar decomposition of

$$A = \begin{bmatrix} 11 & -5 \\ -2 & 10 \end{bmatrix},$$

we begin by finding the SVD of $A = U\Sigma V^*$. It can be shown that

$$v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix} \quad \text{and} \quad v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}$$

are orthonormal eigenvectors of A^*A with corresponding eigenvalues $\sigma_1^2 = 200$ and $\sigma_2^2 = 50$. Thus, we have

$$V = \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{bmatrix} = \begin{bmatrix}10\sqrt{2} & 0\\ 0 & 5\sqrt{2}\end{bmatrix}.$$

Next, we find the columns of U:

$$u_1 = \frac{1}{\sigma_1}Av_1 = \frac{1}{5}\begin{pmatrix}4\\-3\end{pmatrix} \quad \text{and} \quad u_2 = \frac{1}{\sigma_2}Av_2 = \frac{1}{5}\begin{pmatrix}3\\4\end{pmatrix}.$$

Thus,

$$U = \begin{bmatrix}\frac{4}{5} & \frac{3}{5}\\ \frac{-3}{5} & \frac{4}{5}\end{bmatrix}.$$

Therefore, in the notation of the polar decomposition theorem, we have

$$W = UV^* = \begin{bmatrix}\frac{4}{5} & \frac{3}{5}\\ \frac{-3}{5} & \frac{4}{5}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix} = \frac{1}{5\sqrt{2}}\begin{bmatrix}7 & -1\\ 1 & 7\end{bmatrix},$$

and

$$P = V\Sigma V^* = \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}10\sqrt{2} & 0\\ 0 & 5\sqrt{2}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix} = \frac{5}{\sqrt{2}}\begin{bmatrix}3 & -1\\ -1 & 3\end{bmatrix}.$$

Finally, let us illustrate one possible application of SVD's to image processing.

Example 7.5.6. Suppose a satellite takes a picture, and wants to send it to earth. The picture may contain 1000×1000 pixels - a million little squares each with a definite color. We can code the colors, and send back 1000000 numbers. However, it is more convenient if we can find the *essential* information, and send only this.

Suppose we know the SVD, and specifically the matrix of singular values Σ. Typically, some of the σ's are significant, whereas others are extremely small. If we keep, say, 20 singular values, and discard the remaining 980, then we need only send the corresponding 20 columns of U and V. Thus, if only 20 singular values are kept, we send 20×20 numbers rather than a million.

There is, of course, the additional cost of computing the SVD. This has become quite efficient, but is still expensive for big matrices.

7.6. The pseudoinverse

Let V and W be finite-dimensional inner product spaces over the same field $\mathbb{F}$, and let $T : V \to W$ be a linear transformation. It is desirable to have a linear transformation from W to V which captures some of the essence of an inverse of T even if T is not invertible. A simple (but fruitful) approach to this problem is to focus on the "part" of T that is invertible, namely the restriction of T to $\ker(T)^\perp$. Let $L : \ker(T)^\perp \to \text{ran}(T)$ be the linear transformation defined by $L(x) = T(x)$ for all $x \in \ker(T)^\perp$. Then L is invertible, and we can use the inverse of L to construct a linear transformation from W to V which restores some of the benefits of an inverse of T.

Definition 7.6.1. Let V and W be finite-dimensional inner product spaces over the same field, and let $T : V \to W$ be a linear transformation. Let $L : \ker(T)^\perp \to \text{ran}(T)$ be the linear transformation defined by $L(x) = T(x)$ for all $x \in \ker(T)^\perp$. The *pseudoinverse* of T, denoted T^+, is defined as the unique linear transformation from W to V such that

$$T^+(y) = \begin{cases} L^{-1}(y) & \text{for } y \in \text{ran}(T) \\ 0 & \text{for } y \in \text{ran}(T)^\perp \end{cases}.$$

The pseudoinverse of a linear transformation T on a finite-dimensional inner product space exists even if T is not invertible. Furthermore, if T *is* invertible, then $T^+ = T^{-1}$, because $\ker(T)^\perp = V$ and L coincides with T.

Now let $A \in \mathcal{M}_{m\times n}(\mathbb{C})$ be the matrix representation of the linear map T. Then there exists a unique $(n \times m)$ matrix B which represents the pseudoinverse T^+. We call B the *pseudoinverse* of A and denote it by $B = A^+$. It turns out that the pseudoinverse A^+ can be computed with the aid of the singular value decomposition of A.

Theorem 7.23. Let $A \in \mathcal{M}_{m\times n}(\mathbb{C})$ have rank r and singular value decomposition $A = U\Sigma V^*$, where $\sigma_1 \geq \cdots \geq \sigma_r$ are the positive singular values of A. Let Σ^+ be the $(n \times m)$ matrix

$$\Sigma^+_{ij} = \begin{cases} \frac{1}{\sigma_i} & \text{if } i = j \leq r \\ 0 & \text{otherwise} \end{cases}.$$

Then $A^+ = V\Sigma^+U^*$.

We state this result without proof, and focus on its applications.

Let $b \in \mathbb{C}^m$, and consider the system of linear equations

$$Ax = b.$$

We know that this system has either no solution, a unique solution, or infinitely many solutions. It has a unique solution for every $b \in \mathbb{C}^m$ if and only if A is invertible, in which case the solution is given by $A^{-1}b$. Moreover, if A is invertible, then $A^{-1} = A^+$, so we could have written the solution as $x = A^+b$. If, on the other hand, the system $Ax = b$ is underdetermined or inconsistent, then A^+b still exists. This raises the question: How is the vector A^+b related to the system of linear equations $Ax = b$? In order to answer this question, we need the following lemma.

