# Lecture Notes in
# TMA4145 - Linear Methods

*Author:*
Mats Ehrnström

*Compiled by:*
Jon Vegard Venås

August 27, 2013

This is a preliminary PDF version of the online lecture notes that can be found at

`https://wiki.math.ntnu.no/users/ehrnstro/teaching/linearmethods`

# Contents

# Chapter 1

# Sets, spaces and sequences

## 1.1 Sets

### Basic definitions

A **set** is a collection of elements, such as

$$\{1, 2, 3\}, \quad \{a, b, \dagger, \ddagger\}, \quad \text{or} \quad \{\text{all yellow horses}\}.$$

Sets are unordered. Two sets are **equal** if they contain the same elements,

$$\{1, 2, 3\} = \{3, 2, 1\},$$

whence the set containing no elements,

$$\emptyset = \{\}$$

is unique; it is called the **empty set**.

The **cardinality** of a finite set is its number of elements:

$$|\{a, b\}| = 2 \quad \text{and} \quad |\emptyset| = 0.$$

**Ex.** Some well-known infinite sets are the **natural numbers**,[1]

$$\mathbb{N} = \{1, 2, 3, \ldots\},$$

the **integers**,

$$\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\},$$

and the **real**, $\mathbb{R}$, and **complex numbers**, $\mathbb{C}$.

## 1.2 Membership and inclusions

### Membership (possessive relations)

If $x$ is an element in a set $A$ we write

$$x \in A \quad \text{or} \quad A \ni x,$$

and if not

$$x \notin A \quad \text{or} \quad A \not\ni x.$$

[1] In some textbooks also the zero element is included in the set of natural numbers.

**Ex.**

- The **rationals** can be constructed from elements in $\mathbb{Z}$ and $\mathbb{N}$:

$$\mathbb{Q} = \{a/b \colon a \in \mathbb{Z}, b \in \mathbb{N}\}$$

- $\sqrt{2}$ is a real number, but not a rational one:

$$\sqrt{2} \in \mathbb{R}, \qquad \sqrt{2} \notin \mathbb{Q}.$$

### Quantifiers

Quantifiers are used to abbreviate notation. The most important ones are:

- $\forall$ Universal quantifier: 'For any','for all'
- $\exists$ Existential quantifier: 'There exists'
- ! Uniqueness quantifier: 'a unique'

**Ex.** For any real number $x$ there exists a unique real number $-x$ with the property that the sum of $x$ and $-x$ is zero:

$$\forall\, x \in \mathbb{R} \qquad \exists!\ (-x) \in \mathbb{R}; \qquad x + (-x) = 0.$$

### Inclusions

A set $A$ is a **subset** of a set $B$ if any element in $A$ is also an element in $B$:

$$A \subset B \quad (\text{or } A \subseteq B) \quad \overset{\text{def.}}{\Longleftrightarrow} \quad [x \in A \Rightarrow x \in B]$$

A subset $A \subset B$ can also be a **proper subset** of $B$:

$$A \subsetneq B \quad \overset{\text{def.}}{\Longleftrightarrow} \quad A \subset B \text{ but } A \neq B.$$

**Ex.**

- The natural numbers is a subset of the set of non-negative integers, which is a proper subset of the set of real numbers:

$$\mathbb{N} \subset \{0, 1, 2, \ldots\} \subsetneq \mathbb{R}.$$

- The empty set is a subset of any other set (including itself):

$$\emptyset \subset \emptyset \subset \{1, 2, 3\}.$$

- The continuously differentiable real-valued functions on the real line is a subset of the continuous functions:

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

## 1.3 Set operations

### Unions and intersections

The **union** of two sets $A$ and $B$ is the set of elements that are either in A or in B:

$$A \cup B \overset{\text{def.}}{=} \{x \colon x \in A \text{ or } x \in B\}.$$

Their **intersection** is the collection of elements belonging to both A and B:

$$A \cap B \overset{\text{def.}}{=} \{x \colon x \in A \text{ and } x \in B\}.$$

**Ex.**

- For finite sets:

$$\{A, B, C\} \cup \{A, C, D\} = \{A, B, C, D\}, \qquad \{A, B, C\} \cap \{A, C, D\} = \{A, C\}.$$

- For two intervals:

$$(-\infty, 1) \cup (0, \infty) = \mathbb{R}, \qquad (-\infty, 1) \cap (0, \infty) = (0, 1).$$

- For any set $A$,

$$A \cup \emptyset = A, \qquad A \cap \emptyset = \emptyset.$$

### Set differences and complements

The **relative complement** (or **set difference**) of $A$ in $B$ contains any element in $B$ not in $A$:

$$B \setminus A \overset{\text{def.}}{=} \{x \colon x \in B \text{ and } x \notin A\}.$$

When $B$ is understood to be known, this can also be expressed as $\complement(A)$ or $\mathrm{comp}(A)$, the **complement** of $A$ (in $B$).

**Ex.** The complement of the unit ball in three-dimensional Euclidean space is the set of vectors of unit length or larger:

$$\mathrm{comp}(\{x \in \mathbb{R}^3 \colon |x| < 1\}) = \mathbb{R}^3 \setminus \{x \in \mathbb{R}^3 \colon |x| < 1\} = \{x \in \mathbb{R}^3 \colon |x| \geq 1\}.$$

## 1.4 Relations

### Cartesian products

The **Cartesian product** of two sets $A$ and $B$ is the set of **ordered pairs** $(a, b)$ of elements $a \in A$ and $b \in B$:

$$A \times B \overset{\text{def.}}{=} \{(a, b) \colon a \in A, b \in B\}.$$

**Ex.**

- The Cartesian product of $\{1, 2\}$ and $\{\dagger, \ddagger\}$ has four elements:[1]

$$\{1, 2\} \times \{\dagger, \ddagger\} = \{(1, \dagger), (1, \ddagger), (2, \dagger), (2, \ddagger)\}.$$

- The Cartesian product of the set of points on the real line and the set of points in the plane is the set of points in three-dimensional space:

$$\mathbb{R} \times \mathbb{R}^2 = \mathbb{R}^3.$$

[1] In general, $|A \times B| = |A||B|$ for finite sets.

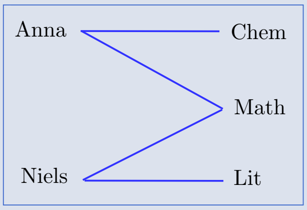
## Relations

A **relation** (or **binary relation**) on two sets $A$ and $B$ is a subset $G$ of $A \times B$:

$$G = \{(a,b) \in A \times B \colon a \text{ satisfying some criteria}, b \text{ satisfying some criteria}\}$$

The set $A$ is called the relation's **domain**, $B$ its **codomain**, and $G$ its **graph**. The graph of a relation can most easily be thought of as connections (**edges**) between 'points' in $A$ and $B$ (**vertices**).

**Ex.**



- Students enlisted for courses at a university is a relation on the set of students and the set of courses. For example,

  $$\{(\text{Anna, Math}), (\text{Anna, Chem}), (\text{Niels, Math}), (\text{Niels, Lit})\}$$

  is the graph of a relation on the domain $\{\text{Anna, Niels}\}$ and the codomain $\{\text{Math, Chem, Lit}\}$.
- Relations can be defined on a product set $A \times A$. For example, '$\leq$' is a relation on $\mathbb{R} \times \mathbb{R}$, whose graph is determined by

$$(a,b) \in G \quad \Longleftrightarrow \quad a \leq b.$$

## Functions

A **function** (or **mapping**) is a relation with the property that for every $a$ in its domain there is a unique $b$ in its codomain such that $(a,b)$ is in the graph.

$$\forall\, a \in A \quad \exists!\, b \in B; \quad (a,b) \in G.$$

To indicate this, one often writes $x, y$ and $X, Y$ instead of $a, b$ and $A, B$. Although functions are completely described by their graphs, it is common to use an extra letter, such as $f$, to express functional relations. One writes

$$f \colon X \to Y, \qquad x \mapsto f(x)$$

or simply

$$y = f(x)$$

to indicate the **argument**, $x$, and **value**, $y$, of a function.[1]

**Ex.**

- The relation with graph

  $$G = \{(x,y) \in \mathbb{R} \times \mathbb{R} \colon y = x^2\}$$

  defines a function $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$.
- The length of a two-vector

  $$|\cdot| \colon \mathbb{R}^2 \to [0, \infty), \qquad (x_1, x_2) \mapsto (x_1^2 + x_2^2)^{1/2}$$

  is a function from the set of vectors in the plane to the set of non-negative real numbers.

[1] Note the difference between the *function*, written $f$, $f(\cdot)$, or $x \mapsto f(x)$, and its *value*, $f(x)$, at a particular point $x$.

## 1.5 Invertibility

### Range and surjectivity

The **range** (or **image**) of a function $f\colon X \to Y$ is the set of elements $y = f(x) \in Y$ in its codomain for which there is an $x \in X$ its domain:

$$\operatorname{ran}(f) \overset{\text{def.}}{=} \{f(x)\colon x \in X\}.$$

A function is **surjective** (or **onto**) if its range equals its codomain,

$$f \text{ surjective} \quad \overset{\text{def.}}{\iff} \quad \operatorname{ran}(f) = Y.$$

This is the same as that, for every $y \in Y$, there is an $x \in X$ with $f(x) = y$.

**Ex.**

- The range of the function

  $$f\colon \mathbb{R} \to \mathbb{R}, \qquad x \mapsto x^2$$

  is $\operatorname{ran}(f) = [0, \infty)$, whence it is *not* surjective.
- Defined differently,

  $$f\colon \mathbb{R} \to [0, \infty), \qquad x \mapsto x^2$$

  *is* surjective.
- The differential operator $\frac{d}{dx}\colon C^1(\mathbb{R}, \mathbb{R}) \to C(\mathbb{R}, \mathbb{R})$ is surjective, since

  $$\text{for any } f \in C(\mathbb{R}, \mathbb{R}) \quad \text{there exists } F = \left[x \mapsto \int_0^x f(t)\,dt\right] \in C^1(\mathbb{R}, \mathbb{R}) \quad \text{such that} \quad \frac{d}{dx}F = f.$$

### Injectivity

A function is **injective** (or **one-to-one**) if different elements in its domain are mapped onto different elements in its codomain,

$$f \text{ injective} \quad \overset{\text{def.}}{\iff} \quad [f(x_1) = f(x_2) \implies x_1 = x_2]$$

Put differently, for any $y \in Y$ there is at most one $x \in X$ with $f(x) = y$.

**Ex.**

- The function

  $$f\colon \mathbb{R} \to \mathbb{R}, \qquad x \mapsto x^2$$

  is not injective, since $x^2 = (-x)^2$.
- The function

  $$f\colon \mathbb{N} \to \mathbb{N}, \qquad n \mapsto 2n$$

  that assigns to each natural number twice its value is injective, since

  $$2m = 2n \quad \implies \quad m = n, \qquad m, n \in \mathbb{N}.$$

- The differential operator $\frac{d}{dx}\colon C^1(\mathbb{R},\mathbb{R}) \to C(\mathbb{R},\mathbb{R})$ is not injective, since, for $f \in C^1(\mathbb{R},\mathbb{R})$,

$$\frac{d}{dx}(f(x)+c) = \frac{d}{dx}f(x) \quad \text{for all} \quad c \in \mathbb{R}.$$

## Invertibility

A function that is both injective and surjective is called **bijective** (or **invertible**). Since its graph covers the entire codomain (surjectivity), and since for each $y \in Y$ there is exactly one $x \in X$ with $f(x) = y$ (injectivity), there exists a function

$$f^{-1}\colon Y \to X, \qquad y \mapsto x,$$

called the **inverse of** $f$. An invertible function satisfies

$$f^{-1}(f(x)) = x \quad \text{for all} \quad x \in X \qquad \text{and} \qquad f(f^{-1}(y)) = y \quad \text{for all} \quad y \in Y,$$

or, shorter,

$$f^{-1} \circ f = \mathrm{id}_X \quad \text{and} \quad f \circ f^{-1} = \mathrm{id}_Y.$$

**N.b.** An injection is always invertible on its range, but not necessarily on the entire codomain.

**Ex.**

- The function $x \mapsto x^2$ is invertible on $[0,\infty)$ (which is easily seen from its graph).
- The map $n \mapsto 2n$, $\mathbb{N} \to 2\mathbb{N}$, is a bijection between the set of positive natural numbers and the set of even numbers. In this sense the cardinality of the set of natural numbers and the cardinality of the set of even numbers are the same.
- The function defined by

$$f(a,b) = a + b\mathrm{i}$$

  is a bijection $\mathbb{R}^2 \to \mathbb{C}$, from the real onto the complex plane.
- One can prove that there is no invertible function from $\mathbb{N}$ to $\mathbb{R}$. In this sense, the cardinality of the real numbers is greater than that of the natural numbers (the natural numbers are said to be **countable**, whereas the real numbers are **uncountable**).
- The differential operator

$$1 - \partial_x^2\colon C^\infty_{2\pi\text{-per}}(\mathbb{R},\mathbb{R}) \to C^\infty_{2\pi\text{-per}}(\mathbb{R},\mathbb{R})$$

  from the set of $2\pi$**-periodic infinitely differentiable real-valued functions** onto itself is a bijection. The operator $1 + \partial_x^2$ on the same set of functions is *not* (can you see why?). This means that the differential equation

$$f'' - f = 0$$

  has exactly one $2\pi$-periodic solution (namely $f \equiv 0$), whereas the equation

$$f'' + f = 0$$

  has many.

- A major question in linear algebra is: When is the matrix $A$ in an equation

$$Ax = b$$

invertible? Here, $A$ is seen as an operator $\mathbb{R}^n \to \mathbb{R}^n$, mapping vectors onto vectors.

## 1.6 Vector spaces

### Definition

A **real vector space** is a set $X$ endowed with an operation called **addition**,

$$X \times X \to X, \qquad (x, y) \mapsto x + y,$$

an operation called **scalar multiplication**,

$$\mathbb{R} \times X \to X, \qquad (\lambda, x) \mapsto \lambda x,$$

an element $\mathbf{0} \in X$ called the **zero vector**, and for each $x \in X$ an **additive inverse** $-x \in X$, such that for any elements $x, y, z \in X$ and real numbers $\lambda, \mu \in \mathbb{R}$ the following properties hold:

$$\begin{array}{lcr}
\text{(i)} & x + \mathbf{0} = x, & \text{(additive identity)} \\
\text{(ii)} & x + (-x) = \mathbf{0}, & \text{(additive inverse)} \\
\text{(iii)} & x + y = y + x, & \text{(symmetry)} \\
\text{(iv)} & x + (y + z) = (x + y) + z, & \text{(associativity)} \\
\\
\text{(v)} & 1x = x, & \text{(multiplicative identity)} \\
\text{(vi)} & \lambda(\mu x) = (\lambda\mu)x, & \text{(compatibility)} \\
\\
\text{(vii)} & \lambda(x + y) = \lambda x + \lambda y, & \text{(distributivity)} \\
\text{(viii)} & (\lambda + \mu)x = \lambda x + \mu x, & \text{(distributivity)}
\end{array}$$

The elements of $X$ are called **vectors**. If the **field of scalars** $\mathbb{R}$ is replaced with $\mathbb{C}$ one obtains instead a **complex vector space**.[1]

**N.b. 1** The notion of a vector space and that of a linear space are identitical.

**N.b. 2** The elements of a real vector space need not be real-valued. It is the *field of scalars* that determines whether a vector space is calld real or complex.

**Ex.**

- $(\mathbb{R}, +, \cdot)$, the set of real numbers $\mathbb{R}$ endowed with the usual addition and multiplication is a real vector space.
- More generally, **Euclidean space**

$$\mathbb{R}^n = \{(x_1, \ldots, x_n) \colon x_j \in \mathbb{R} \text{ for } j = 1, 2, \ldots, n.\}$$

endowed with componentwise addition

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n)$$

[1] It is possible to define a vector space over any field $\mathbb{F}$, but we shall not use this.

and componentwise scalar addition

$$\lambda(x_1, \dots, x_n) = (\lambda x_1, \dots, \lambda x_n)$$

is a real vector space for any natural number $n \in \mathbb{N}$.

- **The set of real-valued continuous functions on an interval** $I \subset \mathbb{R}$,

$$C(I, \mathbb{R}) = \{f \colon I \to \mathbb{R} \text{ such that } f \text{ is continuous}\}$$

is a real vector space with the zero function $f \equiv 0$ as additive identity and $-f$ as additive inverse, when one defines

$$\begin{aligned}(f+g)(t) &:= f(t) + g(t),\\ (\lambda f)(t) &:= \lambda f(t),\\ (-f)(t) &:= -f(t).\end{aligned}$$

- All three examples above can be turned into complex vector spaces by replacing $\mathbb{R}$ with $\mathbb{C}$, i.e., when both the elements in the space and the field of scalars are replaced. These are the spaces

$$\mathbb{C}, \quad \mathbb{C}^n \quad \text{and} \quad C(I, \mathbb{C}).$$

The same spaces can also be considered as real vector spaces if the field of scalars is kept to be $\mathbb{R}$. Often, however, spaces that involve complex numbers are regarded as complex vector spaces.

- The essential property of a vector space is *linearity*: any line

$$\{(x, y) = (r\cos(\theta), r\sin(\theta)) \in \mathbb{R}^2 \colon \theta = \theta_0\}$$

is a vector space (addition and scalar mulitplication as in $\mathbb{R}^2$), whereas a closed ball

$$\{(x, y) = (r\cos(\theta), r\sin(\theta)) \in \mathbb{R}^2 \colon 0 \le r \le \beta\}$$

is not (adding or scaling vectors might get one out of the space).

## 1.7 Normed spaces

### Definition

A **normed space** is a vector space $X$ endowed with a function

$$X \to [0, \infty), \qquad x \mapsto \|x\|,$$

called the **norm** on $X$, which satisfies:

$$\begin{array}{llr}\text{(i)} & \|\lambda x\| = |\lambda|\,\|x\|, & \text{(positive homogeneity)}\\ \text{(ii)} & \|x+y\| \le \|x\| + \|y\|, & \text{(triangle inequality)}\\ \text{(iii)} & \|x\| = 0 \quad \text{if and only if} \quad x = 0, & \text{(positive definiteness)}\end{array}$$

for all scalars $\lambda$ and all elements $x, y \in X$. A vector space may allow for many different norms, but not all vector spaces are normable.[1]

[1] Using the *axiom of choice* it is possible to assign a norm to any vector space, but this norm may not correspond to any natural structure of the space. For example, there is no norm such that $C^\infty(\mathbb{R}, \mathbb{R})$, the set of infinitely differentiable real-valued functions on $\mathbb{R}$, is complete.

**Ex.**

- The vector space $\mathbb{R}^n$ with the usual addition and scalar multiplication allows for several norms, for example:

the **Euclidean norm**

$$\|(x_1,\dots,x_n)\|_{l_2} = \big(x_1^2 + \ldots + x_n^2\big)^{1/2}$$

the **maximum norm**

$$\|(x_1,\dots,x_n)\|_{l_\infty} = \max\{|x_1|,\dots,|x_n|\},$$

and the **summation norm**

$$\|(x_1,\dots,x_n)\|_{l_1} = |x_1| + \ldots + |x_n|.$$

These are all special cases of the (finite-dimensional) $l_p$**-norm** $\|(x_1,\dots,x_n)\|_{l_p} = \big(\sum_{j=1}^n |x_j|^p\big)^{1/p}$, $1 \le p \le \infty$.

**Proof (of the example)**

For both $\|\cdot\|_{l_2}$, $\|\cdot\|_{l_\infty}$ and $\|\cdot\|_{l_1}$, it is clear that they are non-negative functions, and that

$$\|x\| = 0 \quad\Longleftrightarrow\quad x = (x_1,\dots,x_n) = (0,\dots,0).$$

In addition,

$$\|\lambda x\|_{l_2} = \big((\lambda x_1)^2 + \cdots + (\lambda x_n)^2\big)^{1/2} = |\lambda|\big(x_1^2 + \cdots + x_n^2\big)^{1/2} = |\lambda|\|x\|_{l_2},$$

and similarly for $\|\cdot\|_{l_\infty}$ and $\|\cdot\|_{l_1}$.

The **triangle inequality** for $\|\cdot\|_{l_\infty}$ and $\|\cdot\|_{l_1}$ follows from that on $\mathbb{R}$:

$$\|x+y\|_{l_1} = \sum_{j=1}^n |x_j + y_j| \le \sum_{j=1}^n (|x_j| + |y_j|) = \sum_{j=1}^n |x_j| + \sum_{j=1}^n |y_j| = \|x\|_{l_1} + \|y\|_{l_1},$$

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

The triangle inequality for $\|\cdot\|_{l_2}$ is a consequence of the **Cauchy–Schwarz inequality**, which we will prove later.

**Ex.**

- The space of real- (or complex-) valued **bounded and continuous functions** on an interval (open or closed), $BC(I, \mathbb{R})$, becomes a normed vector space when endowed with the **supremum norm**[1]

  $$\|f\|_\infty = \sup_{x\in I} |f(x)|.$$

  If $I = [a, b]$ it follows from the *extreme value theorem* that $BC([a, b], \mathbb{R}) = C([a, b], \mathbb{R})$ (as sets and linear spaces) and

  $$\|f\|_\infty = \sup_{x\in[a,b]} |f(x)| = \max_{x\in[a,b]} |f(x)|.$$

  If $I = (a, b)$ is either infinite or does not contain its end points, then $BC((a, b), \mathbb{R}) \subsetneq C((a, b), \mathbb{R})$. An example of this strict inclusion is the function $x \mapsto 1/x$ on (0,1). It is continuous, but

  $$[x \mapsto 1/x] \notin BC((0, 1), \mathbb{R}),$$

  since $\sup_{x\in(0,1)} |1/x| = \infty$.

### Equivalence of norms

Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a vector space $X$ are said to be **equivalent** if there exists a number $c \in \mathbb{R}$ such that

$$c^{-1}\|x\|_1 \le \|x\|_2 \le c\|x\|_1 \qquad \text{for all} \quad x \in X.$$

**Ex.**

- The maximum and summation norms are equivalent on $\mathbb{R}^n$, since

  $$\max_{1\le j\le n} |x_j| \le \sum_{j=1}^{n} |x_j| \quad \text{and} \quad \sum_{j=1}^{n} |x_j| \le n \max_{1\le j\le n} |x_j|.$$

  Hence

  $$n^{-1}\|x\|_{l_\infty} \le \|x\|_{l_1} \le n\|x\|_{l_\infty} \qquad \text{for} \quad x = (x_1, \dots, x_n).$$

- One can show that, on a *finite-dimensional* vector space, any two norms are equivalent. In particular, any norm on $\mathbb{R}^n$ is equivalent to the Euclidean norm.

## 1.8 Metric spaces

### Definition

Let $X$ be a set and $d\colon X \times X \to [0, \infty)$ a function such that

(i) $d(x, y) = d(y, x)$, (symmetry)

(ii) $d(x, y) \le d(x, z) + d(z, y)$, (triangle inequality)

(iii) $d(x, y) = 0$ if only if $x = y$. (non-degeneracy)

[1] The **supremum** of a set $A \subset \mathbb{R}$ is the smallest $M \in \mathbb{R}$ such that $a \le M$ for all $a \in A$. If no such finite $M$ exists, then $\sup(A) = \infty$. One furthermore defines $\sup(\emptyset) = -\infty$. Thus, the supremum always exists. In a similar fashion, the **infimimum** of a set $A$ is the largest lower bound on the set; it can be defined as $\inf(A) = -\sup(-A)$, where $-A = \{-a \in \mathbb{R} \colon a \in A\}$.

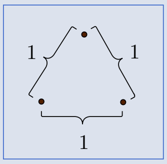
Then the pair $(X, d)$ is called a **metric space** and the function $d$ is called a **metric** or **distance** on $X$.

A subset $M \subset X$ is called a **subspace** of $X$, written $(M, d) \subset (X, d)$, if $M$ is endowed with the same metric as $X$, called the **induced metric** on $M$. Subspaces of metric spaces are themselves metric spaces.

**Ex.**

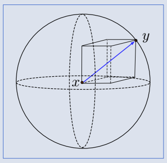
- Any set becomes a metric space when endowed with the **discrete metric**

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



- $\mathbb{R}^n$ becomes a metric space when endowed with the **Euclidean distance**

$$d(x, y) := |x - y| = \left((x_1 - y_1)^2 + \ldots + (x_n - y_n)^2\right)^{1/2}.$$



- The supremum norm induces a metric on the set of bounded and continuous functions on an interval:

$$d(f, g) = \|f - g\|_\infty = \sup_{x \in I} |f(x) - g(x)|.$$

This makes $(BC(I, \mathbb{R}), \|\cdot\|_\infty)$ a metric space.

### ℘ Normed spaces are metric spaces

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

**Proof**

The distance is non-negative and well defined, since

$$0 \leq \underbrace{\|x - y\|}_{d(x,y)} \leq \|x\| + \|y\| < \infty, \qquad \text{for} \qquad x, y \in (X, \|\cdot\|).$$

Symmetry: $d(x, y) = \|x - y\| = \|y - x\| = d(y, x)$.

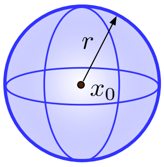
Triangle inequality: $d(x, y) = \|x - y\| \leq \|x - z\| + \|z - y\| = d(x, z) + d(z, y)$.

Non-degeneracy: $d(x, y) = \|x - y\| = 0 \iff x - y = 0 \iff x = y$.

**N.b.** Metric spaces need not be vector spaces. The set of positive real numbers, $\mathbb{R}_+ = (0, \infty)$, with the metric given by $d(x, y) := |x - y|$ is a metric space, but it is not a linear space, since it contains neither an additive identity $(0)$ nor additive inverses $(-x)$.

## 1.9 Balls and spheres

Let $(X, d)$ be a metric space with distance $d\colon X \times X \to [0, \infty)$.



Two important concepts are the **ball of radius** $r > 0$ **centered at** $x_0 \in X$,

$$B_r(x_0) := \{x \in X : d(x, x_0) < r\},$$

and the **sphere of radius** $r > 0$ **centered at** $x_0 \in X$,

$$S_r(x_0) := \{x \in X : d(x, x_0) = r\}.$$

For normed spaces, or other vector spaces that are also metric spaces, we simply write

$$B_r := B_r(0) \quad \text{and} \quad S_r = S_r(0),$$

for balls and spheres centered at the origin (zero element). The sets $B_1$ and $S_1$ are called the **unit ball** and **unit sphere**, respectively.

**Ex.**

- The ball of radius 2 centered at $(1, 0)$ in Euclidean space $\mathbb{R}^2$:

$$B_2((1, 0)) = \{(x, y) \in \mathbb{R}^2 : (x - 1)^2 + y^2 < 4\}.$$

- **Sequence spaces** are spaces in which each element

$$x = \{x_n\}_{n\in\mathbb{N}} = (x_1, x_2, \ldots)$$

  is a sequence (usually of real or complex numbers). The most important ones are the so-called $l_p$-spaces:

  - Let $l_\infty$ be the space of sequences $\{x_j\}_{j\in\mathbb{N}} = (x_1, x_2, \ldots)$ for which

$$\|x\|_{l_\infty} = \sup_{j\in\mathbb{N}} |x_j| < \infty.$$

    Then the sequence $(1/2, 2/3, 3/4, \ldots) \in S_1$ in $l_\infty$ (since $\sup_{j\in\mathbb{N}} |\frac{j}{j+1}| = 1$).

  - For any $p \geq 1$, let $l_p$ be the space of sequences $\{x_j\}_{j\in\mathbb{N}} = (x_1, x_2, \ldots)$ for which

$$\|x\|_{l_p} = \Big(\sum_{j\in\mathbb{N}} |x_j|^p\Big)^{1/p} < \infty.$$

    Let further

$$e_1 = (1, 0, 0, \ldots), \quad e_2 = (0, 1, 0, 0 \ldots) \quad \text{and} \quad e_j = (\ldots, 0, 1, 0, \ldots), \quad j \in \mathbb{N}.$$

    Then

$$e_j \in S_1 \quad \text{for all} \quad j \in \mathbb{N},$$

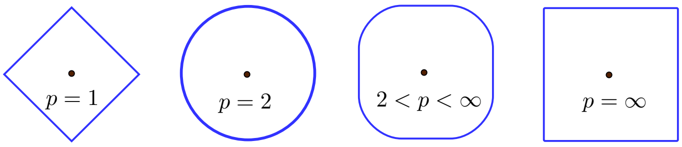
    but

$$d(e_i, e_j) = \|e_i - e_j\|_{l_p} = (|1|^p + |-1|^p)^{1/p} = 2^{1/p} \geq 1 \quad \text{whenever} \quad i \neq j.$$

Note that such a sequence of elements could never exist in $\mathbb{R}^n$ (or any other finite-dimensional vector space).

- The unit ball in $BC([0,1],\mathbb{R})$ consists of all functions whose graph $y = f(x)$ lies strictly between the lines $y = \pm 1$.

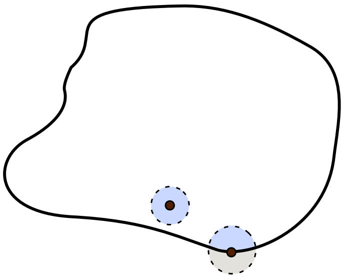
**N.b.** The unit ball may look quite different depending on the underlying metric/norm. The following illustration captures this in the case of the $l_p$-norm on $\mathbb{R}^2$. Homogeneity and the triangle inequality however imply that a ball in any metric given by a norm will always be a convex set in the underlying space.



The unit sphere for different metrics: $\|x\|_{l_p} = 1$ in $\mathbb{R}^2$.

## 1.10 Interior points, boundary points, open and closed sets

Let $(X, d)$ be a metric space with distance $d\colon X \times X \to [0, \infty)$.



- A point $x_0 \in D \subset X$ is called an **interior point in D** if there is a small ball centered at $x_0$ that lies entirely in $D$,

$$x_0 \text{ interior point} \quad \overset{\text{def}}{\iff} \quad \exists\, \varepsilon > 0; \qquad B_\varepsilon(x_0) \subset D.$$

- A point $x_0 \in X$ is called a **boundary point of D** if any small ball centered at $x_0$ has non-empty intersections with both $D$ and its complement,

$$x_0 \text{ boundary point} \quad \overset{\text{def}}{\iff} \quad \forall\, \varepsilon > 0 \quad \exists\, x, y \in B_\varepsilon(x_0); \quad x \in D,\ y \in X \setminus D.$$

- The set of interior points in D constitutes its **interior**, $\text{int}(D)$, and the set of boundary points its **boundary**, $\partial D$. $D$ is said to be **open** if any point in $D$ is an interior point and it is **closed** if its boundary $\partial D$ is contained in $D$; the **closure of D** is the union of $D$ and its boundary:

$$\overline{D} := D \cup \partial D.$$

Alternative notations for the closue of $D$ in $X$ include $\overline{D}^X$, $\text{clos}(D)$ and $\text{clos}(D; X)$.[1]

**Ex.**

- In $\mathbb{R}$ with the usual distance $d(x, y) = |x - y|$, the interval $(0, 1)$ is open, $[0, 1)$ neither open nor closed, and $[0, 1]$ closed.[2]

[1] An alternative to this approach is to take closed sets as complements of open sets. These two definitions, however, are completely equivalent. In particular, a set is open exactly when it does not contain its boundary.

[2] Equivalent norms induce the same **topology** on a space (i.e., the same open and closed sets). Since all norms on $\mathbb{R}^n$ are equivalent, it is unimportant which norm we choose.

- The set

$$D := \{(x,y) \in \mathbb{R}^2 \colon x > 0, y \geq 0\}$$

is neither closed nor open in Euclidean space $\mathbb{R}^2$ (metric coming from a norm, e.g., $d(x,y) = \|x-y\|_{l_2} = ((x_1-y_1)^2 + (x_2-y_2)^2)^{1/2}$), since its boundary contains both points $(x,0)$, $x > 0$, in $D$ and points $(0,y)$, $y \geq 0$, not in $D$. The closure of D is

$$\overline{D} = \{(x,y) \in \mathbb{R}^2 \colon x \geq 0, y \geq 0\}.$$

- An entire metric space is both open and closed (its boundary is empty).
- In $l_\infty$,

$$B_1 \not\ni (1/2, 2/3, 3/4, \ldots) \in \overline{B_1}.$$

- For a general metric space, the **closed ball**

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

may be larger than the closure of a ball, $\overline{B_r(x_0)}$. If we let $X$ be a space with the discrete metric,

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

Then

$$B_1(x_0) = \{x_0\}, \qquad \text{so that} \qquad \overline{B_1(x_0)} = \overline{\{x_0\}} = \{x_0\}.$$

But

$$\tilde{B}_1(x_0) = X.$$

$\wp$ **(Open) balls are open**

Let $(X,d)$ be a metric space, $x_0$ a point in $X$, and $r > 0$. Then $B_r(x_0)$ is open in $X$ with respect to the metric $d$.

**Proof**

Pick $x \in B_r(x_0)$. Then

$$\begin{aligned} d(x,x_0) < r \quad &\Longrightarrow \quad \exists\, \varepsilon > 0; \quad d(x,x_0) < r - \varepsilon \\ &\Longrightarrow \quad d(y,x) < \varepsilon \quad \text{implies} \quad d(y,x_0) \leq d(y,x) + d(x,x_0) < \varepsilon + (r-\varepsilon) = r. \end{aligned}$$

This means: $y \in B_r(x_0)$ if $y \in B_\varepsilon(x)$, i.e. $B_\varepsilon(x) \subset B_r(x_0)$.

## 1.11 Limits

Let $(X, d_X)$ and $(Y, d_Y)$ be metric spaces, and $f \colon X \to Y$ a function between them.

### Sequential limits

A sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ is said to **converge towards** $x_0 \in X$ if for any $\varepsilon > 0$ there is a natural number $n_\varepsilon$ with the property that $x_n \in B_\varepsilon(x_0)$ for all $n \geq n_\varepsilon$:

$$\lim_{n\to\infty} x_n = x_0 \quad \overset{\text{def}}{\Longleftrightarrow} \quad \forall\, \varepsilon > 0 \;\; \exists\, n_\varepsilon \in \mathbb{N}; \qquad x_n \in B_\varepsilon(x_0) \quad \text{for} \quad n \geq n_\varepsilon.$$

We then say that $x_n$ tends to $x_0$ as $n$ tends to infinity, written

$$x_n \to x_0 \quad (\text{as} \quad n \to \infty), \qquad \text{or} \qquad x_n \overset{n\to\infty}{\to} x_0.$$

The point $x_0$ is called the **limit** of the sequence $\{x_n\}_{n\in\mathbb{N}}$.

$\wp$ **Sequential limits are zero limits for the distance function**
Since $\{d_X(x_n, x_0)\}_{n\in\mathbb{N}}$ is a sequence in $\mathbb{R}$ it is easily verified that

$$x_n \to x_0 \quad \iff \quad d_X(x_n, x_0) \to 0.$$

**Proof**

$$\begin{aligned} d_X(x_n,x_0) \to 0 \quad &\iff \quad \forall\varepsilon > 0 \;\; \exists n_\varepsilon; \quad d_X(x_n,x_0) \in B_\varepsilon(0) \qquad (\text{in } \mathbb{R}) \quad \text{for} \quad n \geq n_\varepsilon \\ &\iff \quad \forall\varepsilon > 0 \;\; \exists n_\varepsilon; \quad d_X(x_n,x_0) < \varepsilon \qquad\qquad\qquad \text{for} \quad n \geq n_\varepsilon \\ &\iff \quad \forall\varepsilon > 0 \;\; \exists n_\varepsilon; \quad x_n \in B_\varepsilon(x_0) \qquad\quad (\text{in } X) \quad \text{for} \quad n \geq n_\varepsilon \\ &\iff \quad x_n \to x_0. \end{aligned}$$

## Continuous limits (continuity)

We say that $f(x)$ **converges to** $y_0$ **in** $Y$ **as** $x$ **converges to** $x_0$ **in** $X$ if for any $\varepsilon > 0$ there exists $\delta > 0$ such that $f(x) \in B_\varepsilon(y_0)$ when $x \in B_\delta(x_0)$:

$$\lim_{x\to x_0} f(x) = y_0 \quad \overset{\text{def}}{\iff} \quad \forall\,\varepsilon > 0 \quad \exists\,\delta > 0; \qquad [d_X(x,x_0) < \delta \quad \Longrightarrow \quad d_Y(f(x),y_0) < \varepsilon].$$

Equivalent ways of writing this are

$$f(x) \to y_0 \quad \text{as} \quad x \to x_0, \qquad \text{and} \qquad f(x) \overset{x\to x_0}{\to} y_0.$$

A function $f$ satisfying this is said to be **continuous** at the point $x_0$. It is continuous on a set $D$ if it continuous at all points $x_0 \in D$, and simply continuous if its continuous on all of its domain.

**Ex.**

- The function

$$f\colon x \mapsto \frac{\sin(x)}{x}, \qquad x \in \mathbb{R} \setminus \{0\},$$

may be extended to a bounded and continuous function $\mathbb{R} \to \mathbb{R}$, since

$$f(x) \overset{\text{in } \mathbb{R}}{\to} 1 \qquad \text{as} \qquad x \overset{\text{in } \mathbb{R}}{\to} 0.$$

$\wp$ **In metric spaces continuous and sequential limits agree**

$$(i) \quad \lim_{x\to x_0} f(x) = y \quad \iff \quad (ii) \quad \lim_{n\to\infty} f(x_n) = y \text{ for any sequence such that } \lim_{n\to\infty} x_n = x_0.$$

**Proof**

Assume that (i) holds, i.e.,

$$d_Y(f(x), y) < \varepsilon \quad \text{for} \quad d_X(x, x_0) < \delta.$$

For any sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$ with $\lim_{n\to\infty} x_n = x_0$, there exists $N \in \mathbb{N}$ with

$$d_X(x_n, x_0) < \delta \quad \text{for} \quad n \geq N.$$

Thus, according to (i),

$$d_Y(f(x_n), y) < \varepsilon \quad \text{for} \quad n \geq N.$$

This shows that (i) is sufficient for (ii) to hold.

Now assume that (i) does not hold, i.e., there is an $\varepsilon > 0$ such that *for any* $\delta > 0$ there exists $x_\delta$ with

$$d_X(x_\delta, x_0) < \delta \quad \text{while} \quad d_Y(f(x_\delta), y) \geq \varepsilon.$$

Thus, any sequence of $\delta_n \overset{n\to\infty}{\to} 0$ yields a sequence of numbers $x_n := x_{\delta_n}$ with

$$x_n \to x_0 \text{ as } n \to \infty \quad \text{while} \quad d_Y(f(x_n), y) \geq \varepsilon.$$

This violates (ii) and shows that (i) is necessary for (ii) to hold.

**Ex.**

- The sequence of functions given by

$$f_0(x) = 1, \quad f_1(x) = 1 - \frac{x^2}{3!}, \quad f_n(x) = \sum_{j=0}^{n} \frac{(-1)^j x^{2j}}{(2j+1)!} \quad n \in \mathbb{N},$$

converges in $BC([0,1], \mathbb{R})$. Namely, let $f(x) = \frac{\sin(x)}{x}$, extended to a bounded and continuous function on $\mathbb{R}$ as in the preceding example. Since

$$\sin x \overset{\text{Taylor}}{=} x - \frac{x^3}{3!} + \ldots + \frac{(-1)^n x^{2n+1}}{(2n+1)!} \pm \frac{\cos(\xi)\, x^{2n+3}}{(2n+3)!}, \qquad 0 < |\xi| < |x|,$$

we have

$$\frac{\sin x}{x} = \sum_{j=0}^{n} \frac{(-1)^j x^{2j}}{(2j+1)!} \pm \frac{\cos(\xi)\, x^{2n+2}}{(2n+3)!}, \qquad 0 < |\xi| < |x|,$$

so that

$$\begin{aligned} \|f_n - f\|_{BC([0,1],\mathbb{R})} &= \sup_{x\in[0,1]} \left| \sum_{j=0}^{n} \frac{(-1)^j x^{2j}}{(2j+1)!} - \frac{\sin x}{x} \right| \\ &\leq \sup_{x,\xi\in[0,1]} \left| \frac{\cos(\xi)\, x^{2n+2}}{(2n+3)!} \right| = \frac{1}{(2n+3)!} \to 0 \quad \text{as} \quad n \to \infty. \end{aligned}$$

### ℘ Limits are unique

If $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} x_n = y$, then $x = y$.

**Proof**

In view of the assumptions, and using the triangle inequality,

$$0 \leq d(x,y) \leq d(x, x_n) + d(x_n, y) \to 0 \quad \text{as} \quad n \to \infty.$$

Then $d(x,y) = 0$ implies $x = y$ by the axioms of a metric space.

**Ex.**

- The sequence

$$x_1 = 0.9, \quad x_2 = 0.99, \quad x_3 = 0.999, \qquad \text{and so forth,}$$

converges towards $0.999\ldots$ in $\mathbb{R}$, but also towards 1. Hence

$$0.999\ldots = 1.$$

## Accumulation points

A concept related to convergence is that of an **accumulation point** of a subset $M \subset X$:

$$x_0 \text{ accumulation point for } M \quad \overset{\text{def}}{\Longleftrightarrow} \quad \exists \{x_n\}_{n\in\mathbb{N}} \subset M; \quad x_n \to x_0.$$

Equivalently, any small ball $B_\varepsilon(x_0)$ centered at $x_0$ contains a point in $M$.

**N.b.** The limit of a sequence is always an accumulation point for that sequence, but a (non-convergent) sequence may have several, or no, accumulation points. An accumulation point for a sequence is, per definition, the limit of a subsequence of that sequence.

**Ex.**

- 0 is an accumulation point for $\{1/n\}_{n\in\mathbb{N}}$, but also for $\{1, 1, 2, \frac{1}{2}, 3, \frac{1}{3}, 4, \frac{1}{4}, \ldots\}$.

## Relationship between limits and closures

### $\wp$ Closures are the total of sequential limits of interior points

By comparing the definitions of boundary and interior points with that of a sequential limit, one obtains that

$$\operatorname{clos}(D; X) = \{x \in X \colon x = \lim_{n\to\infty} x_n \text{ for some sequence } \{x_n\}_{n\in\mathbb{N}} \subset D\}$$

**Proof**

Assume that $x \in \overline{D}$. Then, according to the definitions of interior and boundary points, any small ball $B_{1/n}(x)$ contains a point $x_n \in D$. This means that $\{x_n\}_{n\in\mathbb{N}} \subset D$ converges to $x$.

Now, assume instead that there is a sequence

$$\{x_n\}_{n\in\mathbb{N}} \subset D \quad \text{with} \quad \lim_{n\to\infty} x_n = x \quad \text{in } X.$$

Then $d_X(x_n, x) \to 0$ as $n \to \infty$, so that

$$\forall\, \varepsilon > 0 \quad \exists\, x_{n_\varepsilon} \in B_\varepsilon(x).$$

Since $x_{n_\varepsilon} \in D$, either $x$ is an interior point (for small $\varepsilon$ there are only points from $D$ in $B_\varepsilon(x)$), or $x$ is a boundary point ($B_\varepsilon(x)$ contains also points from the complement of $D$); in any case $x \in \overline{D}$.

**Ex.**

- In $l_\infty$,

$$(\tfrac{1}{2}, \tfrac{2}{3}, \tfrac{3}{4}, \ldots) \in \overline{B_1},$$

since it is the limit of the sequence $x_1 = (\frac{1}{2}, \frac{1}{2}, \ldots)$, $x_2 = (\frac{1}{2}, \frac{2}{3}, \frac{2}{3}, \ldots)$, $x_3 = (\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{3}{4}, \ldots)$, and so forth (the elements of which are all in $B_1$ in $l_\infty$). To see this, let $x_0 := \{\frac{j}{j+1}\}_{j\geq 1}$.

Then

$$\|x_0 - x_n\|_{l_\infty} = \sup_{j\geq n}\left|\frac{j}{j+1} - \frac{n}{n+1}\right| = \left|1 - \frac{n}{n+1}\right| = \frac{1}{n+1} \to 0 \quad \text{as} \quad n\to\infty,$$

in view of that the function $j \mapsto \frac{j}{j+1}$ is monotone increasing and bounded by 1. Thus $B_1 \ni x_n \overset{\text{in } l_\infty}{\to} (\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots)$ as $n \to \infty$.

## 1.12 Completeness

### Cauchy sequences

A sequence $\{x_n\}$ in a metric space $(X, d)$ is a **Cauchy sequence** (or simply **Cauchy**) if the distance between its members tends to zero:

$$\{x_n\} \text{ Cauchy} \quad \overset{\text{def}}{\Longleftrightarrow} \quad d(x_n, x_m) \to 0 \quad \text{as} \quad m, n \to \infty.$$

Equivalently,

$$\forall\, \varepsilon > 0 \quad \exists\, n_\varepsilon; \qquad d(x_m, x_n) < \varepsilon \quad \text{whenever} \quad m, n \geq n_\varepsilon.$$

**Ex.**

- The sequence $\{x_n\}_{n\geq 1}$ of rational numbers $x_n = \sum_{k=0}^{n} \frac{1}{k!}$ is a Cauchy sequence with respect to the distance $d(x,y) = |x-y|$. For each $m \geq n \geq 1$,

$$|x_m - x_n| = \sum_{k=n+1}^{m} \frac{1}{k!} \leq \frac{1}{(n+1)!} \sum_{k=0}^{m-(n+1)} \frac{1}{(n+1)^k}$$
$$= \frac{1}{(n+1)!} \frac{1 - (\frac{1}{n+1})^{(m-n)}}{1 - \frac{1}{n+1}} \overset{n\geq 1}{\leq} \frac{2}{(n+1)!} \to 0 \quad \text{as} \quad m \geq n \to \infty.$$

  Note, however, that $\lim_{n\to\infty} x_n = e \notin \mathbb{Q}$.

- The sequence $\{x_n\}_{n\geq 1}$ of functions $x_n : t \mapsto \sum_{k=0}^{n} \frac{t^k}{k!}$ is Cauchy in $BC([0,1], \mathbb{R})$. For each $m \geq n \geq 1$,

$$\|x_m - x_n\|_{BC([0,1],\mathbb{R})} = \sup_{t\in[0,1]} \left| \sum_{k=n+1}^{m} \frac{t^k}{k!} \right| \leq \sum_{k=n+1}^{m} \frac{1}{k!} \to 0 \quad \text{as} \quad m \geq n \to \infty.$$

  In this case $\lim_{n\to\infty} x_n = [t \mapsto e^t] \in BC([0,1], \mathbb{R})$.[1]

### ℘ Convergent sequences are Cauchy sequences

In any metric space

$$x_n \to x \;\text{ as }\; n \to \infty \qquad \text{implies} \qquad d(x_m, x_n) \to 0 \;\text{ as }\; m, n \to \infty.$$

**N.b.** The opposite is not true in general.

[1] The exponential function $t \mapsto e^t$ is often expressed as exp to separate it from the real number $e = \exp(1)$.

**Proof**

Let $\varepsilon > 0$. Since $x_n \to x$ there exists $n_\varepsilon$ such that

$$d(x, x_n) < \varepsilon/2 \quad \text{for} \quad n \geq n_\varepsilon.$$

Hence

$$d(x_m, x_n) \leq d(x_m, x) + d(x, x_n) < \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

for $m, n \geq n_\varepsilon$.

$\wp$ **Cauchy sequences are bounded**

If $(X, \|\cdot\|)$ is a normed space, and $\{x_n\}_n \subset X$ is Cauchy, then

$$\sup_{n\geq 1} \|x_n\| < B \quad \text{for some} \quad B \in \mathbb{R}.$$

**Proof**

According to the definition of a Cauchy sequence, there exists $n_1$ such that

$$\|x_m - x_n\| < 1 \quad \text{for} \quad m, n \geq n_1.$$

If $n \leq n_1$,

$$\|x_n\| \leq \max_{1\leq n\leq n_1} \|x_n\| =: \tilde{B},$$

and, if $n \geq n_1$,

$$\|x_n\| \leq \|x_n - x_{n_1}\| + \|x_{n_1}\| < 1 + \|x_{n_1}\|.$$

Hence, $\|x_n\| \leq B := \max\{\tilde{B}, 1 + \|x_{n_1}\|\}$ for all $n \in \mathbb{N}$.

## Complete metric spaces and Banach spaces

A metric space in which every Cauchy sequence converges is called **complete**. A complete normed space is called a **Banach space**.

**Ex.**

- $\mathbb{R}$ is complete with respect to the metric $d(x, y) = |x - y|$. Hence, $(\mathbb{R}, |\cdot|)$ is a Banach space.
- Both $\mathbb{R}^n$ and $\mathbb{C}^n$ are Banach spaces with respect to the norm $\|x\|_{l_2} = (\sum_{j=1}^n |x_j|^2)^{1/2}$.[1]
- The space of square-summable sequences

$$l_2 = \Big\{\{x_j\}_{j\geq 1} \colon \sum_{j=1}^{\infty} |x_j|^2 < \infty\Big\}$$

is complete with respect to the norm

$$\|x\|_{l_2} = \Big(\sum_{j=1}^{\infty} |x_j|^2\Big)^{1/2}.$$

So is $l_\infty$, and $l_p$, for any $p \geq 1$.

[1] Note that $|z_j|^2 = a_j^2 + b_j^2$ for complex numbers $x_j = a_j + ib_j$.

- For any interval $I \subset \mathbb{R}$, $BC(I,\mathbb{R})$ and $BC(I,\mathbb{C})$, with norms given by $\|x\| = \sup_{t\in I} |x(t)|$, are Banach spaces.

### ℘ Subsets of complete metric spaces are complete if and only if they are closed

Let $M \subset X$ be a subset of a complete metric space $(X,d)$, i.e. $(M,d) \subset (X,d)$. Then

$$M \text{ complete} \quad \iff \quad M \text{ closed}.$$

**Proof**

Assume that $M$ is complete. Closedness means

$$M \ni x_n \to x \in X \quad \implies \quad x \in M.$$

So assume that $\{x_n\} \subset M$ converges towards $x \in X$. Since any convergent sequence is Cauchy, $\{x_n\} \subset M$ is Cauchy. But since $M$ is complete, $\{x_n\}$ converges to an element $y \in M$. By uniqueness of limits, $y = x$. Thus $M$ is closed.

Contrariwise, assume that $M$ is closed and let $\{x_n\} \subset M$ be a Cauchy sequence. Recall that $(M,d) \subset (X,d)$ carry the induced metric $d$. Thus

$$\begin{aligned} \{x_n\} \text{ Cauchy in } M \quad &\implies \quad \{x_n\} \text{ Cauchy in } X \\ &\overset{X \text{ complete}}{\implies} \quad M \ni x_n \to x \in X \\ &\overset{M \text{ closed}}{\implies} \quad x \in M. \end{aligned}$$

Thus $M$ is complete.

**Ex.**

- Euclidean space $\mathbb{R}^n$ can be viewed as a subspace of $l_2$:

$$\mathbb{R}^n = \{(x_1,\ldots,x_n,0,0\ldots) \in l_2\}.$$

  In this respect, $\mathbb{R}^n$ is both a complete and a closed subspace of $l_2$ (any limit of points in $\mathbb{R}^n$ remains in $\mathbb{R}^n$).

### ℘ BC is complete

Let $I \subset \mathbb{R}$ be a non-empty interval. Then $BC(I,\mathbb{R})$ and $BC(I,\mathbb{C})$ are Banach spaces.

**Proof**

The proof is the same for $\mathbb{R}$ and $\mathbb{C}$. Also, we already know that $BC(I,\mathbb{R})$ is a normed and linear space, so we only need to show that it is complete.

Let $\{x_n\}_n$ be a Cauchy sequence in $BC(I,\mathbb{R})$. We want to prove that it converges to a limit function $x_0 \in BC(I,\mathbb{R})$.

**Pointwise convergence:** For any $t \in I$,

$$|x_n(t) - x_m(t)| \le \sup_{t\in I} |x_n(t) - x_m(t)| = \|x_n - x_m\|_{BC(I,\mathbb{R})}$$

Thus

$$\{x_n\}_n \text{ Cauchy in } BC(I, \mathbb{R}) \quad \Longrightarrow \quad \{x_n(t)\}_n \text{ Cauchy in } \mathbb{R}.$$

$\mathbb{R}$ being complete, there exists a limit in $\mathbb{R}$:

$$\forall\, t \in I \quad \exists\, x_0(t) := \lim_{n\to\infty} x_n(t) \qquad \text{in } \mathbb{R}.$$

Define a function $x_0$ by $x_0 := [t \mapsto x_0(t)]$.

**Boundedness**: For each $t \in I$ there exists $n_t \in \mathbb{N}$ such that $|x_0(t) - x_{n_t}(t)| < \varepsilon$. Thus

$$|x_0(t)| \leq |x_0(t) - x_{n_t}(t)| + |x_{n_t}(t)| < \varepsilon + \|x_{n_t}\|_{BC(I,\mathbb{R})} \leq \varepsilon + \sup_{n\in\mathbb{N}} \|x_n\|_{BC(I,\mathbb{R})} < C,$$

since Cauchy sequences are bounded. Taking the supremum over all $t \in I$ yields that

$$\|x_0\|_{BC(I,\mathbb{R})} < C.$$

**Convergence in norm**: A similar argument shows that $x_n \to x_0$ in $BC(I, \mathbb{R})$. Let $\varepsilon > 0$. For any $t \in I$ there exists $m_t \in \mathbb{N}$ such that

$$|x_0(t) - x_m(t)| < \frac{\varepsilon}{2} \qquad \text{for} \qquad m \geq m_t.$$

Also, there exists $n_\varepsilon \in \mathbb{N}$ such that

$$\|x_n - x_m\|_{BC(I,\mathbb{R})} < \frac{\varepsilon}{2} \qquad \text{for} \qquad m, n \geq n_\varepsilon.$$

Choose $m \geq \max\{m_t, n_\varepsilon\}$. Then

$$|x_n(t) - x_0(t)| \leq |x_n(t) - x_m(t)| + |x_m(t) - x_0(t)| \leq \|x_n - x_m\|_{BC(I,\mathbb{R})} + |x_m(t) - x_0(t)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \qquad \text{for} \qquad n \geq$$

Taking the supremum over $t \in I$ yields that

$$\|x_n - x_0\|_{BC(I,\mathbb{R})} < \varepsilon \qquad \text{for} \qquad n \geq n_\varepsilon.$$

Thus $x_n \to x_0$ in $BC(I, \mathbb{R})$.

**Continuity**: To prove that $x_0$ is continuous, pick $t \in I$ and let $\varepsilon > 0$. Since $x_n \to x_0$ in $BC(I, \mathbb{R})$ there exists $n_\varepsilon \in \mathbb{N}$ such that

$$\|x_0 - x_n\|_{BC(I,\mathbb{R})} < \frac{\varepsilon}{3} \qquad \text{for} \qquad n \geq n_\varepsilon.$$

Fix such an $n$. Since $x_n$ is continuous, there exists $\delta := \delta(n, \varepsilon) > 0$ with

$$|x_n(s) - x_n(t)| < \frac{\varepsilon}{3} \qquad \text{for} \qquad |s - t| < \delta.$$

All taken together,

$$\begin{aligned}
|x_0(s) - x_0(t)| &< |x_0(s) - x_n(s)| + |x_n(s) - x_n(t)| + |x_n(t) - x_0(t)| \\
&\leq \|x_0 - x_n\|_{BC(I,\mathbb{R})} + |x_n(s) - x_n(t)| + \|x_n - x_0\|_{BC(I,\mathbb{R})} \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \qquad \text{for} \qquad |s - t| < \delta.
\end{aligned}$$

Hence, $x_0$ is continuous.

This shows that every Cauchy sequence $\{x_n\}_n \subset BC(I,\mathbb{R})$ converges (in $BC(I,\mathbb{R})$) towards an element $x_0 \in BC(I,\mathbb{R})$. Thus $BC(I,\mathbb{R})$ is complete.[1]

### $\wp$ The space of square-summable sequences is complete

The space of square-summable (complex or real) sequences is a Banach space with respect to the norm $\|x\|_{l_2} = (\sum_{j\geq 1} |x_j|^2)^{1/2}$.

**Proof**

This time we pursue the proof for complex-valued sequences (it is no different from the proof for real-valued sequences).

**Pointwise convergence**: Take $\{x_n\}_n$ Cauchy in $l_2$, with $x_n = (x_n(1), x_n(2), \ldots)$. Then

$$|x_n(j) - x_m(j)| \leq \Big(\sum_{j=1}^{\infty} |x_n(j) - x_m(j)|^2\Big)^{1/2} = \|x_n - x_m\|_{l_2} \to 0 \quad \text{as} \quad m,n \to \infty.$$

Thus, for each $j \in \mathbb{N}$, $\{x_n(j)\}_n$ is Cauchy in $\mathbb{C}$, and thus convergent:

$$\forall\, j \in \mathbb{N} \quad \exists\, x_0(j) = \lim_{n\to\infty} x_n(j) \in \mathbb{C}.$$

Let $x_0 := (x_0(1), x_0(2), \ldots)$.

**Boundedness**: For any $j \in \mathbb{N}$ there exists $n_j \in \mathbb{N}$ such that

$$|x_0(j) - x_n(j)|^2 < \frac{1}{2^j} \quad \text{for } n \geq n_j.$$

For any finite $N \in \mathbb{N}$, choose $n \geq \max_{1\leq j\leq N} n_j$. Then

$$\begin{aligned}\Big(\sum_{j=1}^{N} |x_0(j)|^2\Big)^{1/2} &\leq \Big(\sum_{j=1}^{N} |x_0(j) - x_n(j)|^2\Big)^{1/2} + \Big(\sum_{j=1}^{N} |x_n(j)|^2\Big)^{1/2} \\ &< \Big(\sum_{j=1}^{N} \frac{1}{2^j}\Big)^{1/2} + \|x_n\|_{l_2} \leq 1 + \sup_{n\in\mathbb{N}} \|x_n\|_{l_2} < c,\end{aligned}$$

since Cauchy sequences are bounded. The right-hand side is independent of $N$, so we may now let $N \to \infty$, yielding that

$$\|x_0\|_{l_2} < c, \quad \text{whence} \quad x_0 \in l_2.$$

**Convergence in norm**: Let $\varepsilon > 0$. As above, we can find $n_j$ such that

$$|x_0(j) - x_m(j)|^2 < \frac{\varepsilon/2}{2^j} \quad \text{for } m \geq n_j,$$

and

$$\sum_{j=1}^{N} |x_0(j) - x_m(j)|^2 < \sum_{j=1}^{N} \frac{\varepsilon/2}{2^j} \leq \varepsilon/2 \quad \text{for} \quad m \geq \max_{1\leq j\leq N} n_j.$$

[1] The same proof can be used to show that the **space of bounded and uniformly continuous functions** $BUC(I,\mathbb{R})$ is complete. Since continuous functions are uniformly continuous on compact intervals, $BUC([a,b],\mathbb{R}) = BC([a,b],\mathbb{R})$ for any compact interval $[a,b] \subset \mathbb{R}$.