Lemma 7.24. Let V and W be finite-dimensional inner product spaces, and let $T : V \to W$ be linear. Then

i) T^+T is the orthogonal projection of V on $\ker(T)^\perp$.

ii) TT^+ is the orthogonal projection of W on $\text{ran}(T)$.

Proof. As above, we define $L : \ker(T)^\perp \to \text{ran}(T)$ by $L(x) = T(x)$ for $x \in \ker(T)^\perp$. If $x \in \ker(T)^\perp$, then

$$T^+T(x) = L^{-1}L(x) = x,$$

and if $x \in \ker(T)$, then

$$T^+T(x) = T^+(0) = 0.$$

Consequently, T^+T is the orthogonal projection of V on $\ker(T)^\perp$. This proves part i). Part ii) is proved similarly. □

Theorem 7.25. Consider the system of linear equations $Ax = b$, where $A \in \mathcal{M}_{m\times n}(\mathbb{C})$ and $b \in \mathbb{C}^m$. If $z = A^+b$, then z has the following properties.

i) If $Ax = b$ is consistent, then z is the unique solution to the system having minimum norm. That is, z is a solution to the system, and if y is any other solution to the system, then $\|y\| > \|z\|$.

ii) If $Ax = b$ is inconsistent, then z is the unique best approximation to a solution having minimum norm. That is

$$\|Az - b\| \leq \|Ay - b\| \quad \text{for any } y \in \mathbb{C}^n,$$

with equality if and only if $Ay = Az$. Moreover, if $Ay = Az$, then $\|z\| \leq \|y\|$ with equality if and only if $z = y$.

Proof. Let T be the linear map associated to the matrix A

i) Suppose that $Ax = b$ is consistent, and let $z = A^+b$. Observe that $b \in \text{ran}(T)$, and therefore

$$Az = AA^+b = TT^+b = b,$$

by Lemma 7.24ii). Thus, z is a solution to the system $Ax = b$. Now let y be any solution to the system. Then

$$T^+Ty = A^+Ay = A^+b = z.$$

Thus, z is the orthogonal projection of y on $\ker(T)^\perp$. By Theorem 5.5, we have $y = z + v$ with $v \in \ker(T)$, and $\|y\|^2 = \|z\|^2 + \|v\|^2$. It follows that $\|y\| > \|z\|$ unless $v = 0$ and $y = z$.

ii) Suppose that $Ax = b$ is inconsistent. By Lemma 7.24ii), we have that

$$Az = AA^+b = TT^+b$$

is the orthogonal projection of b on $\text{ran}(T)$. Therefore, by Theorem 5.5, Az is the vector in $\text{ran}(T)$ nearest b. If Ay is any other vector in $\text{ran}(T)$, then necessarily

$$\|Az - b\| \leq \|Ay - b\|,$$

with equality if and only if $Az = Ay$. Finally, suppose that y is any vector in $\mathbb{C}^n$ such that $Az = Ay = c$. Then

$$A^+c = A^+Az = A^+AA^+b = A^+b = z,$$

where we have used that $A^+AA^+ = A^+$ (this is easily checked by writing out the SVD of A). Hence, we may apply part i) of this theorem to the system $Ax = c$ to conclude that $\|y\| \geq \|z\|$ with equality if and only if $y = z$.

□

Example 7.6.2. Let us find the minimal norm solution of

$$-x_1 + 2x_2 + 2x_3 = b, \quad \text{for } b \in \mathbb{R}.$$

According to Theorem 7.25i), this is given by

$$z = A^+ b,$$

where A^+ is the pseudoinverse of the (1×3) matrix $A = \begin{bmatrix} -1 & 2 & 2 \end{bmatrix}$. The SVD of A is $A = U\Sigma V^*$, where

$$U = \begin{bmatrix} 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 3 & 0 & 0 \end{bmatrix}, \quad V = \begin{bmatrix} -\frac{1}{3} & \frac{2}{\sqrt{5}} & \frac{2}{3\sqrt{5}} \\ \frac{2}{3} & 0 & \frac{\sqrt{5}}{3} \\ \frac{2}{3} & \frac{1}{\sqrt{5}} & \frac{4}{3\sqrt{5}} \end{bmatrix}.$$

The pseudoinverse of A is thus given by

$$A^+ = V\Sigma^+ U^* = \begin{bmatrix} -\frac{1}{3} & \frac{2}{\sqrt{5}} & \frac{2}{3\sqrt{5}} \\ \frac{2}{3} & 0 & \frac{\sqrt{5}}{3} \\ \frac{2}{3} & \frac{1}{\sqrt{5}} & \frac{4}{3\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{1}{3} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{9} \\ \frac{2}{9} \\ \frac{2}{9}, \end{bmatrix}$$

and it follows that the minimal norm solution of $Ax = b$ is

$$z = A^+ b = \begin{bmatrix} -\frac{1}{9} \\ \frac{2}{9} \\ \frac{2}{9}, \end{bmatrix} b.$$

Any other solution of the system $Ax = b$ is necessarily of the form

$$y = A^+ b + v, \quad v \in \ker(A).$$