Using this,

$$\Big(\sum_{j=1}^{N} |x_0(j) - x_n(j)|^2\Big)^{1/2} \leq \Big(\sum_{j=1}^{N} |x_0(j) - x_m(j)|^2\Big)^{1/2} + \Big(\sum_{j=1}^{N} |x_m(j) - x_n(j)|^2\Big)^{1/2} \leq \varepsilon/2 + \|x_n - x_m\|_{l_2}.$$

Since $\{x_n\}_n$ is Cauchy in $l_2$ there exists $n_\varepsilon \in \mathbb{N}$ such that

$$\|x_n - x_m\|_{l_2} < \varepsilon/2 \quad \text{for} \quad m, n \geq n_\varepsilon.$$

Select $m$ such that this holds (for example, $m \geq n_\varepsilon + \max_{1\leq j\leq N} n_j$). Then

$$\left(\sum_{j=1}^{N} |x_0(j) - x_n(j)|^2\right)^{1/2} < \varepsilon \quad \text{for} \quad n \geq n_\varepsilon.$$

Since $n_\varepsilon$ does not depend on $N$, we may let $N \to \infty$, to obtain that

$$\|x_0 - x_n\|_{l_2} < \varepsilon \quad \text{for} \quad n \geq n_\varepsilon.$$

Hence, $x_n \overset{\text{in } l_2}{\to} x_0$, and $l_2$ is complete.

## 1.13 Completions

Every metric space (and every normed space) can be made complete. To make this precise, recall that an **isomorphism** is a bijective (on-to-one and onto) map that preserves the essential structure of something. Likewise, an **isometry** is a map that preserves distances.[1]

### Isometries

Two metric spaces $(X, d_X)$ and $(Y, d_Y)$ are called **isometric** if there exists a bijective isometry between them, i.e., if there exists an invertible function $\varphi\colon X \to Y$ such that

$$d_X(x_1, x_2) = d_Y(\varphi(x_1), \varphi(x_2)).$$

The function $\varphi$ is called an **isometry**.

**Ex.**

- The set of sequences with only zeros and ones,

  $$X = \{(x_1, x_2, \ldots) \in l_\infty \colon \text{ for each } j,\ x_j = 0 \text{ or } x_j = 1\},$$

  endowed with the $l_\infty$-metric,

  $$d(x, y) = \sup_{j\in\mathbb{N}} |x_j - y_j|$$

  is isometric to $X$ endowed with the discrete metric, because

  $$d(x, y) = 1 \quad \text{unless} \quad x = y.$$

  The isometry is the identity operator, $\varphi\colon x \mapsto x, (X, \|\cdot\|_{l_\infty}) \to (X, d_{\text{discrete}})$.

[1] The word *isos* is Greek for ’same’, ’similar’; *morphe* is ’shape’,’form’; and *metron* is ’measure’.

- Let $a < b$. The normed spaces $BC((0,1),\mathbb{R})$ and $BC((a,b),\mathbb{R})$ of bounded continuous functions on the intervals $(0,1)$ and $(a,b)$, respectively, are isometric, since

$$\varphi\colon BC((0,1),\mathbb{R}) \to BC((a,b),\mathbb{R}), \qquad f(\cdot) \overset{\varphi}{\mapsto} f\Big(\frac{\cdot - a}{b-a}\Big)$$

is an isometry:

$$d_{BC((0,1),\mathbb{R})}(f,g) = \sup_{x\in(0,1)} |f(x)-g(x)| = \sup_{x\in(a,b)} \Big|f\big(\frac{x-a}{b-a}\big) - g\big(\frac{x-a}{b-a}\big)\Big| = d_{BC((a,b),\mathbb{R})}(f,g).$$

(Note that $\varphi$ is invertible with inverse $\varphi^{-1}\colon f(\cdot) \mapsto f(a + \cdot(b-a))$. Hence, studying the metric space $BC((a,b),\mathbb{R})$ is no different from studying $BC((0,1),\mathbb{R})$.

## Isomorphisms

### Vector space isomorphisms

A **vector space isomorphism** is a bijective linear map between two vector spaces, i.e. an invertible function $T\colon X \to Y$ such that

$$T(x+y) = Tx + Ty \quad \text{and} \quad T(\lambda x) = \lambda Tx \qquad \text{for all} \quad x, y \in X,\ \lambda \in \mathbb{R} \text{ (or } \mathbb{C}).$$

Two vector spaces which allow for such a mapping are called **isomorphic**, and we write

$$X \cong Y \quad \overset{\text{def}}{\Longleftrightarrow} \quad \exists \text{ isomorphism } T\colon X \to Y.$$

**N.b.** A set which is the image of a vector space isomorphism automatically becomes a vector space (it inherits its linear structure from $X$).

**Ex.**

- Regarded as a real vector space, the space $\mathbb{C}^n$ of complex $n$-tuples,

$$z = (z_1, \ldots, z_n), \qquad z_1, \ldots, z_n \in \mathbb{C}$$

is isomorphic to Euclidean space $\mathbb{R}^{2n}$ via the isomorphism[1]

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

- **The set of polynomials with real coefficients of degree at most** $n$, $P_n(\mathbb{R})$, is a vector space consisting of elements

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0, \qquad a_0, \ldots, a_n \in \mathbb{R}.$$

Why is this a vector space? Because $P_n(\mathbb{R}) \cong \mathbb{R}^{n+1}$ : The mapping

$$T\colon P_n(\mathbb{R}) \to \mathbb{R}^{n+1}, \qquad a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 \mapsto (a_0, a_1, \ldots, a_n)$$

is both *bijective*,

$$\text{for any } \ (a_0, \ldots, a_n) \in \mathbb{R}^{n+1} \ \text{ there exists a unique } \ p(x) = \sum_{k=0}^{n} a_k x^k \in P_n(\mathbb{R}),$$

[1] This is also an isometry, since $|z|^2 = |x|^2 + |y|^2$.

and *linear*,

$$\begin{aligned}
T\Big(\lambda\sum_{k=0}^{n}a_kx^k+\mu\sum_{k=0}^{n}b_kx^k\Big)&=T\Big(\sum_{k=0}^{n}(\lambda a_k+\mu b_k)x^k\Big)\\
&=(\lambda a_0+\mu b_0,\dots,\lambda a_n+\mu b_n)\\
&=\lambda(a_0,\dots,a_n)+\mu(b_0,\dots,b_n)\\
&=\lambda T\Big(\sum_{k=0}^{n}a_kx^k\Big)+\mu T\Big(\sum_{k=0}^{n}b_kx^k\Big).
\end{aligned}$$

Hence any linear operation in $P_n(\mathbb{R})$ corresponds to a linear operation in $\mathbb{R}^{n+1}$, meaning that $P_n(\mathbb{R})$ and $\mathbb{R}^{n+1}$ are isomorphic as vector spaces.

**Isomorphisms on normed spaces**

If the vector spaces are normed, they are **isomorphic as normed spaces** if they are isomorphic (as vector spaces) and isometric (as metric spaces). Sometimes this is called **isometrically isomorphic** to avoid confusion.

**Ex.**

- The spaces $BC((0,1),\mathbb{R})$ and $BC((a,b),\mathbb{R})$ are isometrically isomorphic, since the isometry $\varphi$ in the example above is also a vector space isomorphism:

$$\varphi(\lambda f+\mu g)(x)=(\lambda f+\mu g)\Big(\frac{x-a}{b-a}\Big)=\lambda f\Big(\frac{x-a}{b-a}\Big)+\mu g\Big(\frac{x-a}{b-a}\Big)=\lambda\varphi(f)(x)+\mu\varphi(g)(x),\qquad x\in(a,b).$$

**Embeddings**

If a (normed) vector space $X$ is isomorphic to a subspace $M\subset Y$ of another (normed) vector space, we say that it is **(continuously) embedded** in $Y$,

$$X\hookrightarrow Y\quad\overset{\text{def}}{\iff}\quad\exists\text{ isomorphism }\;T\colon X\to M\subset Y.$$

To ease terminology we shall use this concept also for isometries between metric spaces.

**Ex.**

- The vector space of polynomials of degree at most 1, $P_1(\mathbb{R})$, is continuously embedded in three-dimensional Euclidean space,

$$P_1\hookrightarrow\mathbb{R}^3,\qquad\text{since}\qquad P_1\cong\mathbb{R}^2\subset\mathbb{R}^3.$$

- 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:[1]

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

## Dense sets

A subset $M\subset X$ of a metric space is **dense** if its closure is the whole space.

$$M\text{dense in}X\quad\overset{\text{def}}{\iff}\quad\overline{M}=X.$$

In this sense, $M$ is 'almost all of $X$'.

[1] Note that, when $I=\mathbb{R}$, or if $I$ is not closed, neither of these spaces are normed. When $I=[a,b]$ is closed and finite, $BC([a,b],\mathbb{R})=(C([a,b],\mathbb{R}),\|\cdot\|_\infty)$. Otherwise, $BC(I,\mathbb{R})$ is strictly smaller than $C(I,\mathbb{R})$.

**Ex.**

- $\mathbb{Q}$ is dense in $\mathbb{R}$:

$$\text{for any } \lambda \in \mathbb{R} \text{ there is a sequence } \{q_n\}_n \subset \mathbb{Q} \quad \text{such that} \quad q_n \to \lambda \quad \text{in } \mathbb{R}.$$

- **Stone-Weierstrass**[1]: Let $I = [a, b]$ be a finite and closed interval. The polynomials, $P(\mathbb{R})$, the **infinitely continuosly differentiable functions**, $C^\infty(I, \mathbb{R})$, and the $k$ times continuously differentiable functions, $C^k(I, \mathbb{R})$, $k \geq 1$, are all dense in the space of bounded and continuous functions $BC(I, \mathbb{R})$:

$$\forall\, \varepsilon > 0,\ f \in BC(I, \mathbb{R}) \quad \exists\, f_{\text{approx}} \in P(\mathbb{R}) \subset C^\infty(I, \mathbb{R}) \subset C^k(I, \mathbb{R}); \quad \sup_{x\in I} |f(x) - f_{\text{approx}}(x)| < \varepsilon.$$

**Separability**

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

$$X \text{ separable} \quad \overset{\text{def}}{\Longleftrightarrow} \quad \exists \{x_n\}_{n\in\mathbb{N}} \subset X; \qquad \overline{\{x_n\}_n} = X.$$

**Ex.**

- Since $\mathbb{Q}$ is countable, and $\overline{\mathbb{Q}} = \mathbb{R}$ (with respect to the distance $d(x, y) = |x - y|$), it follows that $(\mathbb{R}, |\cdot|)$ is separable.
- Using that $\overline{\mathbb{Q}} = \mathbb{R}$ one can show that all the spaces $\mathbb{R}^n, \mathbb{C}^n, l_p(\mathbb{R})$ and $l_p(\mathbb{C})$ for $1 \leq p < \infty, BC([a, b], \mathbb{R})$, and $BC([a, b], \mathbb{C})$ are separable (with respect to their standard norms/metrics).
- Neither $l_\infty$ nor $BC((a, b), \mathbb{R})$ or $BC((a, b), \mathbb{C})$ is separable. (In this respect, these spaces are much 'bigger' than the other spaces considered in this course.)

## $\wp$ Completion theorem

Every metric (normed) space is densely embedded in a complete metric (normed) space.

**Ex.**

- If we complete $\mathbb{Q}$ with respect to the metric $d(x, y) = |x - y|$ we get $\mathbb{R}$:

$$\mathbb{Q} \overset{\text{dense}}{\hookrightarrow} \mathbb{R}.$$

- If we complete $C^\infty([0, 1], \mathbb{R})$ with respect to the supremum norm, $\|\cdot\|_\infty$, we get $BC([0, 1], \mathbb{R})$.
- Let $I \subset \mathbb{R}$ be an open interval, and consider (measurable) functions such that the integral

$$\int_I |f(x)|^2 \, dx \qquad \text{exists and is finite.}$$

[1] There are many versions of this theorem. The classical result states that the set of polynomials are dense in $BC([a, b], \mathbb{C})$.

Then

$$\|f\|_{L_2(I,\mathbb{R})} := \Big(\int_I |f(x)|^2 \, dx\Big)^{1/2}$$

defines a norm on these functions, so we have a normed space[1]. The completion of this space is called **the space of square-integrable functions**, written

$$L_2(I, \mathbb{R}).$$

The same space can be obtained by completing $C(I, \mathbb{R})$ with respect to the $L_2$-norm. Hence

$$(C(I, \mathbb{R}), \|\cdot\|_{L_2}) \overset{\text{dense}}{\hookrightarrow} L_2(I, \mathbb{R}).$$

A deep result in analysis is that $L_2(I, \mathbb{R}) \cong l_2(\mathbb{R})$ (isometrically isomorphic): the elements in $l_2(\mathbb{R})$ may be identified as (generalised) Fourier coefficients of elements in $L_2(I, \mathbb{R})$.

[1] Two functions in this space are equal if they are equal almost everywhere on $I$.

# Chapter 2

# Linear spaces and transformations

## 2.1 Linear subspaces

Let $X$ be a vector space. A subset $S \subset X$ is a **subspace** of $X$ if it is closed under linear operations, i.e.

$$S \subset X \text{ subspace of } X \quad \overset{\text{def}}{\iff} \quad \lambda x + \mu y \in S \quad \text{whenever} \quad x, y \in S \text{ and } \mu, \lambda \in \mathbb{R} \quad (\text{or } \mathbb{C}).$$

In particular, $\mathbf{0} \in S$, and $S$ is itself a vector space (the axioms for a vector space follow from those of $X$).

**Ex.**

- In any vector space, $\{\mathbf{0}\}$ (the set consisting only of the zero element) is a subspace, since

  $$\lambda\mathbf{0} + \mu\mathbf{0} = \mathbf{0} \in \{\mathbf{0}\} \quad \text{for all scalars } \lambda, \mu.$$

- Consider

  $$\mathbb{R} = \{(x, 0, 0) \colon x \in \mathbb{R}\}$$

  as a subset of

  $$\mathbb{R}^3 = \{(x, y, z) \colon x, y, z \in \mathbb{R}\}.$$

  Then $\mathbb{R}$ is a subspace of $\mathbb{R}^3$, since it is non-empty and

  $$\lambda(x_1, 0, 0) + \mu(x_2, 0, 0) = (\lambda x_1 + \mu x_2, 0, 0) \in \mathbb{R} \subset \mathbb{R}^3.$$

- Similarly, the set of real-valued continuous functions on $\mathbb{R}$ which vanish on some set $S \subset \mathbb{R}$ is a subspace of $C(\mathbb{R}, \mathbb{R})$:

  $$\{f \in C(\mathbb{R}, \mathbb{R}) \colon f \equiv 0 \text{ on } S\} \quad \text{is a subspace of} \quad C(\mathbb{R}, \mathbb{R}),$$

  since

  $$\mu f(x) + \lambda g(x) = 0 \quad \text{if} \quad f(x) = 0 \text{ and } g(x) = 0.$$

- The vector space of polynomials of degree at most $n$, $P_n(\mathbb{R})$, endowed with the usual addition and scalar multiplication, is a subspace of the set of polynomials of degree at most $n+1$. Indeed,

$$P_0(\mathbb{R}) \subset P_1(\mathbb{R}) \subset \ldots \subset P_n(\mathbb{R}) \subset P_{n+1}(\mathbb{R}) \subset \ldots \subset P(\mathbb{R}) := \bigcup_{n=0}^{\infty} P_n(\mathbb{R})$$

are all subspaces of each other and, ultimately, of **the vector space of all polynomials**, $P(\mathbb{R})$. The isomorphisms $P_n(\mathbb{R}) \cong \mathbb{R}^{n+1}$ induces a natural representation of $P(\mathbb{R})$:

$$P(\mathbb{R}) = \{(a_0, a_1, \ldots, a_n, 0, 0, \ldots) \colon a_0, \ldots, a_n \in \mathbb{R} \text{ and } n \in \mathbb{N}\}.$$

## 2.2 Linear dependence

Let $X$ be a vector space, and $S \subset X$ any subset of $X$.

### Span

A **linear combination** of vectors $u_1, \ldots, u_n$ is a finite sum

$$\sum_{j=1}^{n} a_j u_j,$$

where $a_1, \ldots, a_n$ are scalars. The **(linear) span** of $S \subset X$ is the set of all linear combinations of vectors in $S$:

$$\operatorname{span}(S) \overset{\text{def.}}{=} \Big\{ \sum_{\text{finite}} a_j x_j \colon x_j \in S, a_j \text{ scalars} \Big\}.$$

For convenience, we define $\operatorname{span}(\emptyset) \overset{\text{def.}}{=} \{\mathbf{0}\}$. If $V = \operatorname{span}(S)$ we say that $S$ **generates** $V$.

$\wp$ **The linear span of a set S is the smallest subspace containing S**

For any $S \subset X$, $\operatorname{span}(S)$ is a subspace of $X$, and

$$\operatorname{span}(S) = \bigcap_{S \subset V} \{V \colon V \text{ is a subspace of } X\}.$$

**Ex.**

- Let $x = (1,0)$, $y = (2,0)$ and $z = (1,1)$ be vectors in $\mathbb{R}^2$. Then

$$\operatorname{span}\{x\} = \operatorname{span}\{y\} = \operatorname{span}\{x,y\} = \{(\lambda, 0) \colon \lambda \in \mathbb{R}\},$$

$$\operatorname{span}\{z\} = \{(\lambda, \lambda) \colon \lambda \in \mathbb{R}\},$$

$$\operatorname{span}\{x,z\} = \operatorname{span}\{y,z\} = \operatorname{span}\{x,y,z\} = \mathbb{R}^2.$$

- The vectors $e_1 = (1, 0, \ldots)$, $e_2 = (0, 1, 0, \ldots)$, …, $e_n = (0, \ldots, 0, 1)$ generate $\mathbb{R}^n$.
- In general, the span of a set differs between real and complex vector spaces:

$$\operatorname{span}_{\mathbb{R}}\{1\} = \mathbb{R} \quad \text{but} \quad \operatorname{span}_{\mathbb{C}}\{1\} = \mathbb{C}.$$

### Linear dependence

A family of vectors $u_1, u_2, \dots$ is called **linearly dependent** if one of them is linear combination of some of the others:

$$\{u_1, u_2, \dots\} \text{ linearly dependent} \quad \overset{\text{def}}{\iff} \quad \sum_{j=1}^{n} a_j u_j = \mathbf{0} \quad \text{for some } n \in \mathbb{N} \text{ and at least one } a_j \neq 0.$$

Else, the family is **linearly independent**:

$$\{u_1, u_2, \dots\} \text{ linearly independent} \quad \overset{\text{def}}{\iff} \quad \left[\text{for any } n \in \mathbb{N} : \ \sum_{j=1}^{n} a_j u_j = \mathbf{0} \implies a_j = 0 \,\forall\, j\right].$$

More generally, a set $S$ is linearly independent if all finite subsets of it are linearly independent.

**Ex.**

- The vectors $x = (1, 0)$, $y = (2, 0)$ and $z = (1, 1)$ are linearly dependent in $\mathbb{R}^2$, since

$$\begin{bmatrix} 2 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

  But both the sets $\{x, z\}$ and $\{y, z\}$ are linearly independent, since

$$a_1 x + a_2 z = \mathbf{0} \iff a_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \iff \begin{bmatrix} a_1 + a_2 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \iff a_1 = a_2 = 0,$$

  and similarly for $\{y, z\}$.
- If $\mathbf{0} \in S$, then $S$ is linearly dependent.
- $\{1, x, x^2, \dots\}$ is linearly independent in $P(\mathbb{R})$.
- $\{1, \cos(x), \sin(x), \cos(2x), \sin(2x), \dots\}$ is linearly independent in $C(I, \mathbb{R})$.

## 2.3 Bases and dimension

Let $X$ be a vector space.

### Hamel Bases

A linearly independent set which generates $X$ is called a **(Hamel) basis** for $X$:

$$S \subset X \text{ Hamel basis for } X \quad \overset{\text{def}}{\iff} \quad \operatorname{span}(S) = X \quad \text{and} \quad S \text{ lin. indep.}$$

Equivalently, $S$ is a Hamel basis for $X$ if every vector $x \in X$ has a *unique and finite* representation

$$x = \sum_{\text{finite}} a_j u_j, \qquad u_j \in S.$$

We shall consider only **ordered** Hamel bases, in which case the scalars $a_j$, called **coordinates**, are well defined.

**Ex.**

- $\{e_1, \dots, e_n\}$, with

$$e_j = (0, \dots, \underbrace{1}_{\text{jth position}}, 0 \dots)$$

  is called the **standard basis** for $\mathbb{R}^n$.

- $\{1, x, x^2, \ldots\}$ is an ordered Hamel basis for $P(\mathbb{R})$: every real polynomial can be uniquely expressed as a finite sum,

$$p(x) = \sum_{\text{finite}} a_j x^j, \qquad a_j \in \mathbb{R}.$$

## Dimension

If $X$ has a basis consisting of finitely many vectors, $X$ is said to be **finite-dimensional**. Else, $X$ is **infinite-dimensional**.

### $\wp$ The dimension of any finite-dimensional vector space is unique

All bases of a finite-dimensional vector space have the same number of elements. This number is called the **dimension** of the space.

**Proof**

Suppose that $\{e_j\}_{j=1}^m$ and $\{f_j\}_{j=1}^n$ are both bases, and that $m > n$. Since a basis is linearly independent, the only solution of

$$\sum_{j=1}^m a_j e_j = \mathbf{0} \quad \text{ought to be} \quad a_1, \ldots, a_m = 0.$$

Since $\{f_j\}_{j=1}^n$ is also a basis, we may represent $e_j = \sum_{k=1}^n b_{j,k} f_k$ in a unique way. Then

$$\sum_{j=1}^m \sum_{k=1}^n a_j b_{j,k} f_k = 0 \quad \text{meaning that} \quad \sum_{j=1}^m a_j b_{j,k} = 0 \;\text{ for }\; k = 1, \ldots, n.$$

This is a linear homogeneous system with $n$ equations and $m > n$ unknowns (the scalars $a_j$). Such a system always has a non-trivial solution (meaning that some $a_j \neq 0$). Hence $\{e_j\}_{j=1}^m$ is not linearly independent, so it cannot be a basis.

**Ex.**

- $\mathbb{R}^n$ has dimension $n$.
- $P_n(\mathbb{R})$, has dimension $n+1$. (Recall that $P_n(\mathbb{R}) \cong \mathbb{R}^{n+1}$.)
- $\mathbb{C}^n$ has dimension $n$ when considered as a *complex* vector space, but $2n$ when considered a *real* vector space.
- The $l_p$-, $BC$-, and $L_2$-spaces are all infinite-dimensional.

### $\wp$ Any finite-dimensional vector space is isomorphic to Euclidean space

Let $X$ be a real vector space with basis $\{e_1, \ldots, e_n\}$. Then $X \cong \mathbb{R}^n$.[1]

**Proof**

By the definition of a basis, any $x \in X$ has a unique representation

$$x = \sum_{j=1}^n a_j e_j.$$

[1] If $X$ is a complex vector space, $X \cong \mathbb{C}^n$.

Let $T\colon X \to \mathbb{R}^n$ be the mapping defined by

$$Tx = (a_1, \ldots, a_n).$$

**$T$ is linear:** if $x = \sum a_j e_j$ and $y = \sum b_j e_j$,

$$T(\lambda x + \mu y) = (\lambda a_1 + \mu b_1, \ldots, \lambda a_n + \mu b_n) = \lambda(a_1, \ldots, a_n) + \mu(b_1, \ldots, b_n) = \lambda Tx + \mu Ty,$$

**$T$ is surjective:**

$$\text{for any } (a_1, \ldots, a_n) \in \mathbb{R}^n \quad \text{there exists} \quad x = \sum_{j=1}^{n} a_j e_j;\ Tx = (a_1, \ldots, a_n)$$

**$T$ is injective:**

$$Tx = Ty \quad \iff \quad \forall j\colon a_j = b_j \quad \implies \quad x = y.$$

Thus $T$ is a vector space isomorphism.

**N.b.** In an $n$-dimensional vector space, $m > n$ vectors are linearly dependent.

## 2.4 Schauder bases

Whereas the concept of a Hamel basis is very general — it applies to any vector space — it is not particularly well suited for infinite-dimensional Banach spaces.

$\wp$ **Infinite-dimensional Banach spaces have only uncountable Hamel bases**
Let $X$ be an infinite-dimensional Banach space. Then a sequence $\{e_j\}_{j\in\mathbb{N}}$ cannot be a Hamel basis for $X$.

### Schauder Bases

Let $(X, \|\cdot\|)$ be a Banach space. A sequence $\{e_j\}_{j\in\mathbb{N}}$ is called a **Schauder basis** (or **countable basis**) for $X$ if every vector $x \in X$ has a *unique* represention

$$x = \sum_{j\in\mathbb{N}} x_j e_j,$$

meaning that $\lim_{N\to\infty} \|x - \sum_{j=1}^{N} x_j e_j\| = 0$. The scalars $x_j$ are the **coordinates** of $x$.

**N.b.** Any space with a Schauder basis is separable.[1]

*From now on, the word* ***basis*** *refers to an ordered basis, Hamel (in the case of any finite-dimensional vector space) or Schauder (in the case of any infinite-dimensional Banach space).* In both cases, a basis assigns to each $x \in X$ unique coordinates $x_1, x_2, \ldots$.

**Ex.**

- Let $e_j = (0, \ldots, \underbrace{1}_{\text{jth position}}, 0 \ldots)$. Then $\{e_j\}_{j=1}^{\infty}$ is a (Schauder) basis for $l_p$, $1 \le p < \infty$:[2]

  **Approximation property:**

$$x = \{x_j\}_{j\in\mathbb{N}} \in l_p \quad \implies \quad \sum_{j=1}^{\infty} |x_j|^p < \infty \quad \implies \quad \sum_{N+1}^{\infty} |x_j|^p \to 0 \text{ as } N \to \infty.$$

[1] The opposite is not true; there are (strange) separable Banach spaces with no Schauder basis.

[2] $l_\infty$ has no Schauder basis.

Thus

$$\Big\|\sum_{j=1}^{N} x_j e_j - x\Big\|_{l_p} = \|(x_1,\dots,x_N,0,0,\dots)-(x_1,\dots,x_N,x_{N+1},\dots)\|_{l_p} = \Big(\sum_{N+1}^{\infty}|x_j|^p\Big)^{1/p} \to 0 \text{ as } N\to\infty.$$

**Uniqueness of coordinates:**

$$\sum_{j=1}^{\infty} x_j e_j = \sum_{j=1}^{\infty} y_j e_j \quad\Longleftrightarrow\quad \lim_{N\to\infty}\sum_{j=1}^{N}|x_j-y_j|^p = 0 \quad\Longrightarrow\quad x_j = y_j,\ \text{for all } j\in\mathbb{N}.$$

- The trigonometric functions $\{\mathrm{e}^{ikx}\}_{k\in\mathbb{Z}}$ is a (Schauder) basis for $L_2((-\pi,\pi),\mathbb{C})$. The coordinates in this basis are known as **Fourier coefficients**.
- The vectors $(1,0,0),(1,1,0),(1,1,1)$ provide a (Hamel) basis for $\mathbb{R}^3$. Find the coordinates $[c_1,c_2,c_3]$ of $(2,0,1)$ in this basis.

$$c_1\begin{bmatrix}1\\0\\0\end{bmatrix}+c_2\begin{bmatrix}1\\1\\0\end{bmatrix}+c_3\begin{bmatrix}1\\1\\1\end{bmatrix}=\begin{bmatrix}2\\0\\1\end{bmatrix} \quad\Longleftrightarrow\quad \begin{bmatrix}c_1+ & c_2+ & c_3\\ & c_2+ & c_3\\ & & c_3\end{bmatrix}=\begin{bmatrix}2\\0\\1\end{bmatrix} \quad\Longleftrightarrow\quad \begin{bmatrix}c_1\\c_2\\c_3\end{bmatrix}=\begin{bmatrix}2\\-1\\1\end{bmatrix}.$$

The coordinates for $(2,0,1)$ in the new basis are $[2,-1,1]$.

## 2.5 Basis transformations

### Change-of-basis matrix

Let $e=\{e_1,\dots,e_n\}$ and $f=\{f_1,\dots,f_n\}$ be two bases for a finite-dimensional real vector space $X$. Pick any element $x\in X$. Then

$$x=\sum_{j=1}^{n} x_j e_j$$

has coordinates $(x_1,\dots,x_n)_e$ in the basis $e$. Since $f$ is also a basis, we may express

$$e_j=\sum_{k=1}^{n} c_{k,j}\, f_k,\quad j=1,\dots,n, \qquad\text{and}\qquad x=\sum_{j=1}^{n} x_j\sum_{k=1}^{n} c_{k,j}\, f_k=\sum_{k=1}^{n}\Big(\underbrace{\sum_{j=1}^{n} c_{k,j}\, x_j}_{\text{coord. in } f}\Big) f_k.$$

Put differently,

$$\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}=\begin{bmatrix}c_{1,1}&c_{1,2}&\dots&c_{1,n}\\c_{2,1}&c_{2,2}&\dots&c_{2,n}\\\vdots&\vdots&\ddots&\vdots\\c_{n,1}&c_{n,2}&\dots&c_{n,n}\end{bmatrix}\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}$$

defines the coordinates $(y_1,\dots,y_n)$ of $x$ in the basis $f$: $x_f = Cx_e$. The $n\times n$ **scalar-valued matrix** $C\in M_{n\times n}(\mathbb{R})$ is called a **change-of-basis matrix**.

- The change-of-basis matrix from $e = \{(1,0,0),(0,1,0),(0,0,1)\}$ to $f = \{(1,0,0),(1,1,0),(1,1,1)\}$ in $\mathbb{R}^n$:

$$\begin{aligned} e_1 = \sum_{k=1}^{3} c_{k,1} f_k \quad &\Longleftrightarrow \quad \begin{bmatrix} c_{1,1} \\ c_{2,1} \\ c_{3,1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\ e_2 = \sum_{k=1}^{3} c_{k,2} f_k \quad &\Longleftrightarrow \quad \begin{bmatrix} c_{1,2} \\ c_{2,2} \\ c_{3,2} \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}. \\ e_3 = \sum_{k=1}^{3} c_{k,3} f_k \quad &\Longleftrightarrow \quad \begin{bmatrix} c_{1,3} \\ c_{2,3} \\ c_{3,3} \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} \end{aligned} \tag{1}$$

The change-of-basis matrix is

$$C = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}.$$

In particular,

$$\begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \qquad \text{yields that} \qquad (2,0,1)_e = (2-1,1)_f.$$

## Change-of-basis matrix as an inverse

If we write (1) in column form, we get:

$$\left| \begin{bmatrix} \cdot \\ e_1 \\ \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ e_2 \\ \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ e_3 \\ \cdot \end{bmatrix} \right| = \left| \begin{bmatrix} \cdot \\ f_1 \\ \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ f_2 \\ \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ f_3 \\ \cdot \end{bmatrix} \right| \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \quad \Longleftrightarrow \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus $I = [f]C$ and $C = [f]^{-1}$, where $[f]$ is the matrix with the basis vectors $f_1, \ldots, f_n$ as column vectors.

**N.b.** Matrices of this form—with zeros below the main diagonal—are called **upper triangular**. More precisely, $(a_{ij})_{ij}$ is upper triangular if $a_{ij} = 0$ for $i > j$. Lower triangular matrices are defined in a similar fashion ($a_{ij} = 0$ for $j > i$).

### $\wp$ The inverse of a basis matrix is its inverse change-of-basis matrix

Let $[f] = [f_1, \ldots, f_n] \in M_{n \times n}(\mathbb{C})$ denote a matrix with column basis vectors $f_1, \ldots, f_n \in \mathbb{C}^n$ expressed in the standard basis $e$. Then

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}_e = \underbrace{\left| \begin{bmatrix} \cdot \\ f_1 \\ \cdot \end{bmatrix} \ldots \begin{bmatrix} \cdot \\ f_n \\ \cdot \end{bmatrix} \right|}_{[f]} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_f \qquad \text{expresses } (y_1, \ldots, y_n)_f \text{ in the basis } e,$$

and

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_f = \underbrace{\left| \begin{bmatrix} \cdot \\ f_1 \\ \cdot \end{bmatrix} \ldots \begin{bmatrix} \cdot \\ f_n \\ \cdot \end{bmatrix} \right|^{-1}}_{[f]^{-1}} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}_e \qquad \text{expresses } (x_1, \ldots, x_n)_e \text{ in the basis } f.$$

$\wp$ **Any basis in a finite-dimensional vector space corresponds to an invertible matrix**

**Proof**

Consider $X \cong \mathbb{F}^n$, $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$. We know from linear algebra that:

$A \in M_{n\times n}(\mathbb{F})$ invertible $\iff$ the columns $A_1, \dots, A_n$ of $A$ are lin.ind. $\iff$ $\{A_1, \dots, A_n\}$ is a basis for $\mathbb{F}$

## 2.6 Gaussian elimination

### Linear systems

Any linear system of equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \quad \vdots \qquad \vdots \quad &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

where $a_{ij}, b_i \in \mathbb{C}$ for $i = 1, \dots, m$, $j = 1, \dots, n$, and $x_1, \dots, x_n$ are unknowns, can be written in matrix form:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}.$$

Finding solutions $(x_1, \dots, x_n$ of the linear system of equations is then equivalent to the following question: given a matrix $A \in M_{m\times n}(\mathbb{C})$ and a vector $b \in \mathbb{C}^m$, is there a vector $x \in \mathbb{C}^n$ such that $Ax = b$?[1]

*The answer depends on $A$ and, for some $A$, also on $b$.*

**Ex.**

- $\begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \iff \begin{aligned} x_1 &= 1 \\ 2x_1 &= 1 \end{aligned}$ has no solution.
- $\begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \iff \begin{aligned} x_1 &= 1 \\ 2x_1 &= 2 \end{aligned}$ has infinitely many solutions: $x_1 = 1, x_2 \in \mathbb{R}$.
- $\begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \iff \begin{aligned} x_1 + x_2 &= b_1 \\ 2x_1 \qquad &= b_2 \end{aligned}$ has a unique solution for any $b_1, b_2$ : $x_1 = \frac{b_2}{2}, x_2 = b_1 - \frac{b_2}{2}$.

### Gaussian elimination and the row echelon form of a matrix

A matrix is in **row echelon form** if i) the left-most non-zero entry of each row (**pivot**) is strictly to the right of the left-most non-zero entry of any row above, and ii) all-zero rows are at the bottom of the matrix:

$$\begin{bmatrix} 1 & \cdots & & & \\ 0 & 2 & \cdots & & \\ 0 & 0 & 0 & 7 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \end{bmatrix}$$

[1] When $A$ and $b$ are real, one looks for $x$ real.

$\wp$ **Any linear system can be brought into row echelon form**

**Proof**

If $A = 0$ is the zero matrix, we are done.

Else, assume for simplicity that $a_{11} \neq 0$ (If not rearrange the rows, or relabel the $x_j$'s, or both). To each row $(a_{i1}, \ldots, a_{in})$ below the first, add $-\frac{a_{i1}}{a_{11}} \times$ the first row:

$$\begin{array}{rr} a_{11}x_1 + a_{12}x_2 + \cdots a_{1n}x_n = b_1 & \qquad -\dfrac{a_{i1}}{a_{11}} \\ a_{i1}x_1 + a_{i2}x_2 + \cdots a_{in}x_n = b_2 & \\ \hline 0 + \left(a_{i2} - \dfrac{a_{i1}}{a_{11}}a_{12}\right)x_2 + \cdots \left(a_{in} - \dfrac{a_{i1}}{a_{11}}a_{1n}\right)x_n = b_2 - \dfrac{a_{i1}}{a_{11}}b_1 & \end{array}$$

Then $a_{i1} = 0$ for all $i = 2, \ldots, m$.

Now, either $a_{ij} = 0$ for all $i, j \geq 2$, or we can restart this procedure (looking at the matrix for indices $i, j \geq 2$). Since there are finitely many rows this procedure must eventually terminate, yielding a matrix $\tilde{A} = (\tilde{a}_{ij})_{ij}$ in row echelon form.

This algorithm is called **Gaussian elimination**.

A neat trick is the following: write $Ax = b$ as $Ax = Ib$:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & 0 & \ddots & \ldots \\ 0 & & \cdots & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \qquad I \in M_{m \times m}.$$

*Gaussian elimination affects only the matrices $A$ and $I$, not the vectors $x$ and $b$.* We therefore introduce the $m \times (n+m)$ **augmented matrix**

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & 1 & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & 0 & \ddots & \ldots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & 0 & & \cdots & 1 \end{bmatrix}.$$

*The linear operations applied to $A$ in Gaussian elimination are 'stored' in the augmented matrix.*

**Ex.**

Solve

$$\underbrace{\begin{bmatrix} 1 & 3 & 1 \\ 2 & 2 & 0 \\ 2 & 2 & -1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}}_{b} \iff \begin{bmatrix} 1 & 3 & 1 \\ 2 & 2 & 0 \\ 2 & 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}.$$

Now, performing the same row operations on the whole augmented matrix,

$$\begin{bmatrix} 1 & 3 & 1 & 1 & 0 & 0 \\ 2 & 2 & 0 & 0 & 1 & 0 \\ 2 & 2 & -1 & 0 & 0 & 1 \end{bmatrix} \iff \begin{bmatrix} 1 & 3 & 1 & 1 & 0 & 0 \\ 0 & -4 & -2 & -2 & 1 & 0 \\ 0 & -4 & -3 & -2 & 0 & 1 \end{bmatrix} \iff \begin{bmatrix} 1 & 3 & 1 & 1 & 0 & 0 \\ 0 & -4 & -2 & -2 & 1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 1 \end{bmatrix},$$

we find the row echelon form

$$\underbrace{\begin{bmatrix}1 & 3 & 1\\ 0 & -4 & -2\\ 0 & 0 & -1\end{bmatrix}}_{U\text{: upper triangular}} \begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} = \underbrace{\begin{bmatrix}1 & 0 & 0\\ -2 & 1 & 0\\ 0 & -1 & 1\end{bmatrix}}_{\tilde{L}\text{: lower triangular}} \begin{bmatrix}1\\ 4\\ 2\end{bmatrix}.$$

Note that $\tilde{L}$ describes the (linear) transformation applied to $A$ to obtain $U$: we have $\tilde{L}A = U$.

## LU-decompositions

As we shall see, if $\tilde{L}A = U$, and if the inverse matrix $\tilde{L}^{-1}$ exists, bringing $\tilde{L}$ into row echelon row form yields $\tilde{L}^{-1}$. Define $L := \tilde{L}^{-1}$. Then

$$\tilde{L}A = U \quad \Longleftrightarrow \quad L\tilde{L}A = LU \quad \Longleftrightarrow \quad A = LU.$$

This is the **LU-decomposition** of the matrix $A$ (it does not always exist).[1]

**N.b.** If the rows of $A$ are not in correct order (for example, if $a_{11} = 0$), they can be rearranged by applying a **permutation matrix** $P$[2]:

$$\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 1 & 0 & 0\end{bmatrix} \begin{bmatrix}\text{row } 1\\ \text{row } 2\\ \text{row } 3\end{bmatrix} = \begin{bmatrix}\text{row } 2\\ \text{row } 3\\ \text{row } 1\end{bmatrix}.$$

This is known as an **LUP-factorization**: $PA = LU$. For square matrices, an LUP-factorization always exists (but it is not necessarily unique).

**Ex.**

Gaussian elimination for $\tilde{L}$ yields

$$\left[\begin{array}{ccc|ccc}1 & 0 & 0 & 1 & 0 & 0\\ -2 & 1 & 0 & 0 & 1 & 0\\ 0 & -1 & 1 & 0 & 0 & 1\end{array}\right] \Longleftrightarrow \left[\begin{array}{ccc|ccc}1 & 0 & 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 2 & 1 & 0\\ 0 & -1 & 1 & 0 & 0 & 1\end{array}\right] \Longleftrightarrow \left[\begin{array}{ccc|ccc}1 & 0 & 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 2 & 1 & 0\\ 0 & 0 & 1 & 2 & 1 & 1\end{array}\right],$$

meaning that

$$\underbrace{\begin{bmatrix}1 & 0 & 0\\ 2 & 1 & 0\\ 2 & 1 & 1\end{bmatrix}}_{L=\tilde{L}^{-1}} \underbrace{\begin{bmatrix}1 & 0 & 0\\ -2 & 1 & 0\\ 0 & -1 & 1\end{bmatrix}}_{\tilde{L}} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}.$$

Hence, the LU-factorization of $A$ is

$$\begin{bmatrix}1 & 3 & 1\\ 2 & 2 & 0\\ 2 & 2 & -1\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\ 2 & 1 & 0\\ 2 & 1 & 1\end{bmatrix} \begin{bmatrix}1 & 3 & 1\\ 0 & -4 & -2\\ 0 & 0 & -1\end{bmatrix}.$$

## Gauss-Jordan elimination

Let $\tilde{L}A = U$. The algorithm that applies Gaussian elimination (backwards) to the matrix $U$ is called **Gauss-Jordan elimination**. It places zeros *below and above* each pivot; this is called the **reduced row echelon form** of the matrix $A$. If $A^{-1}$ exists, Gauss–Jordan elimination finds it.

[1] When it exists, it is unique if one requires the diagonal elements of $L$ to be all ones.

[2] A permutation matrix has (a permutation of) the standard basis $\{e_j\}_j$ as rows.

- Above, $\tilde{L}^{-1}$ gave us $A = \tilde{L}^{-1}U$ (the LU-factorization of $A$).
- In the following, $U^{-1}$ gives us $U^{-1}\tilde{L}A = I$, so that $U^{-1}\tilde{L} = U^{-1}L^{-1} = A^{-1}$ (the inverse of $A$).

**Ex.**

Still solving

$$\underbrace{\begin{bmatrix} 1 & 3 & 1 \\ 2 & 2 & 0 \\ 2 & 2 & -1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}}_{b} \quad \Longleftrightarrow \quad \begin{bmatrix} 1 & 3 & 1 \\ 2 & 2 & 0 \\ 2 & 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}.$$

Gauss:

$$\begin{bmatrix} 1 & 3 & 1 & 1 & 0 & 0 \\ 0 & -4 & -2 & -2 & 1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 1 \end{bmatrix}.$$

Gauss–Jordan[1]:

$$\begin{bmatrix} 1 & 3 & 1 & 1 & 0 & 0 \\ 0 & -4 & -2 & -2 & 1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 1 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 1 & 3 & 0 & 1 & -1 & 1 \\ 0 & -4 & 0 & -2 & 3 & -2 \\ 0 & 0 & -1 & 0 & -1 & 1 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 1 & 0 & 0 & -\frac{1}{2} & \frac{5}{4} & -\frac{1}{2} \\ 0 & -4 & 0 & -2 & 3 & -2 \\ 0 & 0 & -1 & 0 & -1 & 1 \end{bmatrix}$$

Dividing each row with its pivot gives us the inverse of $A$:

$$\begin{bmatrix} 1 & 0 & 0 & -\frac{1}{2} & \frac{5}{4} & -\frac{1}{2} \\ 0 & 1 & 0 & \frac{1}{2} & -\frac{3}{4} & \frac{1}{2} \\ 0 & 0 & 1 & 0 & 1 & -1 \end{bmatrix},$$

meaning that

$$A^{-1} = U^{-1}\tilde{L} = \begin{bmatrix} -\frac{1}{2} & \frac{5}{4} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{3}{4} & \frac{1}{2} \\ 0 & 1 & -1 \end{bmatrix}$$

is the linear transformation that brings $A$ into the unit matrix. The solution to the original problem is

$$x = A^{-1}b \quad \text{for any } b \in \mathbb{R}^3.$$

**N.b.** The product of the final pivots in the Gauss–Jordan elimination (the diagonal elements $(1)(-4)(-1) = 4$ in the example above) is the **determinant** of $A$. It is zero exactly if the Gauss(–Jordan) elimination produces a zero row. In this case the equation $Ax = b$ either has no, or infinitely many, solutions (depending on $b$).

## 2.7 Linear transformations

Let $X$ and $Y$ be vector spaces (both real, or both complex), and $T\colon X \to Y$ a mapping between them.

[1] If we start here with $I$ to the right in the augmented matrix, we obtain $U^{-1}$ instead of $A^{-1}$.

## Linear transformations

We say that $T$ is a **linear transformation** (or just **linear**) if it preserves the linear structure of a vector space:

$$T \text{ linear} \quad \overset{\text{def}}{\Longleftrightarrow} \quad T(\lambda x + \mu y) = \lambda Tx + \mu Ty, \qquad x, y \in X,\ \mu, \lambda \in \mathbb{R}\,(\text{or } \mathbb{C}).$$

**Ex.**

- Any matrix $A \in M_{m\times n}(\mathbb{R})$ defines a linear transformation $\mathbb{R}^n \to \mathbb{R}^m$:

$$\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{x} \mapsto \underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{Ax}.$$

- The integral operator defined by $Tf(t) := \int_0^t f(s)\,ds$ is a linear transformation on $C(I, \mathbb{R})$:

$$T\colon C(I,\mathbb{R}) \to C(I,\mathbb{R}), \quad Tf = \Big[t \mapsto \int_0^t f(s)\,ds\Big].$$

- A slight modification,

$$Tf := \int_0^1 f(s)\,ds,$$

  yields a linear transformation $C(I,\mathbb{R}) \to \mathbb{R}$ (given that $[0,1] \subset I$).[1]

- For any polynomial $p \in P_k(\mathbb{R})$, the differential operator $p(D) := \sum_{j=0}^k a_j D^j$ is a linear transformation:

$$p(D)\colon C^k(I,\mathbb{R}) \to C(I,\mathbb{R}), \qquad p(D)f = \sum_{j=0}^k a_j f^{(j)}.$$

  Here, $D = \frac{d}{dx}$ is the standard differentiation operator.

- The shift operator $T\colon (x_1, x_2, \ldots) \mapsto (0, x_1, x_2, \ldots)$, is a linear transformation $l_p \to l_p$, for any $p$, $1 \le p \le \infty$:

$$T(\lambda x + \mu y) = (0, \lambda x_1 + \mu y_1, \ldots) = \lambda(0, x_1, \ldots) + \mu(0, y_1, \ldots) = \lambda Tx + \mu Ty.$$

  Note that $\|Tx\|_{l_p} = \|x\|_{l_p}$ guarantees that $\operatorname{ran}(T) \subset l_p$.

## The set of linear transformations as a vector space

**The set of linear transformations** $X \to Y$ is denoted by $L(X,Y)$:

$$L(X,Y) \overset{\text{def.}}{=} \{T\colon X \to Y \text{ linear}\}$$

IF $X = Y$, we may abbreviate $L(X,X)$ by $L(X)$.

### $\wp$ L(X,Y) is a vector space

If, for all $S, T \in L(X,Y)$, we define

$$(T+S)(x) := Tx + Sx \quad \text{ and } \quad (\lambda T)x := \lambda(Tx),$$

[1] Such an operator is called a **linear functional**.

for all $x \in X$ and $\lambda \in \mathbb{R}$ (or $\mathbb{C}$), it is easily checked that $L(X,Y)$ becomes a vector space. In particular, $\mu T + \lambda S \in L(X,Y)$ for any $S, T \in L(X,Y)$.

**Ex.**

- The set of $m \times n$-matrices $M_{m\times n}(\mathbb{R})$ forms a real vector space. As we shall see, $M_{m\times n}(\mathbb{R}) \cong L(\mathbb{R}^n, \mathbb{R}^m)$.

### $\wp$ A linear transformation is determined by its action on any basis

Let $X$ be a finite-dimensional[1] vector space with basis $\{e_1, \ldots, e_n\}$. For any values $y_1, \ldots, y_n \in Y$ there exists exactly one linear transformation $T \in L(X,Y)$ such that

$$Te_j = y_j, \qquad j = 1, \ldots, n.$$

**Proof**

Any $x \in X$ has a unique representation $x = \sum_{j=1}^n x_j e_j$. Define $T$ through

$$Tx = \sum_{j=1}^n x_j y_j.$$

Then $Te_j = y_j$, and $T$ is linear since it acts as multiplication with a $1 \times n$ matrix (a dot product with the vector $(y_1, \ldots, y_n)$). Moreover, if $S \in L(X,Y)$ also satisfies $Se_j = y_j$, then

$$Sx = S\big(\sum_{j=1}^n x_j e_j\big) = \sum_{j=1}^n x_j Se_j = \sum_{j=1}^n x_j y_j = Tx, \qquad \text{for all } x \in X,$$

so that $S = T$ in $L(X,Y)$.

**Ex.**

- The columns $A_j$ of an $m \times n$-matrix $A$ are determined by its action on the standard basis $\{e_j\}_{j=1}^n$:

  $$Ae_j = A_j, \qquad j = 1, \ldots, n.$$

  Here $A_j$ plays the role of $y_j$ in the above theorem.

### $\wp$ Linear transformations between finite-dimensional vector spaces correspond to matrices

Let $X, Y$ be real vector spaces of dimension $n$ and $m$, respectively. Then $L(X,Y) \cong M_{m\times n}(\mathbb{R})$.

**N.b.** The corresponding statement holds for complex vector spaces $X, Y$, with $M_{m\times n}(\mathbb{C})$ also complex-valued.

[1] For infinite-dimensional Banach spaces one needs the additional concept of **boundedness** (continuity) of a linear transformation to state a similar result, which then says that the transformation is determined by $Te_j$ (but we cannot choose $Te_j = y_j$ arbitrarily).

**Proof**

Since $X \cong \mathbb{R}^n$ and $Y \cong \mathbb{R}^m$ it suffices to prove the statement for these choices of $X$ and $Y$. Let $\{e_j\}_{j=1}^n$ be the standard basis for $\mathbb{R}^n$. Then

$$T\colon \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \mapsto \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

is a linear transformation $\mathbb{R}^n \to \mathbb{R}^m$ satisfying

$$Te_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix}.$$

According to the above proposition, there is exactly one such $T \in L(\mathbb{R}^n, \mathbb{R}^m)$. Since we can choose the columns of $A = (a_{ij})_{ij}$ to be any elements in $\mathbb{R}^m$, we get all possible $T \in L(\mathbb{R}^n, \mathbb{R}^m)$ in this way.

**Ex.**

- The linear transformation $T\colon (x_1, x_2) \mapsto (-x_2, x_1)$ on $\mathbb{R}^2$ is realized by a rotation matrix $A$:

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -x_2 \\ x_1 \end{bmatrix}.$$

- More generally,

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

rotates a vector $\theta$ radians counterclockwise; the preceeding example is attained for $\theta = \pi/2$. Any such matrix also corresponds to a change of basis[1]: if $f_1 = (\cos(\theta), \sin(\theta))$ and $f_2 = (-\sin(\theta), \cos(\theta))$, then

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

expresses the coordinates $(x_1, x_2)_f$ in the standard basis $e$ as $x_1 f_1 + x_2 f_2$.

- The differential operator $\frac{d}{dx}$ is a linear operator on $P_2(\mathbb{R})$. Since $P_2(\mathbb{R}) \cong \mathbb{R}^3$ via the vector space isomorphism

$$\sum_{j=0}^{2} a_j x^j \overset{\varphi}{\mapsto} (a_0, a_1, a_2),$$

we see that

$$\frac{d}{dx}\colon \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ 2a_2 \\ 0 \end{bmatrix}$$

[1] The determinant of the matrix is 1, so it is invertible, regardless of the value of $\theta$.

expresses the derivation

$$\frac{d}{dx}\big(a_0 + a_1 x + a_2 x^2\big) = a_1 + 2a_2 x + 0x^2$$

using a matrix.[1]

**Digression: representing linear transformations in different bases**

Let $e = \{e_1, \dots, e_n\}$ (standard basis) and $f = \{f_1, \dots, f_n\}$ (new basis) be two bases for $\mathbb{R}^n$, and $[f]$ the matrix with $[f_j]$, $j = 1, \dots, n$, as column vectors (expressed in the standard basis $e$). Then

$$x_e = [f]x_f \quad \text{and} \quad x_f = [f]^{-1}x_e.$$

Hence, if

$$T \in L(\mathbb{R}^n)\colon \quad T \text{ is realised by } A_e \in M_{n\times n}(\mathbb{R}) \text{ in the basis } e,$$

what is its realisation $A_f$ in the basis $f$? We have

$$y_e = A_e x_e \quad \Longleftrightarrow \quad y_f = [f]^{-1}y_e = [f]^{-1}A_e x_e = [f]^{-1}A_e[f]x_f.$$

Thus

$$A_f = [f]^{-1}A_e[f]$$

is the realisation of $T$ in the basis $f$.

**Ex.**

- How do we express the rotation

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_e \mapsto \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_e = \begin{bmatrix} -x_2 \\ x_1 \end{bmatrix}_e$$

in the basis $f = \{(1,1), (-1,0)\}$? Since

$$[f] = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad [f]^{-1} = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix},$$

we have

$$A_f = \underbrace{\begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}}_{[f]^{-1}}\underbrace{\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}}_{A_e}\underbrace{\begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}}_{[f]} = \begin{bmatrix} 1 & -1 \\ 2 & -1 \end{bmatrix}.$$

Check: $(x_1, x_2)_e \overset{[f]^{-1}}{\mapsto} (x_2, x_2 - x_1)_f \overset{A_f}{\mapsto} (x_1, x_1 + x_2)_f \overset{[f]}{\mapsto} (-x_2, x_1)_e$ describes the correct transformation.

## Kernels and ranks

Let $T \in L(X, Y)$. The set of vectors for which $T$ vanishes is called the **kernel of** T.

$$\ker(T) \overset{\text{def.}}{=} \{x \in X\colon Tx = \mathbf{0} \text{ in } Y\}.$$

[1] Note that this matrix is **nilpotent**, meaning that $A^n = 0$ for some $n \in \mathbb{N}$ (in this case $n = 3$). This is because three derivations on any $p \in P_2(\mathbb{R})$ produces the zero polynomial.

### ℘ The kernel and range of a linear transformation are vector spaces

Let $T \in L(X,Y)$. Then $\ker(T) \subset X$ is a linear subspace of $X$, and $\operatorname{ran}(T) \subset Y$ is a linear subspace of $Y$.

**N.b.** The dimension of $\operatorname{ran}(T)$ is called the **rank of** $T$, $\operatorname{rank}(T)$.

**Proof**

For the kernel of $T$: If $x_1, x_2 \in \ker(T)$, then

$$T(\lambda x_1 + \mu x_2) = \lambda T x_1 + \mu T x_2 = \lambda \mathbf{0} + \mu \mathbf{0} = \mathbf{0}.$$

This shows that $\ker(T)$ is a subspace of $X$. (Note, in particular, that the zero element of $X$ is always in $\ker(T)$.)

For the range of $T$: If $y_1, y_2 \in \operatorname{ran}(T)$, then there exists $x_1, x_2 \in X$ such that

$$Tx_1 = y_1, \qquad Tx_2 = y_2.$$

We want to show that, for any scalars $\lambda, \mu$, we have $\lambda y_1 + \mu y_2 \in \operatorname{ran}(T)$. But this follows from that

$$\lambda y_1 + \mu y_2 = \lambda T x_1 + \mu T x_2 = T(\lambda x_1 + \mu x_2) \in \operatorname{ran}(T),$$

where we have used that $\mu x_1 + \lambda x_2 \in X$, by the properties of a vector space.

**Ex.**

- The kernel of $T \in L(\mathbb{R}^2)\colon (x_1, x_2) \mapsto (-x_2, x_1)$ is the trivial subspace $\{(0,0)\} \subset \mathbb{R}^2$. Since $\operatorname{ran}(T) = \mathbb{R}^2$, we have $\operatorname{rank}(T) = 2$.
- The differential operator $\frac{d}{dx}$ is a linear operator $C^1(\mathbb{R}) \to C(\mathbb{R})$[1]. As we know,

$$\ker\big(\frac{d}{dx}\big) = \{f \in C^1(\mathbb{R})\colon f(x) \equiv c \text{ for some } c \in \mathbb{R}\},$$

so that $\ker(\frac{d}{dx}) \cong \mathbb{R}$ is a one-dimensional subspace of $C^1(\mathbb{R})$. Since

$$\frac{d}{dx} \int_0^x f(t)\,dt = f(x) \qquad \text{for any } f \in C(\mathbb{R}),$$

we have $\operatorname{ran}(\frac{d}{dx}) = C(\mathbb{R})$ and $\operatorname{rank}(\frac{d}{dx}) = \infty$.

- *The domain of definition matters*: considered as an operator on $P_n(\mathbb{R})$ the differential operator $\frac{d}{dx}\colon P_n(\mathbb{R}) \to P_n(\mathbb{R})$ still has a one-dimensional kernel (the space of constant polynomials, $P_0(\mathbb{R})$), but its range is now finite-dimensional:

$$\operatorname{ran}\big(\frac{d}{dx}\big) = P_{n-1}(\mathbb{R}) \cong \mathbb{R}^n.$$

This even works for $n = 0$, if we define $P_{-1}(\mathbb{R}) := \mathbb{R}^0 = \{0\}$.[2]

### ℘ A linear transformation is injective if and only if its kernel is trivial

Let $T \in L(X,Y)$. Then

$$T \text{ injective} \quad \Longleftrightarrow \quad \ker(T) = \{\mathbf{0}\}.$$

[1] Here we use the convention that $C^k(\mathbb{R}) = C^k(\mathbb{R},\mathbb{R})$, just as one may write $L(X) = L(X,X)$ for linear transformations on a space $X$.

[2] This is the reason why the degree of the zero polynomial is sometimes taken as $-1$; if $\deg(0) = -1$, then $P_{-1}(\mathbb{R})$ is naturally definied, and the differential operator maps $P_n(\mathbb{R}) \to P_{n-1}(\mathbb{R})$ for all $n \geq 0$.

**Proof**

$$\begin{aligned} T \text{ injective} \quad &\iff \quad \big[Tx = Ty \implies x = y\big] \iff \quad \big[T(x-y) = 0 \implies x - y = 0\big] \\ &\iff \quad \big[Tz = 0 \implies z = 0\big] \quad \iff \quad \ker(T) = \{0\}. \end{aligned}$$

**Ex.**

- A matrix $A \in M_{m\times n}(\mathbb{R})$ describes a linear transformation $\mathbb{R}^n \to \mathbb{R}^m$. This transformation is injective if zero (the zero element in $\mathbb{R}^n$) is the only solution of the corresponding linear homogeneous system:

$$\begin{aligned} &a_{11}x_1 && + \ldots a_{1n}x_n = 0 \\ &\vdots && \quad\vdots \\ &a_{m1}x_1 && + \ldots a_{mn}x_n = 0 \quad \implies \quad (x_1, \ldots, x_n) = (0, \ldots, 0). \end{aligned}$$

## Matrices: null spaces, column spaces and row spaces

Let $A = (a_{ij})_{ij} \in M_{m\times n}(\mathbb{R})$ be the matrix realisation of a linear map $\mathbb{R}^n \to \mathbb{R}^m$.[1]

### The null space of a matrix

In this case the kernel of $A$ is also called the **null space of** $A$:

$$\begin{aligned} x \in \ker(A) \quad &\iff \quad Ax = 0 \quad \iff \quad \sum_{j=1}^n a_{ij}x_j = 0 \quad \forall i = 1, \ldots, m \\ &\iff \quad (x_1, \ldots, x_n) \perp (a_{i1}, \ldots, a_{in}) \qquad \text{for all } i = 1, \ldots, n. \end{aligned}$$

Thus, *the kernel is the space of vectors $x \in \mathbb{R}^n$ which are orthogonal to the row vectors of $A$.*

### The column space of a matrix

The **column space of** $A$ is the range of $A$: since

$$Ax = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1 \begin{bmatrix} a_{11} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ \vdots \\ a_{m2} \end{bmatrix} + \ldots + x_n \begin{bmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{bmatrix},$$

we have that

$$\operatorname{ran}(A) = \{Ax \colon x \in \mathbb{R}^n\} = \Big\{ \sum_{j=1}^n x_j A_j \colon (x_1, \ldots, x_n) \in \mathbb{R}^n \Big\} = \operatorname{span}\{A_1, \ldots, A_n\}$$

is the subspace of $\mathbb{R}^m$ spanned by the column vectors $A_j$, $j = 1, \ldots, n$, of $A$.

### The row space of a matrix

Similarly, define the **row space of** $A$ to be the space spanned by the row vectors of $A$. Then

$$\text{row space of } A = \text{ column space of } A^t,$$

where $A^t = (a_{ji})$ is the transpose of $A = (a_{ij})$.

[1] This realisation is unique as long we have agreed upon a choice of bases for $\mathbb{R}^n$ and $\mathbb{R}^m$. If nothing else is said, we assume that vectors in $\mathbb{R}^n$, $n \in \mathbb{N}$, are expressed in the standard basis.

### ℘ The kernel of a matrix is perpendicular to the range of its transpose

Let $A \in M_{m\times n}(\mathbb{R})$. Then

$$\ker(A) \perp \operatorname{ran}(A^t),$$

meaning that if $x \in \ker(A)$ and $y \in \operatorname{ran}(A^t)$, then $x \cdot y = \sum_{j=1}^{n} x_j y_j = 0$.

**Proof**

As shown above, the null space of $A$ is perpendicular to the row space of $A$. The row space of $A$ equals the column space of $A^t$ (this is the definition of the matrix transpose). The proposition follows.

## The rank–nullity theorem and its consequences

### ℘ The rank–nullity theorem

Let $T \in L(\mathbb{R}^n, \mathbb{R}^m)$. Then

$$\dim\ker(T) + \dim\operatorname{ran}(T) = n.$$

**N.b.** The name comes from that $\dim\ker(T)$ is the **nullity of** $T$. Thus, the sum of the rank and the nullity of $T$ equals the dimension of its ground space (domain).

**Proof**

Pick a basis $e = \{e_1, \dots, e_k\}$ for $\ker(T)$. If $k = n$ and $\ker(T) = \mathbb{R}^n$ we are done, since then

$$\operatorname{ran}(T) = \{Tx \colon x \in \mathbb{R}^n\} = \{0\},$$

so that $\dim\ker(T) + \dim\operatorname{ran}(T) = n$.

Hence, assume that $k < n$ and extend $e$ to a basis $\{e_1, \dots, e_k, f_1, \dots, f_m\}$ for $\mathbb{R}^n$.

This can be done in the following way: pick $f_1 \notin \operatorname{span}\{e_1, \dots, e_k\}$. Then $\{e_1, \dots, e_n, f_1\}$ is linearly independent. If $\operatorname{span}\{e_1, \dots, e_k, f_1\} = \mathbb{R}^n$ we stop. Else, pick $f_2 \notin \operatorname{span}\{e_1, \dots, e_k, f_1\}$. Since $l > n$ vectors are always linearly dependent in $\mathbb{R}^n$, this process stops when $k + m = n$ (it cannot stop before, since then $\mathbb{R}^n$ would be spanned by a set of dimension $< n$, which is impossible; see the definition of vector space dimension).

We now prove that $Tf = \{Tf_1, \dots Tf_m\}$ is a basis for $\operatorname{ran}(T)$.

$Tf$ is linearly independent:

$$\sum_{j=1}^{m} a_j T f_j = 0 \quad\iff\quad T\big(\sum_{j=1}^{m} a_j f_j\big) = 0 \quad\iff\quad \sum_{j=1}^{m} a_j f_j \in \ker(T) \quad\iff\quad a_j = 0\, \forall j = 1, \dots, m,$$

since $T$ is linear, and since, by the construction of $f$, no non-zero linear combination of elements $f_j$ is in $\ker(T)$.

Furthermore, $Tf$ spans $\operatorname{ran}(T)$:

$$\begin{aligned}\mathrm{ran(T)} = \{Tx \colon x \in \mathbb{R}^n\} &= \Big\{T\big(\sum_{j=1}^{k} a_j e_j + \sum_{j=1}^{m} b_j f_j\big) \colon a_j, b_j \in \mathbb{R}\Big\} \\ &= \Big\{T\big(\sum_{j=1}^{k} a_j e_j\big) + T\big(\sum_{j=1}^{m} b_j f_j\big) \colon a_j, b_j \in \mathbb{R}\Big\} = \Big\{\sum_{j=1}^{m} b_j T f_j \colon b_j \in \mathbb{R}\Big\},\end{aligned}$$

since $T$ is linear and $e \subset \ker(T)$.

Hence, $\{Tf_1, \dots, Tf_m\}$ is a basis for $\operatorname{ran}(T)$, and

$$\dim\ker(T) + \dim\operatorname{ran}(T) = k + m = n.$$

### ℘ For finite-dimensional linear transformations, injective means surjective

Let $T \in L(\mathbb{R}^n)$ be a linear transformation $\mathbb{R}^n \to \mathbb{R}^n$. Then the following are equivalent:

- $T$ is injective $\quad (\ker(T) = \{\mathbf{0}\})$
- $T$ is surjective $\quad (\operatorname{ran}(T) = \mathbb{R}^n)$
- $T\colon \mathbb{R}^n \to \mathbb{R}^n$ is invertible
- The matrix representation $A$ of $T$ (in any given basis) is invertible.
- For any $b \in \mathbb{R}^n$ the system $Ax = b$ has a unique solution $x$.

**Proof**

**(i) $\Longleftrightarrow$ (ii):** When $m = n$, the rank–nullity theorem says that $\operatorname{ran}(T) = \mathbb{R}^n$ (so that $T$ is surjective) exactly when $\ker(T) = \{\mathbf{0}\}$ (so that $T$ is injective).

**(i,ii) $\Longleftrightarrow$ (iii):** A function is bijective exactly if it is both invertible and surjective.

**(iii) $\Longleftrightarrow$ (iv):** Given any basis for $\mathbb{R}^n$, $T$ has a unique matrix representation $A$ (defined by its action on the basis vectors). If the inverse matrix $A^{-1}$ exists, then there exists a corresponding linear transformation $S$ such that $ST = ST = \mathrm{id}$ (since $A^{-1}A = AA^{-1} = I$, and the identity map $\mathrm{id}\colon \mathbb{R}^n \to \mathbb{R}^n$ has the identity matrix $I$ as representation in all bases). Thus $S = T^{-1}$ is the inverse of $T$. If, on the other hand, $T^{-1}$ exists, it must by the same argument have a matrix representation $B$ such that $AB = BA = I$. Hence, $A^{-1} = B$ exists.

**(iv) $\Longleftrightarrow$ (v):** If $A$ is invertible it is immediate that $x = A^{-1}b$ is the unique solution. If, on the other hand, $Ax = b$, has a unique solution $x$ for any $b$, we construct a matrix $B$ by taking as its columns $x_j$ such that $Ax_j = e_j$, where $\{e_1, \dots, e_n\}$ is the standard basis. This guarantees that $B = A^{-1}$ is the inverse matrix of $A$. (A less constructive argument would be to note that $Ax = b$ is uniquely solvable for all $b \in \mathbb{R}^n$ exactly if $T$ is invertible.)

### Geometric interpretation of the rank–nullity theorem

Define the **direct sum** $X \oplus Y$ of two vector spaces (both real, or both complex) as the space of pairs $(x, y)$ with the naturally induced vector addition and scalar multiplication:

$$X \oplus Y \overset{\text{def.}}{=} \{(x, y) \in X \times Y\},$$

where

$$(x_1, y_1) + (x_2, y_2) \overset{\text{def.}}{=} (x_1 + x_2, y_1 + y_2) \quad \text{and} \quad \lambda(x, y) \overset{\text{def.}}{=} (\lambda x, \lambda y).$$

If $X, Y \subset V$ are subspaces of a vector space $V$, then

$$X \oplus Y = V \quad \Longleftrightarrow \quad X \cap Y = \{0\} \quad \text{and} \quad X + Y \overset{\text{def.}}{=} \{x + y\colon x \in X, y \in Y\} = V,$$

where the equality $X \oplus Y = V$ should be interpreted in terms of isomorphisms ($V$ can be represented as $X \oplus Y$). Note that

$$\dim(X \oplus Y) = \dim(X) + \dim(Y).$$

With these definitions, the rank–nullity theorem can be expressed as a geometric description of the underlying space ($\mathbb{R}^n$) in terms of the matrix $A$.

### ℘ Rank–nullity theorem: geometric version

Let $A \in M_{m\times n}(\mathbb{R})$. Then

$$\mathbb{R}^n = \ker(A) \oplus \operatorname{ran}(A^t).$$

**N.b.** A consequence of this is that $\operatorname{rank}(A) = \operatorname{rank}(A^t)$; another is that $\mathbb{R}^m = \ker(A^t) \oplus \operatorname{ran}(A)$.

**Proof**

We have already showed that $\ker(A) \perp \operatorname{ran}(A^t)$ in $\mathbb{R}^n$, so that

$$\ker(A) \cap \operatorname{ran}(A^t) = \{0\};$$

this is a consequence of that $|x|^2 = x \cdot x = 0$ for any $x \in \ker(A) \cap \operatorname{ran}(A^t)$.

It remains to show that $\ker(A) \oplus \operatorname{ran}(A^t)$ make up of all $\mathbb{R}^n$. Since $\ker(A) \perp \operatorname{ran}(A^t)$ in $\mathbb{R}^n$, we have

$$\operatorname{rank}(A^t) \leq n - \dim(\ker(A) = \operatorname{rank}(A),$$

where the last equality follows is the rank–nullity theorem. But this argument is not dependent on $A$; hence

$$\operatorname{rank}(A) = \operatorname{rank}((A^t)^t) \leq \operatorname{rank}(A^t),$$

and

$$\operatorname{rank}(A) = \operatorname{rank}(A^t).$$

This shows that

$$\ker(A) \oplus \operatorname{ran}(A^t) = \mathbb{R}^n$$

constitute all of $\mathbb{R}^n$.

### ℘ Summary on linear equations (the Fredholm alternative)

Let $A \in M_{m\times n}(\mathbb{R})$ be the realisation of a linear transformation $\mathbb{R}^n \to \mathbb{R}^m$, and consider the linear equation

$$Ax = b.$$

- Either $b \in \operatorname{ran}(A)$ and the equation is solvable, or $b \in \ker(A^t)$ and there is no solution.
- In case $\ker(A) = \{0\}$ any solution is unique, else the solutions can be described as

  $$x_p + \ker(A),$$

  where $x_p$ is any (particular) solution of $Ax = b$.

**Proof**

The first statement is a reformulation of the geometric version of the rank–nullity theorem; the second follows from that

$$\begin{cases} Ax = b \\ Ay = b \end{cases} \iff \begin{cases} Ax = b \\ A(x-y) = 0 \end{cases} \iff \begin{cases} Ax = b \\ y = x + z, \quad z \in \ker(A). \end{cases}$$

If $\ker(A) = \{0\}$ there is at most one solution, $x$, else the solution space is an affine space[1] of the same dimension as $\ker(A)$.

## 2.8 Bounded linear transformations

Let $X$ and $Y$ be normed spaces (both real, or both complex), and $T \in L(X, Y)$ a linear mapping between them.

### Boundedness

A linear mapping $T\colon X \to Y$ is called **bounded**, $T \in B(X, Y)$, if $T$ maps bounded sets into bounded sets:

$$T \in B(X, Y) \quad \overset{\text{def}}{\iff} \quad \exists C; \quad \|Tx\|_Y \le C\|x\|_X \quad \text{for all } x \in X.$$

Thus, if $T$ is bounded, the number

$$\|T\| \overset{\text{def.}}{=} \sup_{x \ne 0} \frac{\|Tx\|_Y}{\|x\|_X}$$

is finite; it is the **(operator) norm of** $T$.

**N.b.** In some sources $B(X, Y)$ is denoted by $L(X, Y)$; to us, $L(X, Y)$ is the space of linear transformations between two (not necessarily normed) vector spaces; if $X$ and $Y$ are normed spaces, then $B(X, Y) \subset L(X, Y)$.

*From now on, we will not always write out the indices for the norms; just recall that $x \in X$ and $Lx \in Y$.*

**Ex.**

- The integral $\int_0^t f(s)\,ds$ defines a linear transformation on the space of bounded and continuous functions $f\colon [0,1] \to \mathbb{R}$,

$$T\colon BC([0,1], \mathbb{R}) \to BC([0,1], \mathbb{R}), \qquad (Tf)(t) = \int_0^t f(s)\,ds.$$

  This transformation is bounded, since

$$\|Tf\|_{BC([0,1],\mathbb{R})} = \sup_{t\in[0,1]} \left| \int_0^t f(s)\,ds \right| \le \int_0^1 \max_{s\in[0,1]} |f(s)|\,ds = \|f\|_{BC([0,1],\mathbb{R})},$$

  so that $\|T\| \le 1$. In fact, $\|T\| = 1$ (can you see why?).

- If $g \in BC([0,1], \mathbb{R})$, a similar argument yields that

$$T\colon BC([0,1], \mathbb{R}) \to BC([0,1], \mathbb{R}), \qquad (Tf)(t) = \int_0^t f(s)g(s)\,ds$$

  is bounded too, with $\|T\| \le \max_{t\in[0,1]} |g(t)| = \|g\|_{BC([0,1],\mathbb{R})}$ (see if you can make this better).

[1] An affine space is a 'translation' of a vector space (or a vector space which has lost its origin); affine spaces are not themselves vector spaces, since they do not have any zero element.

- By the same argument, with $t = 1$ and $g \in BC([0,1], \mathbb{R})$, the definite integral $\int_0^1 f(s)g(s)\,ds$ defines a **bounded linear functional**[1]

$$T \colon BC([0,1], \mathbb{R}) \to \mathbb{R}, \qquad f \mapsto \int_0^1 f(s)g(s)\,ds.$$

Remember the form of this functional – it is the inner product for the Hilbert space $L_2((0,1), \mathbb{R})$.

- The derivative $\frac{d}{dx}$ is in general *not* a bounded operator.[2] To see why, consider a sequence of functions like

$$f_n(x) := \sin(nx).$$

These functions are uniformly bounded, but not their derivatives. This indicates that, to solve differential equations, it is better to reformulate them as integral operators.

**Equivalence of norm expressions**
For $T \in B(X,Y)$,

$$\|T\| = \sup_{x \neq 0} \frac{\|Tx\|}{\|x\|} = \sup_{\|x\|=1} \|Tx\| = \sup_{\|x\| \leq 1} \|Tx\|$$

all describe the least possible bound on $C$ such that $\|Tx\| \leq C\|x\|$ for all $x \in X$.

**Proof**

Since $T$ is linear, and since norms are positively homogeneous, we get

$$\frac{\|Tx\|}{\|x\|} = \Big\| \frac{1}{\|x\|} Tx \Big\| = \Big\| T\big(\frac{x}{\|x\|}\big) \Big\|, \qquad x \neq 0.$$

Note that the mapping $S_\lambda \to S_1$, $x \mapsto \frac{x}{\|x\|}$, is bijective. Thus, if $\lambda > 0$, we have

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

By considering all fixed, but different, $\lambda > 0$, we see that

$$\sup_{\|x\|=\lambda>0} \frac{\|Tx\|}{\|x\|} = \sup_{\|x\|=1} \|Tx\|, \qquad \text{so that} \qquad \sup_{x \neq 0} \frac{\|Tx\|}{\|x\|} = \sup_{\|x\|=1} \|Tx\|.$$

Then, consider $\lambda \leq 1$ to see that

$$\sup_{\|x\| \leq 1} \|Tx\| \overset{\lambda \leq 1}{\leq} \sup_{\|x\|=1} \|Tx\| \leq \sup_{\|x\| \leq 1} \|Tx\|,$$

where the last inequality follows from the definition of the supremum. This proves the assertion.

### ℘ B(X,Y) is a normed space

$$B(X,Y) = \{T \in L(X,Y) \colon T \text{ is bounded}\}$$

is a normed space when equipped with the operator norm $\|\cdot\|$.

[1] A functional is a function from a vector space to its field of scalars ($\mathbb{R}$ or $\mathbb{C}$).

[2] Unless we consider it on some special space of functions, as the finite-dimensional space of real polynomials of degree less than $n$.

**Proof**

We already know that $L(X,Y)$ is a linear space, so it remains to show that $B(X,Y)$ is a subspace and $\|\cdot\|$ a norm on $B(X,Y)$.

**Subspace property**. take $T,S\in B(X,Y)$ and $\mu,\lambda$ scalars. Then

$$\begin{aligned}\sup_{\|x\|\le 1}\|(\mu T+\lambda S)(x)\|_Y &\le \sup_{\|x\|\le 1}\big(|\mu|\|Tx\|_Y+|\lambda|\|Sx\|_Y\big)\\ &\le |\mu|\sup_{\|x\|\le 1}\|Tx\|_Y+|\lambda|\sup_{\|x\|\le 1}\|Sx\|_Y=|\mu|\|T\|+|\lambda|\|S\|\end{aligned}$$

is finite by choice of $S,T$. Thus $\mu T+\lambda S$ is bounded if $T$ and $S$ are bounded, so that $B(X,Y)$ is a subspace of $L(X,Y)$.

**Norm properties**. These are consequences of that the operator norm $\|\cdot\|$ is defined using (primarily) the norm of $Y$.

**Positive definiteness**:

$$\|T\|=0 \quad\Longleftrightarrow\quad \|Tx\|_Y=0 \quad \forall x\in X \quad\Longleftrightarrow\quad T\equiv 0 \quad \text{in } L(X,Y).$$

**Positive homogeneity**:

$$\|\lambda T\|=\sup_{\|x\|_X\le 1}\|\lambda Tx\|_Y=|\lambda|\sup_{\|x\|_X\le 1}\|Tx\|_Y=|\lambda|\|T\|.$$

**Triangle inequality**:

$$\begin{aligned}\|T+S\|=\sup_{\|x\|_X\le 1}\|(T+S)x\|_Y &\le \sup_{\|x\|_X\le 1}\big(\|Tx\|_Y+\|Sx\|_Y\big)\\ &\le \sup_{\|x\|_X\le 1}\|Tx\|_Y+\sup_{\|x\|_X\le 1}\|Sx\|_Y=\|T\|+\|S\|.\end{aligned}$$

**Ex.**

- As we shall see, $B(\mathbb{R}^n,\mathbb{R}^m)=L(\mathbb{R}^n,\mathbb{R}^m)$ (as sets and linear spaces): given bases for $\mathbb{R}^n,\mathbb{R}^m$ there is a bijective correspondence between matrices $A\in M_{m\times n}(\mathbb{R})$ and bounded linear transformations $T\in B(\mathbb{R}^n,\mathbb{R}^m)$.
- Let $X$ be a real normed space. The space $X':=B(X,\mathbb{R})$ is called the **dual of** $X$; its elements are **bounded linear functionals** on $X$. If $X$ is complex, $B(X,\mathbb{C})$ is its dual.[1]
- The dual of $\mathbb{R}$ is $\mathbb{R}$: each bounded linear functional $T\in B(\mathbb{R},\mathbb{R})$ is realized by multiplication with a real constant:

$$T\in B(\mathbb{R},\mathbb{R}) \quad\Longleftrightarrow\quad Tx=\lambda x, \quad \lambda\in\mathbb{R}.$$

- **Riesz representation theorem**: Let $L_2(I,\mathbb{R})$ be the space of real-valued square-integrable functions on an open interval $I\subset\mathbb{R}$, with norm

$$\|f\|_{L_2(I,\mathbb{R})}=\Big(\int_I |f(t)|^2\,dt\Big)^{1/2}.$$

The Riesz representation theorem asserts that each bounded linear functional $T$ on $L_2(I,\mathbb{R})$ can be identified with an element $g\in L_2(I,\mathbb{R})$, via

$$Tf=\int_I f(t)g(t)\,dt.$$

Thus $L_2(I,\mathbb{R}) \cong B(L_2(I,\mathbb{R}),\mathbb{R})$ is its own dual.[2]

### $\wp$ B(X,Y) is Banach for Y Banach

If $Y$ is complete, so is $B(X,Y)$.

**N.b.** Note that $X$ has no role in the completeness of $B(X,Y)$.[3]

**Proof**

**Pointwise convergence.** Let $\{T_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence in $B(X,Y)$. Then, for each fixed $x \in X$,

$$\|(T_n - T_m)x\|_Y \le \|T_m - T_n\|\|x\|_X \overset{m,n\to\infty}{\to} 0,$$

so that $\{T_n x\}_{n\in\mathbb{N}}$ is Cauchy in $Y$. By assumption, $Y$ is complete, so $\{T_n x\}_{n\in\mathbb{N}}$ is convergent. Define

$$Tx := \lim_{n\to\infty} T_n x, \qquad x \in X.$$

**The pointwise limit defines a linear and bounded transformation.** With this construction $T\colon X \to Y$ is linear,

$$T(\lambda x + \mu y) = \lim_{n\to\infty} T_n(\lambda x + \mu y) = \lim_{n\to\infty} (\lambda T_n x + \mu T_n y) = \lambda \lim_{n\to\infty} T_n x + \mu \lim_{n\to\infty} T_n y = \lambda Tx + \mu Ty,$$

and for each fixed $x \in X$ there exists $n_\varepsilon$ (depending also on $x$), such that, for all $n \ge n_\varepsilon$,

$$\begin{aligned}\|Tx\|_Y \le \|(T - T_n)x\|_Y + \|T_n x\|_Y &\le \varepsilon + \|T_n x\|_Y \\ &\le \varepsilon + \underbrace{\sup_{n\in\mathbb{N}} \|T_n\|}_{\text{finite}} \|x\|_X,\end{aligned}$$

where we have used that Cauchy sequences are bounded (so that $\|T_n\|$ is bounded, uniformly for $n \in \mathbb{N}$). By taking the supremum over all $x$ with $\|x\|_X = 1$, we obtain that $T$ is bounded.

**Convergence in** $B(X,Y)$. It remains to show that $T_n \to T$ in $B(X,Y)$. Similarly to the above argument, if $m \ge n_{\varepsilon/2}$ (depending also on $x$), we have

$$\begin{aligned}\|(T - T_n)x\|_Y \le \|(T - T_m)x\|_Y + \|(T_m - T_n)x\|_Y &\le \tfrac{\varepsilon}{2} + \|(T_m - T_n)x\|_Y \\ &< \tfrac{\varepsilon}{2} + \|T_m - T_n\|\|x\|_X.\end{aligned}$$

Since $\{T_n\}_{n\in\mathbb{N}}$ is Cauchy, there exists $N_{\varepsilon/2}$ such that

$$\|T_n - T_m\| < \frac{\varepsilon}{2} \quad \text{for} \quad m,n \ge N_{\varepsilon/2}.$$

Choose $n \ge N_{\varepsilon/2}$ and, for each $x$, an appropriate $m \ge \max\{n_{\varepsilon/2}, N_{\varepsilon/2}\}$ By taking the supremum over all $x$ with $\|x\|_X = 1$ we thus find

$$\|T - T_n\| < \varepsilon \quad \text{for} \quad n \ge N_{\varepsilon/2}.$$

Hence, $\lim_{n\to\infty} T_n = T$ in $B(X,Y)$.

[1] This notion of dual coincides with that of a **continuous dual**; it is possible to define more general duals.

[2] So far, this identification is in terms of linear spaces, but we will see later that it extends to inner products (and therefore to norms).

[3] If, however, $X$ is non-trivial, i.e., if $X \neq \{\mathbf{0}\}$, then the converse also holds, so that $B(X,Y)$ is complete if and only if $Y$ is complete.

## Boundedness and continuity

### Continuity

A mapping $f\colon X \to Y$ between to metric spaces is said to be **continuous at** $x_0$ if

$$f(x_n) \to f(x_0) \;\text{ in } Y \qquad \text{as} \qquad x_n \to x_0 \;\text{ in } X.$$

Since continuous and sequential limits agree, this is the same as

$$\forall\, \varepsilon > 0 \quad \exists\, \delta > 0; \qquad d_Y(f(x), f(x_0)) < \varepsilon \quad \text{for} \quad d_X(x, x_0) < \delta.$$

A mapping that is continuous at all points in $X$ is called **continuous**.

**Ex.**

- In a normed space, $(X, \|\cdot\|)$, the norm is a continuous function $X \to \mathbb{R}$: if $x_n \to x_0$ in $X$, then

  $$d_{\mathbb{R}}(\|x_n\|, \|x_0\|) = \big|\|x_n\| - \|x_0\|\big| \le \|x_n - x_0\| = d_X(x_n, x_0) \to 0,$$

  by the reverse triangle inequality.

### ℘ For linear operators, continuity means boundedness

Let $T \in L(X, Y)$. Then the following statements are equivalent:

- $T$ is everywhere continuous.
- $T$ is continuous at $x = 0$.
- $T$ is bounded.

**Proof**

First, note that for any fixed $x_0 \in X$,

$$\begin{aligned} Tx_n \to Tx_0 \quad \text{as} \quad x_n \to x_0 \quad &\Longleftrightarrow \quad T(x_n - x_0) \to \mathbf{0}_Y \quad \text{as} \quad (x_n - x_0) \to \mathbf{0}_X \\ &\overset{z_n = x_n - x_0}{\Longleftrightarrow} \quad Tz_n \to \mathbf{0}_Y \quad \text{as} \quad z_n \to \mathbf{0}_X, \end{aligned}$$

so that, for linear operators, continuity at the origin is the same as continuity everywhere ($x_0$ is arbitrary).

To see that boundedness and continuity at the origin are equivalent, assume first that $T$ is bounded. Then

$$\|Tx\|_Y \le \|T\| \|x\|_X \to 0 \quad \text{as} \quad \|x\|_X \to 0,$$

so that $T$ is also continuous. Contrariwise, assume that $T$ is continuous at the origin. Then

$$\|Tx\|_Y = \|Tx - T\mathbf{0}\|_Y \le \varepsilon \quad \text{for} \quad \|x\|_X = \|x - \mathbf{0}\|_X \le \delta.$$

But $T$ is linear, so by scaling $x$ (replace $x$ with $\delta x$) we obtain

$$\|Tx\|_Y \le \frac{\varepsilon}{\delta} \quad \text{for} \quad \|x\|_X \le 1.$$

Thus $T$ is bounded.

**Ex.**

- Any linear operator $T \in L(X,Y)$ defined on a *finite-dimensional* normed space $X$ is continuous. Reason: identify $X \cong \mathbb{R}^n$ and note that

$$\operatorname{ran}(T) = \operatorname{span}\{Te_1, \ldots, Te_n\} \cong \mathbb{R}^m \quad \text{for some } m \leq n,$$

where $\{e_1, \ldots, e_n\}$ is a basis for $\mathbb{R}^n$. Hence,

$$T\colon X \cong \mathbb{R}^n \to \mathbb{R}^m \cong \tilde{Y} \subset Y$$

is a linear transformation onto a finite-dimensional subspace $\tilde{Y}$ of $Y$, and, as such, has a matrix representation

$$T\colon x \mapsto Ax = \big(\sum_{j=1}^{n} a_{ij}x_j\big)_{i=1}^{m}.$$

All norms on a finite-dimensional vector space are equivalent, so whatever the norms of $X$ and $Y$, we can consider any suitable norms for $\mathbb{R}^n \cong X$ and $\mathbb{R}^m \cong \tilde{Y}$. Choose, for example, the $l_\infty$-norm: then

$$\|Ax\|_{l_\infty} = \max_{1\leq i\leq m} \big|\sum_{j=1}^{n} a_{ij}x_j\big| \leq n \max_{i,j} |a_{ij}| \max_{j} |x_j| = n \max_{i,j} |a_{ij}| \|x\|_{l_\infty}.$$

This means that $T$ is bounded with $\|T\| \leq n \max_{i,j} |a_{ij}|$, and therefore also continuous.

**N.b.** Equivalent norms yield the same open and closed sets, the same convergence, but *not the same constants* in the estimates – in particular, the exact value of $\|T\|$ depends on the norms for $X$ and $Y$.

### ℘ The kernel of a bounded operator is closed

Let $T \in B(X,Y)$. Then $\ker(T)$ is a closed subspace of $X$. In particular, if $X$ is a Banach space, so is $\ker(T)$.

**Proof**

Take

$$\{x_n\}_{n\in N} \subset \ker(T); \qquad \lim_{n\to\infty} x_n = x_0 \in X.$$

We want to show that $x_0 \in \ker(T)$. But this follows from the continuity of $T$:

$$\|Tx_0\|_Y = \|Tx_0 - Tx_n\|_Y \leq \|T\|\|x_0 - x_n\|_X \to 0 \quad \text{as} \quad x_n \to x_0 \qquad \Longrightarrow \qquad Tx_0 = 0.$$

If, furthermore, $X$ is complete, so is the closed subspace $\ker(T) \subset X$.

**Ex.**

- The null space of a matrix $A \in M_{m\times n}(\mathbb{R})$ is a closed subspace of $\mathbb{R}^n$.
- In $L_2((-\pi,\pi),\mathbb{R})$, the kernel of the bounded linear functional

  $$T\colon f \mapsto \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin(t)\,dt$$

  is a closed subspace; it consists of all functions with zero Fourier coefficient before $\sin(t)$ in its Fourier expansion.

# Chapter 3

# Solving differential equations

## 3.1 General existence theorems

Let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^n$.

**Initial-value problems**
Let $(t_0, x_0)$ be a fixed point in an open subset $I \times U \subset \mathbb{R} \times \mathbb{R}^n$, and $f \in C(I \times U, \mathbb{R}^n)$ a continuous vector-valued function on this subset. The problem of finding $x \in C^1(J, U)$ such that

$$\dot{x}(t) = f(t, x(t)), \qquad x(t_0) = x_0, \qquad\qquad \text{(IVP)}$$

for some possibly smaller interval $J \subset I$ is called an **initial-value problem**. Here, $\dot{x} = \frac{d}{dt}x$.

$\wp$ **Reformulation of real-valued ODEs as first-order systems**
Any ordinary differential equation

$$x^{(n)}(t) = g(t, x(t), \dot{x}(t), \ldots, x^{(n-1)}(t)),$$

with initial conditions

$$x(t_0) = x_1, \quad \dot{x}(t_0) = x_2, \quad \ldots, \quad x^{(n-1)}(t_0) = x_n,$$

and $g$ continuous in some open set $I \times U \subset \mathbb{R} \times \mathbb{R}^n$ containing $(t_0, x_1, \ldots, x_n)$, can be reformulated in the form (IVP).

**Proof**

Let

$$y_0 := x, \quad y_1 := \dot{x}, \quad \ldots, \quad y_{n-1} := x^{(n-1)}.$$

Then

$$\begin{bmatrix} \dot{y}_0 \\ \vdots \\ \dot{y}_{n-2} \\ \dot{y}_{n-1} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n-1} \\ g(t, y_0, \ldots, y_{n-1}) \end{bmatrix}$$

describes $y = (y_0, \ldots, y_{n-1})$ as a function $I \to U \subset \mathbb{R}^n$ for some interval $I \subset \mathbb{R}$; the function $f \in C(I \times U, \mathbb{R}^n)$ is the vector-valued function given by the right-hand side of this system. The initial condition is $y(t_0) = (x_1, \ldots, x_n)$.

**Ex.**

- The second-order ordinary differential equation

$$\ddot{x} + \sin(x) = 0, \qquad x(0) = 1, \quad \dot{x}(0) = 2,$$

is equivalent to the system

$$\begin{bmatrix}\dot{y}_0\\ \dot{y}_1\end{bmatrix} = \begin{bmatrix} y_1\\ -\sin(y_0)\end{bmatrix} \quad \text{with} \quad \begin{bmatrix} y_0\\ y_1\end{bmatrix}_{t=0} = \begin{bmatrix}1\\2\end{bmatrix}.$$

In this case

$$f\colon \mathbb{R}^2 \to \mathbb{R}^2, \qquad \begin{bmatrix} y_0\\ y_1\end{bmatrix} \mapsto \begin{bmatrix} y_1\\ -\sin(y_0)\end{bmatrix}$$

is independent of time.

### $\wp$ The Peano existence theorem

For any $(t_0, x_0) \in I \times U$ there exists $\varepsilon > 0$ such that the initial-value problem (IVP) has a solution defined for $|t - t_0| < \varepsilon$. The solution $x = x(\cdot; t_0, x_0) \in C^1(B_\varepsilon(t_0), U)$.

**N.b.** The Peano existence theorem guarantees the existence of (local) solutions, but not their uniqueness.

**Ex.**

- The initial-value problem

$$\begin{cases} \dot{x} = \frac{3}{2}x^{1/3}, & t \geq 0,\\ \dot{x} = 0, & t < 0, \end{cases} \qquad x(0) = 0,$$

has the trivial solution $x \equiv 0$, but also the ones given by

$$\begin{cases} x(t) = \pm t^{3/2}, & t \geq 0,\\ x(t) = 0, & t < 0. \end{cases}$$

To remedy this lack of uniqueness in Peano's theorem one needs the concept of Lipschitz continuity.

### Lipschitz continuity

A continuous function $f \in C(I \times U, \mathbb{R}^n)$ is said to be locally **Lipschitz continuous** with respect to its second variable $x \in U$ if for any $(t_0, x_0) \in I \times U$ there exists $\varepsilon, L > 0$ with

$$|f(t,x) - f(t,y)| \leq L\,|x - y|, \qquad \text{for all} \quad (t,x), (t,y) \in B_\varepsilon(t_0, x_0).$$

The set of locally Lipschitz continuous functions on $I \times U$ form a vector space, $Lip(I \times U, \mathbb{R}^n)$. *If the Lipschitz constant $L$ does not depend on the point $(t_0, x_0)$, then the Lipschitz condition is said to be* ***uniform*** ($f$ is then uniformly Lipschitz continuous). *A locally Lipschitz continuous function is uniformly Lipschitz continuous on any compact set.*[1]

**N.b.** Any continuously differentiable function is also locally Lipschitz continuous, and hence unformly Lipschitz on any compact set.

[1] A compact set in a metric space is a set in which every sequence has a convergent subsequence (converging to an element in the set). The Bolzano–Weierstrass theorem states that, in $\mathbb{R}^n$, the compact sets are exactly those that are closed and bounded.

**Proof (for anyone interested in ODEs)**

If $f$ is continuously differentiable, then

$$f(t,x) - f(t,y) = \int_0^1 \frac{d}{ds} f(t, sx + (1-s)y)\, ds = \left( \int_0^1 D_x f(t, sx + (1-s)y)\, ds \right) (x-y),$$

where $D_x f$ is the Jacobian matrix of $f(t,\cdot)$, and $(x-y) \in \mathbb{R}^n$ is a vector. For each fixed triple $(t,x,y)$ the matrix

$$T(t,x,y) := \int_0^1 D_x f(t, sx + (1-s)y)\, ds$$

defines a bounded linear transformation on $\mathbb{R}^n$, so that

$$|T(t,x,y)z| \leq \|T(t,x,y)\| |z|, \qquad z \in \mathbb{R}^n.$$

Since $D_x f$ is continuous, so is $T$ (in all its variables); and since norms are continuous, the composition

$$(t,x,y) \mapsto \|T(t,x,y\| \text{ is continuous} \qquad \Longrightarrow \qquad C_{t_0,x_0,y_0} := \max_{\substack{|t-t_0| \leq \varepsilon \\ |x-x_0| \leq \varepsilon \\ |y-y_0| \leq \varepsilon}} \|T(t,x,y)\| < \infty.$$

Taken together, we have that

$$|f(t,x) - f(t,y)| = |T(t,x,y)(x-y)| \leq \|T(t,x,y)\| |x-y| \leq C_{t_0,x_0,y_0} |x-y|$$

for all $(t,x), (t,y) \in B_\varepsilon(t_0,x_0)$. (The ball $B_\varepsilon(t_0,x_0)$ is actually a bit smaller than the square described by $|x-x_0| \leq \varepsilon, |t-t_0| \leq \varepsilon$, where we have proved the statement, but we do not need more.)

**Ex.**

Consider $f\colon \mathbb{R} \to \mathbb{R}$ (one spatial variable, no time).

- $x \mapsto \sin(x)$ is continuously differentiable. It is also uniformly Lipschitz, since

$$|\sin(x) - \sin(y)| \leq \max_{\xi \in \mathbb{R}} |\cos(\xi)||x - y|.$$

- $x \mapsto x^2$ is continuously differentiable. It is locally Lipschitz, since

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

- $x \mapsto |x|$ is not continuously differentiable. It is however (uniformly) Lipschitz, since

$$||x| - |y|| \leq |x - y|.$$

- $x \mapsto \sqrt{|x|}$ is continuous but not locally Lipschitz, since it cannot have a finite Lipschitz constant at $x_0 = 0$:

$$\frac{\sqrt{|x|}}{|x|} \to \infty \text{ as } x \to 0.$$

In particular, this shows that $C^1(\mathbb{R}) \subsetneq \mathit{Lip}(\mathbb{R}) \subsetneq C^0(\mathbb{R})$.[1]

## The Banach fixed-point theorem and its applications

To solve the initial-value problem (IVP) we shall reformulate it as

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds, \qquad x \in BC(I, U),$$

where the right-hand side defines a (not necessarily linear) mapping

$$T\colon BC(J, U) \to BC(J, U), \quad x \mapsto x_0 + \int_{t_0}^{t} f(s, x(s))\, ds,$$

for some smaller interval $J = [t_0 - \varepsilon, t_0 + \varepsilon] \subset I$. This is because, if $x$ and $f$ are continuous, so is $s \mapsto f(s, x(s))$, so the integral $\int_{t_0}^{t} f(s, x(s))\, ds$ is continuous (even $C^1$) and bounded on compact intervals. The idea then is that, if $f$ is also Lipschitz, then $T$ **contracts** points for small $|t - t_0| \leq \varepsilon$:

$$\begin{aligned}
|Tx(t) - Ty(t)| = \left| \int_{t_0}^{t} \big(f(s, x(s)) - f(s, y(s))\big)\, ds \right| &\leq \int_{t_0}^{t} \big|f(s, x(s)) - f(s, y(s))\big|\, ds \\
&\leq \int_{t_0}^{t} L|x(s) - y(s)|\, ds \leq L|t - t_0|\|x - y\|_{BC(J,U)}
\end{aligned}$$

Thus, if $\varepsilon L < 1$, taking the maximum over $t \in J$ yields

$$\|Tx - Ty\|_{BC(J,U)} \leq \lambda \|x - y\|_{BC(J,U)}, \qquad \text{for } \lambda = \varepsilon L < 1,$$

so that $Tx$ and $Ty$ are closer to each other than $x$ and $y$. As we shall now see, that gives us a local and unique solution of our problem.

[1] This is the reason why Lipschitz continuity is sometimes denoted by $C^{1-}$; Lipschitz is just slightly worse than being continuously differentiable. In this notation $\mathit{Lip}(I \times U, \mathbb{R}^n) = C^{0,1-}(I \times U, \mathbb{R}^n)$, to clarify that $f$ is continuous with respect to its first variable and Lipschitz with respect to its second.

**Contractions**
Let $(X, d)$ be metric space. A mapping $T : X \to X$ is called a **contraction** if there exists $\lambda < 1$ such that

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

In particular, contractions are continuous.

**N.b.** The uniformity of the constant $\lambda < 1$ is important; it is not enough that $d(T(x), T(y)) < d(x, y)$ for each pair $(x, y) \in X \times X$.

$\wp$ **The Banach fixed-point theorem**
Let $T$ be a contraction on a complete metric space $(X, d)$ with $X \neq \emptyset$. Then there exists a unique $x \in X$ such that $T(x) = x$.

**Proof (*if you are to learn one proof, this is the one*)**

**Existence of a candidate for** $x$: Let $x_0 \in X$,

$$x_1 := T(x_0), \qquad x_{n+1} := T(x_n) = T^{n+1}(x_0), \quad n \in \mathbb{N}.$$

For $n > m \geq n_0$, we have that

$$\begin{aligned} d(x_n, x_m) &\overset{\Delta\text{-ineq.}}{\leq} \sum_{k=m+1}^{n} d(x_k, x_{k-1}) \overset{\text{def.} x_n}{=} \sum_{k=m+1}^{n} d\left(T^k(x_0), T^{k-1}(x_0)\right) \\ &\overset{\text{contr.}}{\leq} \sum_{k=m+1}^{n} \lambda^{k-1} d(x_1, x_0) = d(x_1, x_0)\, \lambda^m \sum_{k=0}^{n-m-1} \lambda^k \\ &\overset{\text{geom. series}}{=} d(x_1, x_0)\, \lambda^m \, \frac{1 - \lambda^{n-m}}{1 - \lambda} \leq \frac{\lambda^{n_0}}{1 - \lambda} d(x_1, x_0) \overset{n_0 \to \infty}{\to} 0. \end{aligned}$$

Thus $\{x_n\}_n$ is Cauchy. By assumption, $(X, d)$ is complete, so there exists $x := \lim_{n\to\infty} x_n \in X$.

$x$ **is a fixed point for** $T$:

$$\begin{aligned} 0 \leq d(x, T(x)) &\leq d(x, x_n) + d(x_n, T(x_n)) + d(T(x_n), T(x)) \\ &\leq d(x, x_n) + d(x_n, x_{n+1}) + \lambda\, d(x_n, x) \\ &\leq \underbrace{d(x, x_n)}_{\to 0} + \underbrace{\lambda^n}_{\to 0} d(x_0, x_1) + \lambda \underbrace{d(x_n, x)}_{\to 0} \to 0, \quad \text{as} \quad n \to \infty. \end{aligned}$$

That is: $x = T(x)$ is a fixed point for $T$.

**Uniqueness**: Assume that $y = T(y)$. Then

$$0 \leq d(x, y) = d(T(x), T(y)) \leq \lambda\, d(x, y) \quad \overset{\lambda<1}{\Longrightarrow} \quad d(x, y) = 0 \quad \Longrightarrow \quad y = x.$$

**Well-posedness for the initial-value problem (IVP)**
$\wp$ **The Picard–Lindelöf theorem**
Let $f : I \times U \to \mathbb{R}^n$ be locally Lipschitz continuous with respect to its second variable and $(t_0, x_0)$ a point in $I \times U$ determining the initial data. Then, for each $\eta > 0$, there exists $\varepsilon > 0$ such that the initial-value problem (IVP) has a unique solution $x \in C^1(\overline{B_\varepsilon(t_0)}, \overline{B_\eta(x_0)})$.

**Proof (as outlined above)**

**Equivalence of formulations:** If $x \in C^1(I,U)$ solves (IVP), then

$$x(t) = x_0 + \int_{t_0}^{t} f(s,x(s))\,ds, \tag{1}$$

by integration. Contrariwise, if $x \in C(I,U)$ fulfils (1), then $x$ is a $C^1(I,U)$-solution (this follows from the Fundamental Theorem of Calculus). Thus, the initial-value problem (IVP) for $x \in C^1(I,U)$ is equivalent to (1) for $x \in C^0(I,U)$.

**Some constants:** Let $\delta > 0$ be such that $[t_0-\delta, t_0+\delta] \subset I$. Fix an arbitrary constant $\eta > 0$. Let

$$R := [t_0-\delta, t_0+\delta] \times \overline{B_\eta(x_0)}, \qquad M := \max_{(t,x)\in R} |f(t,x)|,$$
$$\varepsilon := \min\left\{\delta, \frac{\eta}{M}, \frac{1}{2L}\right\}, \qquad J := [t_0-\varepsilon, t_0+\varepsilon],$$

where $L$ denotes the Lipschitz constant for $T$ in $R$ (since $R$ is compact, $f$ is uniformly Lipschitz continuous on $R$).

**Definition of T**: For $v \in BC(J,\mathbb{R}^n)$, define

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

and consider

$$X := \{v \in BC(J,\mathbb{R}^n) \colon v(t_0) = x_0,\ \sup_{t\in J} |x_0 - v(t)| \le \eta\},$$

which is a closed subset of $BC(J,\mathbb{R}^n)$. Note that $BC(J,\mathbb{R}^n)$ is a complete metric space with respect to the metric

$$d(v_1,v_2) := \max_{t\in J} |v_1(t) - v_2(t)|,$$

so that $(X,d)$—by virtue of being a closed metric subspace of a complete metric space—is a complete metric space itself.

**T maps X into X**: If $v \in X$, then $T(v)(t_0) = x_0$ and

$$\begin{aligned}|x_0 - T(v)(t)| &= \left|\int_{t_0}^{t} f(s,v(s))\,ds\right| \\ &\le |t-t_0| \max_{t\in J} |f(t,v(t))| \overset{v(t)\in\overline{B_\eta(x_0)}}{\le} \varepsilon M \le \eta,\end{aligned}$$

by the definitions of $R, M, \varepsilon$ and $J$.

**T is a contraction on X**: Let $v_1, v_2 \in X$. Then

$$\begin{aligned}|T(v_1)(t) - T(v_2)(t)| &= \left|\int_{t_0}^{t} \big(f(s,v_1(s)) - f(s,v_2(s))\big)\,ds\right| \\ &\le \varepsilon \max_{|s-t_0|\le|t-t_0|} \big|f(s,v_1(s)) - f(s,v_2(s))\big| \\ &\le \varepsilon L \max_{|s-t_0|\le|t-t_0|} |v_1(s) - v_2(s)| \\ &\le \tfrac{1}{2} \max_{s\in J} |v_1(s) - v_2(s)|,\end{aligned}$$

by the definition $\varepsilon$. Now, taking the maximum over all $t \in J$ yields

$$d(T(v_1), T(v_2)) \leq \frac{1}{2} d(v_1, v_2).$$

Thus, according to the Banach fixed-point theorem, there exists a unique solution $x \in BC(J, \overline{B_\eta(x_0)})$.

$\wp$ **Picard iteration**

Under the assumptions of the Picard-Lindelöf theorem, the sequence given by

$$x_0 = x(t_0), \qquad x_n = Tx_{n-1}, \quad n \in \mathbb{N}; \qquad (Tx)(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds,$$

converges uniformly and exponentially fast to the unique solution $x$ on $J = [t_0 - \varepsilon, t_0 + \varepsilon]$:

$$\|x_n - x\|_{BC(J,\mathbb{R}^n)} \leq \frac{\lambda^n}{1-\lambda} \|x_1 - x_0\|_{BC(J,\mathbb{R}^n)},$$

where $\lambda = \varepsilon L$ is the contraction constant used in the proof of the Picard–Lindelöf theorem.[1]

**Proof**

According to the proof of the Banach fixed-point theorem, if $m \geq n$ one has

$$d(x_n, x_m) \leq \frac{\lambda^n}{1-\lambda} d(x_1, x_0),$$

where $\lambda \in (0,1)$ is the contraction constant. We apply this to the operator $T$, the metric $d$, and the constants $\varepsilon$ and $L$ as defined in the proof of the Picard–Lindelöf theorem. Since $\lim_{m\to\infty} x_m = x$ and $d(x_n, \cdot) = \|x_n - \cdot\|_{BC(J,\mathbb{R}^n)}$ is continuous, the proposition follows.

**Ex.**

The first Picard iteration for the initial-value problem

$$\dot{x} = \sqrt{x} + x^3, \qquad x(1) = 2,$$

is given by

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

The second is

$$x_2(t) = 2 + \int_1^t \big(\sqrt{x_1(s)} + (x_1(s))^3\big)\, ds.$$

(This indicates that Picard iteration, in spite of its simplicity and fast convergence, is better suited as a theoretical and computer-aided tool, than as a way to solve ODE's by hand.)

[1] It is possible to refine both the estimate for $\lambda$ and the size of the interval in different ways, but we will not pursue this here.

## 3.2 Spectral theory

Let $A \in M_{n\times n}(\mathbb{C})$ be the realization of a bounded linear transformation $\mathbb{C}^n \to \mathbb{C}^n$ (standard basis assumed), and let $|\cdot|$ denote the standard unitary norm on $\mathbb{C}^n$,

$$|(z_1,\ldots,z_n)| = \big(\sum_{j=1}^{n} |z_j|^2\big)^{1/2}; \qquad |z_j|^2 = |x_j + iy_j|^2 = |x_j|^2 + |y_j|^2.$$

In this section, most entities considered will be complex. You can think of $A$ as real, but, if so, still describing a bounded linear map $\mathbb{C}^n \to \mathbb{C}^n$.

### Existence theory for constant-coefficient linear ODE's

Consider

$$\dot{u} = Au, \qquad u(0) = u_0 \in \mathbb{C}^n. \qquad (1)$$

(The choice $t_0 = 0$ is irrelevant, since $u(\cdot - t_0)$ is a solution exactly if $u$ is.) Note that the right-hand side $f(u) = Au$ is uniformly Lipschitz with

$$|Au - Av| \le \|A\||u - v|, \qquad \|A\| = \sup_{|u|=1} |Au|,$$

so that this problem is locally and uniquely solvable. As we shall see, any solution can be globally continued on $\mathbb{R}$, and even explicitly constructed.

**The spectrum of an operator**

Let $T \in B(X)$ be a bounded linear transformation $X \to X$ (for example, $T\colon \mathbb{C}^n \to \mathbb{C}^n$ given by $A$).

- $\lambda \in \mathbb{C}$ is called an **eigenvalue** of $T$ if there exists a nonzero $v \in X$ such that

  $$Tv = \lambda v.$$

  The vector $v$ is called an **eigenvector** corresponding to the eigenvalue $\lambda$.
- The set of values $\lambda \in \mathbb{C}$ for which $(T - \lambda I)$ is invertible with a bounded inverse $(T-\lambda I)^{-1} \in B(X)$ is called the **resolvent set** of $T$. Its complement in $\mathbb{C}$, denoted $\sigma(T)$, is called the **spectrum** of $T$.

**For matrices, the spectrum consists only of eigenvalues**

For $A \in M_{n\times n}(\mathbb{C})$,

$$\sigma(A) = \{\lambda \in \mathbb{C}\colon \det(A - \lambda I) = 0\}$$

consists of the roots $(\lambda_1,\ldots,\lambda_n)$ of the **characteristic polynomial** $p_A(\lambda) \overset{\text{def.}}{=} \det(A - \lambda I)$; these are identical with the eigenvalues of $A$.

**N.b.** Defining properties of the determinant are not treated in this course; the determinant of a square matrix is the product of the final diagonal pivots in its reduced row echelon form (at the end of the Gauss–Jordan elimination).

**Proof**

Since $\mathbb{C}^n$ is finite-dimensional, we have that

$$\begin{aligned} \exists v \neq 0;\quad (A-\lambda)v = 0 \quad &\Longleftrightarrow \quad \ker(A - \lambda I) \text{ nontrivial} \\ &\Longleftrightarrow \quad (A - \lambda I) \text{ not invertible} \quad \Longleftrightarrow \quad \det(A - \lambda I) = 0, \end{aligned}$$

where the last equivalence follows, either from linear algebra, or from considering the reduced row echelon form of the matrix $A - \lambda I$. The proposition then follows by noting that $\det(A - \lambda I)$

is a polynomial in $\lambda$ of degree $n$ (which, according to the fundamental theorem of algebra, has $n$ roots).

**Multiplicity**

- The multiplicity of a root $\lambda$ of $p_A(\lambda)$ is the **algebraic multiplicity** of the the eigenvalue $\lambda$, denoted $\operatorname{mult}(\lambda)$.
- The eigenvectors corresponding to an eigenvalue $\lambda$ span a subspace of $\mathbb{C}^n$,

  $$\ker(A - \lambda I),$$

  called the **eigenspace** of $\lambda$. The dimension of this space is the **geometric multiplicity** of $\lambda$.
- An eigenvalue $\lambda$ is called **simple** if $\lambda$ is simple as a root of $p_A(\lambda)$,

  $$\lambda \text{ simple} \quad \overset{\text{def}}{\iff} \quad \operatorname{mult}(\lambda) = 1;$$

  it is **semi-simple** if the geometric and algebraic multiplicity coincide,

  $$\lambda \text{ semi-simple} \quad \overset{\text{def}}{\iff} \quad \operatorname{mult}(\lambda) = \dim\ker(A - \lambda I).$$

As we will see, if all eigenvalues of $A$ are semi-simple, then $A$ can be diagonalized.

$\wp$ **Characterization of solution spaces**

The solution set of $\dot{u} = Au$ is a vector space isomorphic to $\mathbb{C}^n$. If $A$ is real and only real initial data $u_0 \in \mathbb{R}^n$ is considered, then the solution space is isomorphic to $\mathbb{R}^n$.

**Proof**

By the Picard–Lindelöf theorem, the solution map $u_0 \overset{\varphi}{\mapsto} u(\cdot; u_0)$ is well defined.

**Injectivity**. If $\varphi(u_0) = \varphi(v_0)$ are two identical solutions, then clearly $\varphi(u_0)(0) = \varphi(v_0)(0)$, meaning $u_0 = v_0$.

**Surjectivity**. The map $\varphi$ is surjective onto the set of solutions: any solution $v$ of $\dot{v} = Av$ gives rise to initial data $v_0 := v(0)$, which in turn generates a solution $u(\cdot; v_0)$. By uniqueness $v = u = \varphi(v_0)$, so that $v \in \operatorname{ran}(\varphi)$.

**Linearity**. $\varphi$ is linear: if $v$ solves (1) with $v(0) = v_0$, and $w$ solves (1) with $w(0) = w_0$, then $u = \lambda v + \mu w$ solves (1) with $u(0) = \lambda v_0 + \mu w_0$.

**Conclusion**: Thus $\varphi \colon \mathbb{C}^n \to \varphi(\mathbb{C}^n)$ is a vector space isomorphism onto its image, which consequently is a complex vector space of dimension $n$. Since we have shown that the image $\varphi(\mathbb{C}^n)$ consists of all solutions of $\dot{u} = Au$, the proposition follows.

**Fundamental matrix**

- A basis $\{u_j\}_{j=1}^n$ of solutions is called a **fundamental system** for $\dot{u} = Au$; the corresponding matrix $(u_j)_j$ is a **fundamental matrix**.

**N.b.** According to the above characterization, a set of solutions $\{u_j\}_j$ is a fundamental system exactly if $\{u_j(0)\}_j$ is a basis for $\mathbb{C}^n$ (or $\mathbb{R}^n$ if we are considering real solutions).

**The exponential map for square matrices**

- The map

  $$\exp(A) \overset{\text{def.}}{=} \sum_{j=0}^{\infty} \frac{A^j}{j!} = I + A + \frac{A^2}{2} + \frac{A^3}{3!} + \ldots, \qquad A \in L(\mathbb{C}^n),$$

  also written $e^A$, is called the **exponential** of $A$.

**The exponential map is well defined**

Let $A \in L(\mathbb{C}^n)$. Then $\exp(A) \in L(\mathbb{C}^n)$ (in particular, $\exp(A)$ is a matrix).

**Proof**

Recall that $L(\mathbb{C}^n) = B(\mathbb{C}^n)$ as linear spaces, and that $B(\mathbb{C}^n)$ is a Banach space with the operator norm as norm. The statement $\exp(A) \in L(\mathbb{C}^n)$ is thus equivalent with that

$$\lim_{N\to\infty} \underbrace{\sum_{j=0}^{N} \frac{A^j}{j!}}_{=:y_N}$$

is well-defined as a limit in $B(\mathbb{C}^n)$. For $N \geq m$ we have

$$\|y_N - y_m\| \leq \sum_{j=m+1}^{N} \frac{\|A^j\|}{j!} \leq \sum_{j=m+1}^{\infty} \frac{\|A\|^j}{j!} \to 0,$$

as $N \geq m \to \infty$. Hence $\{y_N\}_N$ is Cauchy and converges in $B(\mathbb{C}^n)$. The same argument without $y_m$ shows that

$$\|\exp(A)\| \leq e^{\|A\|}.$$

**Properties of the exponential map**

- If $AB = BA$, then

$$B\,\exp(A) = \exp(A)\,B \qquad \text{and} \qquad \exp(A+B) = \exp(A)\exp(B).$$

- If $T \in M_{n\times n}(\mathbb{C}^n)$ is invertible, then

$$T\exp(A)T^{-1} = \exp(TAT^{-1}).$$

- $\exp(A)$ is invertible with

$$\big(\exp(A)\big)^{-1} = \exp(-A).$$

- $[t \mapsto \exp(tA)]$ is continuously differentiable with

$$\frac{d}{dt}\exp(tA) = A\exp(tA).$$

**Solution formula**

The unique solution of (1) is

$$u(t; u_0) = \exp(tA)u_0,$$

and $\exp(tA)$ is a fundamental matrix with $\exp(tA)|_{t=0} = I$.

**Proof**

Since

$$\frac{d}{dt}\exp(tA) = A\exp(tA),$$

$\exp(tA)$ solves the matrix equation $\dot{X} = AX$. This means that each column of $\exp(tA)$ solves $\dot{u} = Au$. Since $\exp(tA)$ is invertible, the columns are linearly independent, so they must span the solution space (it is $n$-dimensional, as we have proved). Thus $\exp(tA)$ is a fundamental matrix. That $\exp(\mathbf{0}) = \exp(tA)|_{t=0} = I$ follows immediately from the definition of the exponential map.

Now, multiplying $\exp(tA)$ from the right with $u_0$ yields the solution of the initial-value problem (1), since

$$\frac{d}{dt}\exp(tA)u_0 = A\exp(tA)u_0,$$

and

$$u(0) = \exp(tA)|_{t=0}u_0 = Iu_0 = u_0.$$

## Spectral decompositions

If $A$ is nilpotent, i.e., if

$$A^{n_0} = 0 \quad \text{for some } n_0 \in \mathbb{N},$$

then $\exp(A)$—and therefore $\exp(tA)$—is a finite sum:

$$\exp(A) = \sum_{j=0}^{\infty}\frac{A^j}{j!} = \sum_{j=0}^{n_0-1}\frac{A^j}{j!} = I + A + \ldots + \frac{A^{n_0-1}}{(n_0-1)!}.$$

In general, other methods must be employed.

### Cayley–Hamilton

A matrix satisfies its characteristic polynomial: $p_A(A) = 0$.

**N.b.** Since $p_A$ is a polynomial of degree $n$, this implies that $A^n$ can be replaced with a polynomial of degree at most $n-1$. Hence $\exp(A)$ can be reduced to a polynomial in $A$ of degree at most $n-1$. This is the basis for the spectral decomposition below.

### Algebraic description of the solution space

Let $\lambda \in \mathbb{C}$ denote an eigenvalue of $A$.

- A vector $v \neq 0$ is called a **generalized eigenvector** if $(A-\lambda I)^k v = 0$ for some $k \in \mathbb{N}$; we call

$$N_k := \ker\left((A-\lambda I)^k\right)$$

a **generalized eigenspace** corresponding to the eigenvalue $\lambda$.

### Each eigenvalue has a maximal generalized eigenspace

There exists a minimal integer, $k_\lambda \in \mathbb{N}$, such that $N_k = N_{k_\lambda}$ for all $k \geq k_\lambda$.

**Proof**

Since

$$(A-\lambda)v = 0 \quad \Longrightarrow \quad (A-\lambda)^2 v = 0,$$

we have

$$\{0\} \subset N_k \subset N_{k+1} \subset \mathbb{C}^n, \qquad k \in \mathbb{N},$$

and since $N_k$ are linear spaces, all contained in the $n$-dimensional space $\mathbb{C}^n$, this chain must end:

$$\exists \text{ minimal } k_\lambda \geq 1; \qquad N_k = N_{k_\lambda} \quad \text{for all } k \geq k_\lambda.$$

Let

$$R_k := \operatorname{ran}((A-\lambda I)^k), \qquad k \in \mathbb{N}.$$

The **Riesz index** $m_\lambda$ is the minimal natural number that ends the chain $\{R_k\}_k$:

$$m_\lambda \overset{\text{def.}}{=} \min\{k \in \mathbb{N} \colon R_k = R_{k+1}\}.$$

The vector spaces $R_k$ satisfy

$$\{0\} \subset R_{k+1} \subset R_k \subset \mathbb{C}^n, \qquad \text{with} \qquad R^k = R^{m_\lambda} \quad \text{for all } k \geq m_\lambda.$$

**Riesz decomposition**
We have $k_\lambda = m_\lambda$, and

$$\mathbb{C}^n = N(\lambda) \oplus R(\lambda) := \ker\left((A-\lambda I)^{m_\lambda}\right) \oplus \operatorname{ran}\left((A-\lambda I)^{m_\lambda}\right).$$

**Spectral decomposition**
$\mathbb{C}^n$ can be decomposed into maximal generalized eigenspaces $N(\lambda_k)$ with $\dim(N(\lambda_k)) = \operatorname{mult}(\lambda_k)$:

$$\mathbb{C}^n = \oplus_{k=1}^m N(\lambda_k).$$

Here $m$ is the number of different eigenvalues (not counted with multiplicity).

**Sketch of steps**

Let $n_k := \operatorname{mult}(\lambda_k)$. Using Cayley–Hamilton one can show that

$$\mathbb{C}^n = \oplus_{k=1}^m \ker((A-\lambda_k I)^{n_k}),$$

and then furthermore (using the Riesz decomposition) that

$$n_k = \dim(N(\lambda_k)) \quad \text{and} \quad n_k \geq m_{\lambda_k},$$

so that

$$\ker((\lambda_k I - A)^{n_k}) = N(\lambda_k).$$

The fact that the algebraic multiplicity is the dimension of the maximal generalized eigenspace implies that $\lambda$ is sempi-simple exactly if $m_\lambda = 1$.

**The matrix form of the spectral decomposition**
According to the above, $\mathbb{C}^n = \oplus_{k=1}^m N(\lambda_k)$ has a basis of generalized eigenvectors. Let

$$A_k := A|_{N(\lambda_k)}, \qquad I_k := I|_{N(\lambda_k)}, \qquad \tilde{N}_k := A_k - \lambda_k I_k, \quad k = 1, \dots, m,$$

be the restrictions of the mappings $A$, $I$ and $A - \lambda_k I$ onto the eigenspaces $N(\lambda_k)$ (meaning that they act only on the basis vectors of the corresponding eigenspaces). Then $\tilde{N}_k$ is nilpotent, since

$$\tilde{N}_k^{m_{\lambda_k}} = 0$$

on the generalized eigenspace $N(\lambda_k)$ (this is the definition of $m_{\lambda_k}$). *In our basis of generalized eigenvectors* $A$ takes the form

$$\begin{bmatrix} [A_1] & 0 & 0 & \dots & 0 \\ 0 & [A_2] & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & \dots & 0 & [A_m] \end{bmatrix}_{n\times n} = \begin{bmatrix} [\lambda_1 I_1 + \tilde{N}_1] & 0 & 0 & \dots & 0 \\ 0 & [\lambda_2 I_2 + \tilde{N}_2] & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & & 0 & [\lambda_m I_m + \tilde{N}_m] \end{bmatrix}_{n\times n}.$$

Because $\tilde{N}_k I_k = I_k \tilde{N}_k$, $\exp(t\lambda_k I_k) = e^{t\lambda} I_k$, and $\tilde{N}_k^{m_{\lambda_k}} = 0$ one has

$$\exp(tA_k) = \exp\big(t(\lambda_k I_k + \tilde{N}_k)\big) = \exp\big(t\lambda_k I_k\big)\exp\Big(t\tilde{N}_k\Big) = e^{t\lambda_k}\big(I_k + t\tilde{N}_k + \ldots + \frac{(t\tilde{N}_k)^{m_{\lambda_k}-1}}{(m_{\lambda_k}-1)!}\big),$$

and then

$$\exp(t \underbrace{T[A_k]_k T^{-1}}_{tA \text{ in original basis}}) = T\exp(t[A_k]_k)T^{-1} \qquad (T \text{ change-of-basis matrix}).$$

One only needs to find suitable bases for $N(\lambda_k)$, $k = 1, \ldots, m$.

**Ex.**

The matrix

$$A := \begin{bmatrix} 0 & -8 & 4 \\ 0 & 2 & 0 \\ 2 & 3 & -2 \end{bmatrix}$$

has eigenvalues $\lambda_{1,2} = 2$ and $\lambda_3 = -4$. Its generalized eigenvectors solve

$$(A - 2I)^2 v = \begin{bmatrix} 12 & 28 & -24 \\ 0 & 0 & 0 \\ -12 & -28 & 24 \end{bmatrix} v = 0 \quad\Leftrightarrow\quad v = s\begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} + t\begin{bmatrix} 0 \\ 6 \\ 7 \end{bmatrix} \qquad s, t \in \mathbb{C},$$

and

$$(A + 4I)\,v = 0 \quad\Leftrightarrow\quad v = s\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \qquad s \in \mathbb{C}.$$

Let

$$T := \begin{bmatrix} 2 & 0 & -1 \\ 0 & 6 & 0 \\ 1 & 7 & 1 \end{bmatrix} \quad \text{so that} \quad T^{-1} = \frac{1}{18}\begin{bmatrix} 6 & -7 & 6 \\ 0 & 3 & 0 \\ -6 & -14 & 12 \end{bmatrix}, \quad T^{-1}AT = \begin{bmatrix} 2 & -10 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -4 \end{bmatrix}.$$

In the basis given by $T$ we have

$$I_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \tilde{N}_1 = \begin{bmatrix} 0 & -10 \\ 0 & 0 \end{bmatrix} \quad \text{with} \quad A_1 = 2I_1 + \tilde{N}_1,$$

$$\exp(tA_1) = \exp(2tI_1)\exp(t\tilde{N}_1) = e^{2t}(I_1 + t\tilde{N}_1) = e^{2t}\begin{bmatrix} 1 & -10t \\ 0 & 1 \end{bmatrix},$$

and

$$\exp(tT^{-1}AT) = \begin{bmatrix} e^{2t} & -10te^{2t} & 0 \\ 0 & e^{2t} & 0 \\ 0 & 0 & e^{-4t} \end{bmatrix}.$$

Expressed in the original basis,

$$\exp(tA) = T\exp(tT^{-1}AT)T^{-1} = \frac{1}{9}e^{2t}\begin{bmatrix} 6 & -7 & 6 \\ 0 & 9 & 0 \\ 3 & 7 & 3 \end{bmatrix} - \frac{1}{3}te^{2t}\begin{bmatrix} 0 & 10 & 0 \\ 0 & 0 & 0 \\ 0 & 5 & 0 \end{bmatrix} + \frac{1}{9}e^{-4t}\begin{bmatrix} 3 & 7 & -6 \\ 0 & 0 & 0 \\ -3 & -7 & 6 \end{bmatrix}.$$

## Applications: the Jordan normal form and finite-dimensional spectral theorem

### The Jordan normal form

The Jordan normal form corresponds to a spectral decomposition in which the bases for $N(\lambda_k)$ are chosen such that the nilpotent matrices $\tilde{N}_k$ have the special form

$$\tilde{N}_{\lambda_k} = \begin{bmatrix} 0 & j_1 & 0 & \dots & 0 \\ 0 & 0 & j_2 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & j_{n_k-1} \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}, \qquad j_l \in \{0,1\}, \quad l = 1, \dots, n_k - 1,$$

with $n_k = \operatorname{mult}(\lambda_k)$, and

$$A_k = \lambda_k I_k + \tilde{N}_{\lambda_k} = \begin{bmatrix} \lambda_k & j_1 & 0 & \dots & 0 \\ 0 & \lambda_k & j_2 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & 0 & \dots & \lambda_k & j_{n_k-1} \\ 0 & 0 & \dots & 0 & \lambda_k \end{bmatrix}.$$

To obtain this, given an eigenvalue $\lambda$, pick a generalized eigenvector

$$v_{m_\lambda} \in \ker(A - \lambda I)^{m_\lambda}, \quad v_{m_\lambda} \notin \ker(A - \lambda I)^{m_\lambda - 1}$$

and set

$$v_{m_\lambda - 1} := (A - \lambda I) v_{m_\lambda}, \quad \dots \quad , v_1 := (A - \lambda I)^{m_\lambda - 1} v_{m_\lambda},$$

so that

$$v_j \in \ker\big((A - \lambda I)^j\big), \quad v_j \notin \ker\big((A - \lambda I)^{j-1}\big), \qquad j = 1, \dots, m_\lambda.$$

The **Jordan chain** $\{v_1, \dots, v_{m_\lambda}\}$ is a basis for a subspace of $N(\lambda)$, on which

$$\tilde{N} v_j = (A - \lambda I) v_j = v_{j-1}, \qquad j = 1, \dots, m_\lambda,$$

if we let $v_0 := 0$. Hence, the $j$:th column of $\tilde{N}$ is $v_{j-1}$. This gives the nilpotent part of a so-called **Jordan block** (with ones above the diagonal, all other elements zero). If $m_\lambda < n_k$ additional Jordan chains need to be added. Each chain gives rise to a Jordan block; adding the different chains into a basis for $N(\lambda)$ gives the form of $\tilde{N}$ above.

**Ex. (continued from above)**

The eigenvalues of

$$A := \begin{bmatrix} 0 & -8 & 4 \\ 0 & 2 & 0 \\ 2 & 3 & -2 \end{bmatrix}$$

are $\lambda_1 = 2$ (double) and $\lambda_2 = -4$ (simple).

Since $\operatorname{mult}(2) = 2$, we have $k_2 = m_2 \leq 2$, so we can start the Jordan chain by looking for a vector in $N_2$ (which equals the maximal generalized eigenspace $N(2)$). Candidates are (cf. above):

$$u = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad w = \begin{bmatrix} 0 \\ 6 \\ 7 \end{bmatrix}.$$

Since $(A-2I)u=0$ we have $u \in N_1$, whereas

$$(A-2I)w = \begin{bmatrix} -20 \\ 0 \\ -10 \end{bmatrix} = -10u$$

implies that

$$v_2 := w \in N_2 \setminus N_1 \qquad \text{whereas} \qquad v_1 := (A-2I)w = -10u \in N_1.$$

The Jordan block corresponding to the simple eigenvalue $-4$ consists of just the eigenvalue itself, and the eigenvector spanning the one-dimensional eigenspace $N(-4)$ is $\tilde{v}_1 := (-1,0,1)$, as calculated above.

The change-of-basis matrix is thus given by

$$T := [v_1 \; v_2 \; \tilde{v}_1] = \begin{bmatrix} -20 & 0 & -1 \\ 0 & 6 & 0 \\ -10 & 7 & 1 \end{bmatrix},$$

in which the linear transformation expressed by $A$ in the original basis takes the Jordan normal form[1]

$$\begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -4 \end{bmatrix}.$$

### The spectral theorem for Hermitian matrices

- For $A \in M_{n\times n}(\mathbb{C})$ the matrix $A^*$ defined by $a^*_{ij} := \overline{a_{ji}}$ is called its **adjoint** or **conjugate transpose**. Equivalently,

  $$A^* = \overline{A^t},$$

  where $A^t$ is the transpose of $A$.
- $A$ is said to be **Hermitian** (or **self-adjoint**) if $A = A^*$.

The **spectral theorem** says that any Hermitian matrix admits a basis of eigenvectors in which $A$ can be diagonalized, and that this basis can be chosen to be **orthonormal**, meaning that the basis vectors are of unit length and perpendicular to each other.[2]

**N.b.** *If a matrix can be diagonalized, this can always be achieved using the spectral (or Jordan) decomposition* (though the corresponding basis need not be orthonormal).

[1] As can be seen, the spectral decomposition above brought as very close to the Jordan normal form, which will typically happen if the Jordan chains are few or short (low algebraic multiplicity).

[2] We will come back to this later.

# Chapter 4

# Hilbert space theory

## 4.1 Inner-product spaces

Let $X$ be a vector space over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$.

### Inner-product spaces

An **inner product** $\langle \cdot, \cdot \rangle$ on $X$ is a map $X \times X \to \mathbb{K}$, $(x, y) \mapsto \langle x, y \rangle$, that is **conjugate symmetric**

$$\langle x, y \rangle = \overline{\langle y, x \rangle},$$

**linear in its first argument**,

$$\begin{aligned} \langle \lambda x, y \rangle &= \lambda \langle x, y \rangle, \\ \langle x + y, z \rangle &= \langle x, z \rangle + \langle y, z \rangle, \end{aligned}$$

and **non-degenerate (positive definite)**,

$$\langle x, x \rangle > 0 \quad \text{for} \quad x \neq 0,$$

with $x, y, z \in X$ and $\lambda \in \mathbb{K}$ arbitrary. The pair $(X, \langle \cdot, \cdot \rangle)$ is called an **inner-product space**.

**Ex.**

- The canonical inner product is the dot product in $\mathbb{R}^n$:

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

- For matrices in $M_{n \times n}(\mathbb{R})$ one can define a dot product by setting

$$\langle A, B \rangle := \mathrm{tr}(B^t A),$$

where $\mathrm{tr}(C) = \sum_{j=1}^{n} c_{jj}$ is the trace of a matrix $C$, and $B^t$ is the transpose of $B$. Then

$$B^t A = \sum_{j=1}^{n} b_{ij}^t a_{jk} = \sum_{j=1}^{n} b_{ji} a_{jk},$$

and

$$\operatorname{tr}(B^t A) = \sum_{k=1}^{n}\sum_{j=1}^{n} b_{jk}a_{jk} = \sum_{1\le j,k\le n} a_{jk}b_{jk}$$

coincides with the dot product on $\mathbb{R}^{nn} \cong M_{n\times n}(\mathbb{R})$.

**Properties of the inner product**
An inner product satisfies

$$\begin{array}{ll}
\text{(i)} & \langle x, y+z\rangle = \langle x,y\rangle + \langle x,z\rangle, \\
\text{(ii)} & \langle x, \lambda y\rangle = \bar{\lambda}\langle x,y\rangle, \\
\text{(iii)} & \langle x, 0\rangle = \langle 0, x\rangle = 0, \\
\text{(iv)} & \text{If } \langle x, z\rangle = 0 \text{ for all } z\in X \quad \text{then} \quad x = 0.
\end{array}$$

**N.b.** By linearity, the last property implies that if $\langle x,z\rangle = \langle y,z\rangle$ for all $z \in X$, then $x = y$.

**Proof**

**(i)**

$$\langle x, y+z\rangle = \overline{\langle y+z, x\rangle} = \overline{\langle y,x\rangle + \langle z,x\rangle} = \overline{\langle y,x\rangle} + \overline{\langle z,x\rangle} = \langle x,y\rangle + \langle x,z\rangle.$$

**(ii)**

$$\langle x, \lambda y\rangle = \overline{\langle \lambda y, x\rangle} = \overline{\lambda\langle y,x\rangle} = \bar{\lambda}\langle x,y\rangle.$$

**(iii)**

$$\langle \mathbf{0}, x\rangle = \langle 0x, x\rangle = 0\langle x,x\rangle = 0,$$

and

$$\langle x, 0\rangle = \overline{\langle 0, x\rangle} = 0.$$

**(iv)**

$$\langle x,z\rangle = 0 \text{ for all } z\in X \quad\Longrightarrow\quad \langle x,x\rangle = 0 \quad\Longrightarrow\quad x = 0.$$

**Inner-product spaces as normed spaces**
An inner-product space $(X, \langle\cdot,\cdot\rangle)$ carries a natural norm given by $\|x\| := \langle x,x\rangle^{1/2}$. To prove this, we need:

**The Cauchy–Schwarz inequality**
For all $x, y \in (X, \langle\cdot,\cdot\rangle)$,

$$|\langle x,y\rangle| \le \|x\|\|y\|,$$

with equality if and only if $x$ and $y$ are linearly dependent.

**Proof**

**Linearly dependent case**: Without loss of generality, assume that $x = \lambda y$ (if $y = \lambda x$ we can always relabel the vectors). Then

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

**Linearly independent case**: If $x - \lambda y \neq 0$ and $y - \lambda x \neq 0$ for all $\lambda \in \mathbb{K}$, then also $x, y \neq 0$, and

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

If $\langle x, y\rangle = 0$ the Cauchy–Schwarz inequality is trivial, so assume that $\langle x, y\rangle \neq 0$. Let $\lambda := tu$ with $u := \frac{\langle x,y\rangle}{|\langle x,y\rangle|}$, so that

$$\bar{\lambda}\langle x, y\rangle = t\frac{\overline{\langle x, y\rangle}\langle x, y\rangle}{|\langle x, y\rangle|} = t|\langle x, y\rangle| \qquad \text{and} \qquad |\lambda|^2 = t^2.$$

Hence,

$$0 < \|x\|^2 + 2t|\langle x, y\rangle| + t^2\|y\|^2 = \Big(\|y\|t + \frac{|\langle x, y\rangle|}{\|y\|}\Big)^2 + \|x\|^2 - \Big(\frac{|\langle x, y\rangle|}{\|y\|}\Big)^2.$$

By choosing $t = -|\langle x, y\rangle|/\|y\|^2$, we obtain that

$$\frac{|\langle x, y\rangle|^2}{\|y\|^2} < \|x\|^2,$$

which proves the assertion.

**Inner-product spaces are normed**

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**

**Positive homogeneity**:

$$\|\lambda x\| = \langle \lambda x, \lambda x\rangle^{1/2} = \big(\lambda\bar{\lambda}\langle x, x\rangle\big)^{1/2} = \big(|\lambda|^2\|x\|^2\big)^{1/2} = |\lambda|\|x\|.$$

**Triangle inequality**: By the Cauchy–Schwarz inequality,

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

**Non-degeneracy**:

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

according to the positive definiteness of the inner product.

**Parallelogram law and polarization identity**
Let $(X, \|\cdot\|)$ be a normed space. Then the **parallelogram law**

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

holds exactly if $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$ can be defined using an inner product on $X$. If so,

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

if $X$ is real, and

$$\langle x, y\rangle = \frac{1}{4}\sum_{k=0}^{3} i^k \|x + i^k y\|^2,$$

if $X$ is complex.

**Proof**

We only show that the parallelogram law and polarization identity hold in an inner product space; the other direction (starting with a norm and the parallelogram identity to define an inner product) is left as an exercise.

**Parallelogram law**: If $X$ is an inner-product space, then

$$\|x \pm y\|^2 = \|x\|^2 \pm 2\Re\langle x, y\rangle + \|y\|^2;$$

the parallelogram law follows from adding these two equations to each other.

**Polarization identity**: When $X$ is a real inner-product space, it follows directly that

$$\|x+y\|^2 - \|x-y\|^2 = \big(\|x\|^2 + 2\langle x,y\rangle + \|y\|^2\big) - \big(\|x\|^2 - 2\langle x,y\rangle + \|y\|^2\big) = 4\langle x, y\rangle.$$

If $X$ is complex, the corresponding calculcation yields that

$$\begin{aligned}\sum_{k=0}^{3} i^k \|x + i^k y\|^2 &= \sum_{k=0}^{3} i^k \big(\|x\|^2 + 2\Re\langle x, i^k y\rangle + \|i^k y\|^2\big)\\ &= \big(\|x\|^2 + 2\Re\langle x,y\rangle + \|y\|^2\big) - \big(\|x\|^2 - 2\Re\langle x,y\rangle + \|y\|^2\big)\\ &+ i\big(\|x\|^2 - 2\Re i\langle x,y\rangle + \|y\|^2\big) - i\big(\|x\|^2 + 2\Re i\langle x,y\rangle + \|y\|^2\big).\end{aligned}$$

Since $\Re iz = -\Im z$ for any $z \in \mathbb{C}$, we obtain

$$\sum_{k=0}^{3} i^k \|x + i^k y\|^2 = 4\Re\langle x,y\rangle + 4\Im\langle x,y\rangle = 4\langle x, y\rangle.$$

**Ex.**

- **Pythagoras' theorem**: If $\langle x, y\rangle = 0$ in an inner-product space, then

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

which, in $\mathbb{R}^2$, we recognize as

$$a^2 + b^2 = c^2,$$

with $a, b, c$ the sides of a right-angled triangle.

- If we define $\langle x, y\rangle := \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$ in $\mathbb{R}^2$ using the polarization identity , we see that

$$\begin{aligned}\langle x, y\rangle &= \frac{1}{4}\left((x_1+y_1)^2 + (x_2+y_2)^2\right) - \frac{1}{4}\left((x_1-y_1)^2 + (x_2-y_2)^2\right)\\ &= \frac{1}{4}\left(x_1^2 + 2x_1y_1 + y_1^2 + x_2^2 + 2x_2y_2 + y_2^2\right) - \frac{1}{4}\left(x_1^2 - 2x_1y_1 + y_1^2 + x_2^2 - 2x_2y_2 + y_2^2\right)\\ &= x_1y_1 + x_2y_2\end{aligned}$$

is the standard dot product.

## Hilbert spaces

- A complete inner-product space is called a **Hilbert space**. Similarly, inner-product spaces are sometimes called **pre-Hilbert spaces**.

**Ex.**

- The Banach spaces $\mathbb{R}^n$, $l_2(\mathbb{R})$ and $L_2(I, \mathbb{R})$, as well as their complex counterparts $\mathbb{C}^n$, $l_2(\mathbb{C})$ and $L_2(I, \mathbb{C})$, all have norms that come from inner products:

$$\langle x, y\rangle_{\mathbb{C}^n} = \sum_{j=1}^{n} x_j \bar{y}_j \quad \text{in} \quad \mathbb{C}^n,$$

$$\langle x, y\rangle_{l_2} = \sum_{j=1}^{\infty} x_j \bar{y}_j \quad \text{in} \quad l_2,$$

and

$$\langle x, y\rangle_{L_2} = \int_I x(s)\overline{y(s)}\, ds \quad \text{in} \quad L_2.$$

(If the spaces are real, there are no complex conjugates.) Thus, they are all Hilbert spaces. In particular, this proves the $l_2$- and $L_2$-norms defined earlier in this course are indeed norms.

- The space of real-valued bounded continuous functions on a finite open interval, $BC((a, b), \mathbb{R})$, can be equipped with the $L_2$-inner product. This is a pre-Hilbert space, the completion of which is $L_2((a, b), \mathbb{R})$.

### Convex sets and the closest point property

- Let $X$ be a linear space. A subset $M \subset X$ is called **convex** if

$$x, y \in M \quad \Longrightarrow \quad tx + (1-t)y \in M \quad \text{for all} \quad t \in (0, 1),$$

i.e., if all points in $M$ can be joined by line segments in $M$.

**Ex.**

- Any **hyperbox** $\{x \in \mathbb{R}^n \colon a_j \leq x_j \leq b_j\}$ is convex.
- Intuitively, any region with a 'hole', like $\mathbb{R}^n \setminus B_1$, is *not* convex.
- Linear subspaces are convex:

$$x, y \in M \quad \Longrightarrow \quad \mu x + \lambda y \in M \quad \text{for all scalars } \mu, \lambda,$$

clearly implies that $tx + (1-t)y \in M$ for all $t \in (0,1)$.

**Closest point property (Minimal distance theorem)**

Let $H$ be a Hilbert space, and $M \subset H$ a non-empty, closed and convex subset of $H$. For any $x_0 \in H$ there is a unique element $y_0 \in M$ such that

$$\|x_0 - y_0\| = \inf_{y \in M} \|x_0 - y\|.$$

**N.b.** The number $\inf_{y \in M} \|x_0 - y\|$ is the **distance from** $x_0$ **to** $M$, denoted $\operatorname{dist}(x_0, M)$.

**Proof**

**A minimizing sequence**: Since $M \neq \emptyset$, the number $d := \inf_{y \in M} \|x_0 - y\|$ is finite and non-negative, and by the definition of infimum, there exists a minimizing sequence $\{y_j\}_{j \in \mathbb{N}} \subset M$ such that

$$\lim_{j \to \infty} \|x_0 - y_j\| = d.$$

$\{y_j\}_{j \in \mathbb{N}}$ **is Cauchy**: By the parallelogram law applied to $x_0 - y_n$, $x_0 - y_m$, we have

$$\|2x_0 - (y_m + y_n)\|^2 + \|y_m - y_n\|^2 = 2\|x_0 - y_m\|^2 + 2\|x_0 - y_n\|^2 \to 4d^2, \qquad m, n \to \infty.$$

In view of that $M$ is convex and $d$ minimal, we also have that

$$\|2x_0 - (y_m + y_n)\|^2 = 4\Big\|x_0 - \frac{y_m + y_n}{2}\Big\|^2 \geq 4d^2.$$

Consequently,

$$\|y_m - y_n\|^2 \to 0 \quad \text{as} \quad m, n \to \infty.$$

Since $M \subset H$ is closed and $H$ is complete, there exists

$$y_0 = \lim_{j \to \infty} y_j \in M \quad \text{with} \quad \|x_0 - y_0\| = \lim_{j \to \infty} \|x_0 - y_j\| = d.$$

**Uniqueness**: Suppose that $z_0 \in M$ satisfies $\|x_0 - z_0\| = d$. Then $\frac{y_0 + z_0}{2} \in M$ and the parallelogram law (applied to $x_0 - y_0$, $x_0 - z_0$) yields that

$$\|y_0 - z_0\|^2 = 2\|x_0 - y_0\|^2 + 2\|x_0 - z_0\|^2 - 4\Big\|x_0 - \frac{y_0 + z_0}{2}\Big\|^2 \leq 2d^2 + 2d^2 - 4d^2 = 0,$$

so that $z_0 = y_0$.

**Ex.**

- In the Hilbert space $\mathbb{R}^2$:
  - The closed unit disk $\{x_1^2 + x_2^2 \leq 1\}$ contains a unique element that minimizes the distance to the point $(2,0)$ (namely $(1,0)$).
  - The subgraph $\{x_2 \leq x_1^2\}$ is closed but not convex; it has more than one point minimizing the distance to the point $(0,1)$.
  - The open unit ball $\{x_1^2 + x_2^2 < 1\}$ is convex but not closed; it has no element minimizing the distance to a point outside itself.
- Let

$$M_n := \operatorname{span}\{e^{ikx}\}_{k=-n}^{n}$$

be the closed linear span of trigonometric functions $1, e^{ix}, e^{-ix} \dots, e^{inx}, e^{-inx} \in L_2((-\pi,\pi),\mathbb{C})$. For any $n \in \mathbb{N}$ and any $f \in L_2((-\pi,\pi),\mathbb{C})$ there is a unique linear combination of such functions that minimizes the $L_2$-distance to $f$:

$$\int_{-\pi}^{\pi} \Big| f(x) - \sum_{k=-n}^{n} c_k e^{ikx} \Big|^2 \, dx = \min_{g \in M_n} \int_{-\pi}^{\pi} \big| f(x) - g(x) \big|^2 \, dx.$$

The coefficients $c_k$ are known as (complex) **Fourier coefficients** of the function $f$.

## 4.2 Orthogonality

Consider an inner-product space $(X, \langle \cdot, \cdot \rangle)$ over a field $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. When $X$ is complete, we shall write $H$ to indicate that it is a Hilbert space.

### The projection and Riesz representation theorems

**Orthogonal vectors and set**

- Two vectors $x, y \in X$ are said to be **orthogonal**,

$$x \perp y \quad \overset{\text{def}}{\iff} \quad \langle x, y \rangle = 0.$$

- A vector $x \in X$ is said to be **orthogonal** to a set $M \subset X$,

$$x \perp M \quad \overset{\text{def}}{\iff} \quad \langle x, y \rangle = 0 \quad \text{for all } y \in M.$$

- Two sets $M, N \subset X$ are said to be **orthogonal**,

$$M \perp N \quad \overset{\text{def}}{\iff} \quad \langle x, y \rangle = 0 \quad \text{for all} \quad x \in M, y \in N.$$

- The **orthogonal complement** of a set $M \in X$ consists of all vectors orthogonal to $M$:

$$M^{\perp} \quad \overset{\text{def.}}{=} \quad \{x \in X \colon x \perp M\}.$$

The word **perpendicular** is sometimes used interchangeably with 'orthogonal', but mostly in $\mathbb{R}^n$.

**Ex.**

- In $\mathbb{R}^3$, the vector $(1,2,1)$ is orthogonal to the plane $\{x_1 + 2x_2 + x_3 = 0\}$.
- In $L_2((-\pi,\pi),\mathbb{C})$ the vectors $e^{ikx}$, $k \in \mathbb{Z}$, are all orthogonal to each other:

$$\langle e^{ik_1x}, e^{ik_2x}\rangle = \int_{-\pi}^{\pi} e^{ik_1x}\overline{e^{ik_2x}}\,dx = \int_{-\pi}^{\pi} e^{i(k_1-k_2)x}\,dx = \frac{e^{i(k_1-k_2)x}}{i(k_1-k_2)}\Big|_{-\pi}^{\pi} = 0 \quad \text{for} \quad k_1 \neq k_2,$$

  by periodicity of $e^{ikx} = \cos(kx) + i\sin(kx)$.
- In $l_2(\mathbb{R})$, with $e_1 = (1,0,0,\ldots)$,

$$\{e_1\}^\perp = \{x \in l_2(\mathbb{R})\colon x = (0, x_2, x_3, \ldots)\}.$$

* In a Hilbert space, $H = \oplus_{j=1}^m H_j$ for subspaces $H_j \subset H$ means that

$$x = \sum_{j=1}^m x_j, \quad x_j \in H_j, \qquad H_j \perp H_k \text{ for } j \neq k,$$

which gives a unique representation of any $x \in H$ in terms of elements in orthogonal subspaces.[1]

**Ex.**

- $$\mathbb{R}^3 = \operatorname{span}\{(1,0,0)\} \oplus \operatorname{span}\{(0,1,0),(0,0,1)\},$$

  but also

$$\mathbb{R}^3 = \operatorname{span}\{(1,2,1)\} \oplus \{x_1 + 2x_2 + x_3 = 0\}.$$

### The projection theorem

Let $M \subset H$ be a closed linear subspace of a Hilbert space $H$. Then $H = M \oplus M^\perp$.

**Proof**

**Existence of $y_0 \in M$:** Pick $x_0 \in H$. By the minimal distance theorem, there exists a unique point $y_0 \in M$ with

$$\|x_0 - y_0\| = \inf_{y \in M} \|x_0 - y\|.$$

**Existence of $x_0 - y_0 \in M^\perp$:** Since $M$ is a subspace, $y_0 + \lambda y \in M$ for any $y \in M$, $\lambda \in \mathbb{K}$. Hence

$$\|x_0 - y_0\|^2 \leq \|x_0 - y_0 - \lambda y\|^2 = \|x_0 - y_0\|^2 - 2\Re(\lambda\langle y, x_0 - y_0\rangle) + |\lambda|^2\|y\|^2,$$

and

$$-2\Re(\lambda\langle y, x_0 - y_0\rangle) + |\lambda|^2\|y\|^2 \geq 0.$$

By taking $\lambda = \varepsilon \ll 1$, we see that

$$\Re(\lambda\langle y, x_0 - y_0\rangle) \leq 0,$$

[1] Note that, in general, not all direct sums describe orthogonal subspaces.

and, similarly, by taking $\lambda = -i\varepsilon$, that

$$\Im(\lambda\langle y, x_0 - y_0\rangle) \le 0.$$

Since $y \in M$ is arbitrary, by exchanging $-y$ for $y$, we obtain

$$\langle y, x_0 - y_0\rangle = 0 \quad \text{for any} \quad y \in M.$$

Thus we can write

$$x_0 = y_0 + (x_0 - y_0), \quad \text{where} \quad y_0 \in M, \quad x_0 - y_0 \in M^\perp.$$

**Uniqueness**: If we have two representations $x_0 = y_0 + z_0$ and $x_0 = \tilde{y}_0 + \tilde{z}_0$, then

$$M \ni y_0 - \tilde{y}_0 = \tilde{z}_0 - z_0 \in M^\perp,$$

but only the zero vector is orthogonal to itself, implying that $y_0 = \tilde{y}_0$ and $z_0 = \tilde{z}_0$.

**Ex.**

- The null space of a matrix $A \in M_{m\times n}(\mathbb{R})$ is closed linear subspace, so that $\mathbb{R}^n = \ker(A) \oplus (\ker(A))^\perp$. The geometric rank–nullity theorem characterizes the orthogonal complement as the range of the transpose matrix:

$$\mathbb{R}^n = \ker(A) \oplus \operatorname{ran}(A^t).$$

### Corollary: strict subspace characterization

If $M \subsetneq H$ is a closed linear subspace of $H$, there exists a non-zero vector $z_0 \in H$ with $z_0 \perp M$.

**Proof**

Since $M \neq H$ there exists $x_0 \in H \setminus M$. According to the projection theorem, $x_0 = y_0 + z_0$ with $y_0 \in M$, $z_0 \in M^\perp$. Then $z_0 \neq 0$, and $z_0 \perp M$ is the vector we are looking for.

**Ex.**

- Let $M = \overline{l_0}$ be the closure of

$$l_0 = \{x \in l_2 \colon \{x_j\}_{j\in\mathbb{N}} \text{ has finitely many non-zero entries}\}$$

in $l_2$. Is $M = l_2$? Say there exists $z \in l_2$ such that $z \perp M$. Since $\{e_j\}_{j\in\mathbb{N}} \subset M$, we have

$$\langle z, e_j\rangle = z_j = 0 \quad \text{for all} \quad j \in \mathbb{N}.$$

Thus $z = 0$, and $\overline{l_0} = l_2$.

### The Riesz representation theorem

A Hilbert space is its own dual: every bounded linear functional $T \in B(H, \mathbb{K})$ is given by an inner product,

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

for a unique $z \in H$. Moreover, $\|T\|_{B(H,\mathbb{K})} = \|z\|_H$.

**N.b.** Note that any function $x \mapsto \langle x, y\rangle$ defines a bounded linear functional on $H$.

**Proof**

**Existence**: Let

$$N = \ker(T).$$

Then $N$ is a closed linear subspace of $H$. If $N = H$, we have $T = 0$ in $B(H, \mathbb{K})$ and $Tx = \langle x, 0 \rangle$.

Assume now that $N \neq H$. According to the corollary above, there exists $z_0 \in N^\perp$, $z_0 \neq 0$. Since $z_0 \perp \ker(T)$ we have $Tz_0 \neq 0$. Consequently,

$$x - \frac{Tx}{Tz_0} z_0 \in \ker(T) \quad \text{for all} \quad x \in H,$$

implying

$$\Big\langle x - \frac{Tx}{Tz_0} z_0, z_0 \Big\rangle = 0 \quad \Leftrightarrow \quad Tx \Big\langle \frac{1}{Tz_0} z_0, z_0 \Big\rangle = \langle x, z_0 \rangle \quad \Leftrightarrow \quad Tx = \frac{Tz_0}{\|z_0\|^2} \langle x, z_0 \rangle = \Big\langle x, \frac{\overline{Tz_0}}{\|z_0\|^2} z_0 \Big\rangle.$$

Thus

$$Tx = \langle x, z \rangle \quad \text{for} \quad z := \frac{\overline{Tz_0}}{\|z_0\|^2} z_0.$$

**Uniqueness**: If, in addition,

$$Tx = \langle x, w \rangle \quad \text{for all} \quad x \in H,$$

then

$$\langle x, z - w \rangle = Tx - Tx = 0 \quad \text{for all} \quad x \in H,$$

so that $z = w$.

**Equality of norms**: We have

$$\|T\| = \sup_{\|x\|=1} |Tx| = \sup_{\|x\|=1} |\langle x, z \rangle| \leq \sup_{\|x\|=1} \|x\| \|z\| = \|z\|,$$

by the Cauchy–Schwarz inequality. Contrariwise,

$$\|z\|^2 = \langle z, z \rangle = |Tz| \leq \|T\| \|z\| \quad \Longrightarrow \quad \|z\| \leq \|T\|.$$

Thus $\|T\| = \|z\|$.

**Ex.**

- $\mathbb{C}^n$ is its own dual: every bounded linear functional on $\mathbb{C}^n$ is realized by a dot product: $Tx = x \cdot \overline{y}$.
- $L_2(\mathbb{R},\mathbb{R})$ is its own dual: every bounded linear functional on $L_2(\mathbb{R},\mathbb{R})$ is realized by an inner product: $Tf = \int_{\mathbb{R}} f(s)\overline{g(s)}\,ds$. By the Cauchy–Schwarz inequality,

$$|Tf| = |\langle f, g\rangle| = \Big|\int_{\mathbb{R}} f(s)\overline{g(s)}\,ds\Big| \leq \Big(\int_{\mathbb{R}} |f(s)|^2\,ds\Big)^{1/2}\Big(\int_{\mathbb{R}} |g(s)|^2\,ds\Big)^{1/2} = \|f\|\|g\|,$$

with equality for $f = \lambda g$; hence, $\|T\| = ||g||$.

## Orthonormal systems, Bessel's inequality and the Fourier series theorem

- A sequence $\{e_j\}_{j\in\mathbb{N}}$ is called **orthogonal** if $e_j \perp e_k$ for $j \neq k$. If, in addition, $\|e_j\| = 1$ for all $j \in \mathbb{N}$, it is called **orthonormal**.
- An orthonormal sequence is called **complete** if there are no non-zero vectors orthogonal to it:

$$K \text{ complete} \quad \overset{\text{def}}{\Longleftrightarrow} \quad K^{\perp} = \{0\}.$$

- If $\{e_j\}_{j\in\mathbb{N}} \subset H$ is an orthonormal sequence, the projection $\langle x, e_j\rangle$ is called the $j$**th Fourier coeffecient of** $x$. The series $\sum_{j\in\mathbb{N}}\langle x, e_j\rangle e_j$ is called the **Fourier series of** $x$ with respect to the sequence $\{e_j\}_{j\in\mathbb{N}}$.

**N.b.** All the above definition carry over to general (finite or infinite) sets, called **orthonormal systems**.

**Ex.**

- The canonical basis $\{e_j\}_j$ is an orthonormal basis in $\mathbb{R}^n$, $\mathbb{C}^n$ and $l_2$ (real or complex).
- The sequence $\{\frac{1}{\sqrt{2\pi}}e^{ikx}\}_{k\in\mathbb{Z}}$ is an ortonormal sequence in $L_2((-\pi,\pi),\mathbb{C})$, since

$$\|\tfrac{1}{\sqrt{2\pi}}e^{ikx}\| = \Big(\int_{-\pi}^{\pi} \tfrac{1}{\sqrt{2\pi}}e^{ikx}\overline{\tfrac{1}{\sqrt{2\pi}}e^{ikx}}\,dx\Big)^{1/2} = \Big(\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ikx}e^{-ikx}\,dx\Big)^{1/2} = 1.$$

### Bessel's inequality

An othonormal sequence satisfies $\sum_{j\in\mathbb{N}} |\langle x, e_j\rangle|^2 \leq \|x\|^2$, for all $x \in X$.

**Proof**

$$\begin{aligned}
0 &\leq \Big\|x - \sum_{j=1}^{N}\langle x, e_j\rangle e_j\Big\|^2 \\
&= \|x\|^2 - 2\Re\Big\langle x, \sum_{j=1}^{N}\langle x, e_j\rangle e_j\Big\rangle + \Big\langle \sum_{j=1}^{N}\langle x, e_j\rangle e_j, \sum_{k=1}^{N}\langle x, e_k\rangle e_k\Big\rangle \\
&= \|x\|^2 - 2\Re\sum_{j=1}^{N}\overline{\langle x, e_j\rangle}\langle x, e_j\rangle + \sum_{j=1}^{N}\sum_{k=1}^{N}\langle x, e_j\rangle\overline{\langle x, e_k\rangle}\langle e_j, e_k\rangle \\
&= \|x\|^2 - \sum_{j=1}^{N}|\langle x, e_j\rangle|^2.
\end{aligned}$$

Thus

$$\sum_{j=1}^{N} |\langle x, e_j\rangle|^2 \leq \|x\|^2,$$

irrespective of $N \in \mathbb{N}$. Bessel's inequality is obtained by letting $N \to \infty$.

**Fourier coefficients are best possible coefficients**
An orthonormal sequence satisfies

$$\Big\|x - \sum_{j=1}^{N} \lambda_j e_j\Big\| \geq \Big\|x - \sum_{j=1}^{N} \langle x, e_j\rangle e_j\Big\|,$$

for any $N \in \mathbb{N}$ and any scalars $\lambda_1, \ldots, \lambda_N \in \mathbb{K}$. Equality holds if and only if $\lambda_j = \langle x, e_j\rangle$ for all $j \in \mathbb{N}$.

**Proof**

$$\begin{aligned}
\Big\|x - \sum_{j=1}^{N} \lambda_j e_j\Big\|^2 &= \|x\|^2 - 2\Re\Big\langle x, \sum_{j=1}^{N} \lambda_j e_j\Big\rangle + \Big\|\sum_{j=1}^{N} \lambda_j e_j\Big\|^2 \\
&= \|x\|^2 - 2\Re\sum_{j=1}^{N} \overline{\lambda_j}\langle x, e_j\rangle + \sum_{j=1}^{N} |\lambda_j|^2 \\
&= \|x\|^2 + \sum_{j=1}^{n} |\langle x, e_j\rangle - \lambda_j|^2 - \sum_{j=1}^{N} |\langle x, e_j\rangle|^2 \\
&\geq \|x\|^2 - \sum_{j=1}^{N} |\langle x, e_j\rangle|^2 \\
&= \Big\|x - \sum_{j=1}^{N} \langle x, e_j\rangle e_j\Big\|^2,
\end{aligned}$$

where the last equality comes from the proof of Bessel's inequality.

**Corollary: closest point**
If $\{e_1, \ldots, e_n\}$ is an orthonormal system, then $y = \sum_{j=1}^{n} \langle x, e_j\rangle e_j$ is the closest point to $x$ in $\operatorname{span}\{e_1, \ldots, e_n\}$, with $d = \|x - y\|$ given by

$$d^2 = \|x\|^2 - \sum_{j=1}^{N} |\langle x, e_j\rangle|^2.$$

**N.b.** In particular, if $x \in \operatorname{span}\{e_1, \ldots, e_n\}$, then $x = \sum_{j=1}^{N} \langle x, e_j\rangle e_j$.

**Proof**

Since Fourier coefficients are best possible, there is no better approximation of $x$ in $\operatorname{span}\{e_1, \ldots, e_n\}$. The distance formula follows from the proof of Bessel's inequality.

**Ex.**

- In $\mathbb{R}^3$, what is the closest point in the plane spanned by $e_1 := \frac{1}{\sqrt{2}}(1,1,0)$ and $e_2 := (0,0,1)$ to the point $x = (2,1,1)$? We have

$$\begin{aligned}\langle x, e_1\rangle e_1 + \langle x, e_2\rangle e_2 &= \big((2,1,1)\cdot \tfrac{1}{\sqrt{2}}(1,1,0)\big)\tfrac{1}{\sqrt{2}}(1,1,0) + \big((2,1,1)\cdot(0,0,1)\big)(0,0,1)\\ &= \tfrac{3}{2}(1,1,0) + (0,0,1) = (\tfrac{3}{2}, \tfrac{3}{2}, 1).\end{aligned}$$

  The distance is

$$\big(\|x\|^2 - |\langle x, e_1\rangle|^2 - |\langle x, e_2\rangle|^2\big)^{1/2} = \big(6 - \tfrac{9}{2} - 1\big)^{1/2} = \frac{1}{\sqrt{2}},$$

  which can be checked to fit with $|(2,1,1) - (\frac{3}{2}, \frac{3}{2}, 1)|$.

**Convergence as an l2-property (in Hilbert spaces)**

Let $\{e_j\}_{j\in\mathbb{N}}$ be an orthonormal sequence in a Hilbert space $H$, and $\{\lambda_j\}_{j\in\mathbb{N}}$ a sequence of scalars. Then

$$\exists \lim_{N\to\infty} \sum_{j=1}^{N} \lambda_j e_j \quad \text{in } H \quad \Longleftrightarrow \quad \sum_{j=1}^{\infty} |\lambda_j|^2 < \infty.$$

In that case, $\| \sum_{j\in\mathbb{N}} \lambda_j e_j\|^2 = \sum_{j\in\mathbb{N}} |\lambda_j|^2$.

**N.b.** A consequence of this is that every infinite-dimensional separable Hilbert space can be identified with $l_2$. If the Hilbert space is finite, it can be identified with $\mathbb{R}^n$ or $\mathbb{C}^n$; if it is not separable, it is bigger than $l_2$.

**Proof**

Let $x_n := \sum_{j=1}^{n} \lambda_j e_j$. For $m > n$,

$$\|x_m - x_n\|^2 = \Big\| \sum_{j=n+1}^{m} \lambda_j e_j \Big\|^2 = \sum_{j,k=n+1}^{m} \lambda_j \overline{\lambda_k} \langle e_j, e_k\rangle = \sum_{j=n+1}^{m} |\lambda_j|^2,$$

meaning that $\{x_n\}_{n\in\mathbb{N}}$ is Cauchy exactly if $\sum_{j=1}^{\infty} |\lambda_j|^2$ converges in $\mathbb{R}$. Since $H$ is complete, this happens exactly if $\{x_n\}_{n\in\mathbb{N}}$ converges in $H$. A similar calculation shows that

$$\Big\| \sum_{j=1}^{m} \lambda_j e_j \Big\|^2 = \sum_{j=1}^{m} |\lambda_j|^2.$$

When (one of) these sums converge we may let $m \to \infty$ to obtain the desired equality.

- An orthonormal system $\{e_j\}_j \subset H$ is called an **orthonormal basis** for $H$ if

$$x = \sum_j \langle x, e_j\rangle e_j \qquad \text{for all } x \in H.$$

**Ex.**

- In $\mathbb{R}^n$, $\mathbb{C}^n$, $l_2(\mathbb{R})$ and $l_2(\mathbb{C})$, the canonical basis $\{e_j\}_j$ is also an orthonormal basis.
- The vectors

$$\tfrac{1}{\sqrt{2}}(1,1,0), \quad \tfrac{1}{\sqrt{2}}(1,-1,0), \quad (0,0,1)$$

  form an orthonormal basis for $\mathbb{R}^3$.

- $\{\frac{1}{\sqrt{2}}, \cos(x), \sin(x), \cos(2x), \sin(2x), \ldots\}$ is an orthonormal basis for $L_2((-\pi, \pi), \mathbb{R})$ if we equip it with the inner product

$$\langle f, g \rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)g(x)\, dx.$$

(One may also use the standard inner product and scale the functions with $1/\sqrt{\pi}$.)

- $\{e^{ikx}\}_{k\in\mathbb{Z}}$ is an orthonormal basis for $L_2((-\pi, \pi), \mathbb{C})$ if we equip it with the inner product

$$\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\overline{g(x)}\, dx.$$

Equivalently, one may use the standard inner product and scale the functions with $1/\sqrt{2\pi}$.

**The Fourier series theorem**

Let $M = \{e_j\}_{j\in\mathbb{N}}$ be an orthonormal sequence in a Hilbert space $H$. Then the following are equivalent:

- $M$ is complete.
- $\overline{\text{span}(M)} = H$.
- $M$ is an orthonormal basis for $H$.
- For all $x \in H$, $\|x\|^2 = \sum_{j\in\mathbb{N}} |\langle x, e_j\rangle|^2$.

**N.b.**

- An analog result holds for orthonormal systems (in particular: for finite sets).
- The last equality is known as **Parseval's identity**.

**Proof**

**(i)** $\Longrightarrow$ **(ii)**: If $M$ is complete, then $M^{\perp} = \{0\}$, so that $\overline{\text{span}(M)} = H$ (else, there would exists a non-zero vector in its orthogonal complement.

**(ii)** $\Longrightarrow$ **(iii)**: If $\overline{\text{span}(M)} = H$, then, for any $x \in H$, there exist $\{\lambda_j\}_{j\in\mathbb{N}}$ such that

$$\lim_{N\to\infty} \sum_{j=1}^{N} \lambda_j e_j = x.$$

But

$$\Big\| \sum_{j=1}^{N} \lambda_j e_j - x \Big\|^2 \geq \Big\| \sum_{j=1}^{N} \langle x, e_j\rangle e_j - x \Big\|^2 \geq 0,$$

so that $x = \sum_{j=1}^{\infty} \langle x, e_j\rangle e_j$.

**(iii)** $\Longrightarrow$ **(iv)**: If $M$ is an orthonormal basis, it is immediate that

$$\|x\|^2 = \Big\langle \sum_{j\in\mathbb{N}} \langle x, e_j\rangle e_j, \sum_{j\in\mathbb{N}} \langle x, e_j\rangle e_j \Big\rangle = \sum_{j\in\mathbb{N}} |\langle x, e_j\rangle|^2.$$

**(iv)** $\Longrightarrow$ **(i)**: Finally, if $\|x\|^2 = \sum_{j\in\mathbb{N}} |\langle x, e_j\rangle|^2$ for all $x \in H$, and $x \perp M$, then $\|x\| = 0$. Hence, there is no non-zero vector in $M^{\perp}$, which is the definition of $M$ being complete.

**Ex.**

- Consider $L_2((-\pi,\pi),\mathbb{C})$ with the orthonormal basis $\{e^{ikx}\}_{k\in\mathbb{Z}}$ and the inner product

$$\langle f,g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx.$$

The Fourier coefficients are given by

$$\hat{f}_k := \langle f, e^{ik\cdot}\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}\,dx,$$

and Parseval's identity states that

$$\|f\|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |f(x)|^2\,dx = \sum_{k=-\infty}^{\infty} |\hat{f}_k|^2 = \sum_{k=-\infty}^{\infty} |\langle f, e^{ik\cdot}\rangle|^2.$$

## 4.3 Adjoints and decompositions

Consider a Hilbert space $H$ over a field $\mathbb{K}\in\{\mathbb{R},\mathbb{C}\}$. Most of the results in this section will be for the cases $H=\mathbb{R}^n$ and $H=\mathbb{C}^n$.

### Adjoints

Let $T\in B(H)$ be a bounded linear operator $H\to H$.

- The **adjoint** of $T$ is the operator $T^*\in B(H)$ defined by

$$\langle Tx,y\rangle = \langle x,T^*y\rangle \quad \text{for all} \quad x,y\in H.$$

- $T$ is called **self-adjoint** if $T=T^*$.

**Properties of the adjoint**

The adjoint is well defined: for each $T\in B(H)$, there exists a unique $T^*\in B(H)$. The map $^*\colon B(H)\to B(H)$, $T\mapsto T^*$ satisfies the following properties:

- It is anti-linear: $(\mu S+\lambda T)^* = \overline{\mu}S^* + \overline{\lambda}T^*$, for all $S,T\in B(H)$ and $\mu,\lambda\in\mathbb{K}$.
- It is bounded with unit norm: $\|T^*\| = \|T\|$.
- It is invertible with itself as inverse: $(T^*)^* = T$.

**N.b.** We adopt the convention that $T^{**} \overset{\text{def.}}{=} (T^*)^*$.

**Proof**

**Existence of the adjoint**: For each $y\in H$, the map $x\mapsto\langle Tx,y\rangle$ is a bounded linear functional, since by the Cauchy–Schwarz inequality and the boundedness of $T$,

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

The Riesz representation theorem thus guarantees that there exists a unique element $y^*\in H$ with

$$\langle Tx,y\rangle = \langle x,y^*\rangle.$$

Define $T^*y := y^*$.

**Linearity and boundedness of the adjoint**: Since $\lambda y_1^* + \mu y_2^*$ is the unique element corresponding to the functional $\langle Tx, \lambda y_1 + \mu y_2\rangle$, the map $y \mapsto T^*y$ is linear. It is also bounded: we have

$$|\langle x, T^*y\rangle| = |\langle Tx, y\rangle| \le \|T\|\|x\|\|y\|,$$

so by choosing $x = T^*y$ we obtain

$$\|T^*y\|^2 \le \|T\|\|T^*y\|\|y\| \quad \Longrightarrow \quad \|T^*y\| \le \|T\|\|y\|,$$

so that $\|T^*\| \le \|T\|$. But since $T^{**} = T$, we also obtain that $\|T\| \le \|T^*\|$, whence $\|T^*\| = \|T\|$.

**Anti-linearity and boundedness of *:** Since

$$\langle x, (\mu S + \lambda T)^* y\rangle = \langle (\mu S + \lambda T)x, y\rangle = \mu\langle Sx, y\rangle + \lambda\langle Tx, y\rangle = \mu\langle x, S^*y\rangle + \lambda\langle x, T^*y\rangle = \langle x, (\overline{\mu}S^* + \overline{\lambda}T^*)y\rangle,$$

the map $^*$ is anti-linear:

$$(\mu S + \lambda T)^* = \overline{\mu}S^* + \overline{\lambda}T^*.$$

In view of that $\|T\| = \|T^*\|$ it follows that $^*$ is bounded with norm 1.

**Invertibility of ***:

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

so that $\langle (T - T^{**})x, y\rangle = 0$ for all $x, y \in H$. Choose $y = (T - T^{**})x$. Then

$$\|(T - T^{**})x\|^2 = 0 \quad \text{for all} \quad x \in H,$$

meaning that $T = T^{**}$ in $B(H)$.

### The adjoint of a matrix

For matrices, we extend this definition (in that $A$ need not map $H$ to $H$):

- The adjoint of $A \in M_{m\times n}(\mathbb{R})$ is its **transpose** $A^t \in M_{n\times m}(\mathbb{R})$.
- The adjoint of $A \in M_{m\times n}(\mathbb{C})$ is its **conjugate transpose** $A^* \in M_{n\times m}(\mathbb{C})$.

**N.b.** Note that, in the case $m = n$, this fits with the above definition if one adopts the standard inner product on $\mathbb{K}^n$.

### Self-adjoint matrices are symmetric or hermitian

- If $T \in B(\mathbb{R}^n)$ is realized by a matrix $A \in M_{n\times n}(\mathbb{R})$, $T$ is self-adjoint if and only if $A$ is symmetric.
- If $T \in B(\mathbb{C}^n)$ is realized by a matrix $A \in M_{n\times n}(\mathbb{C})$, $T$ is self-adjoint if and only if $A$ is hermitian.

**Proof**

**Real case**:

$$\langle Tx, y\rangle = \langle x, Ty\rangle \quad \Longleftrightarrow \quad y^tAx = (Ay)^tx = y^tA^tx.$$

By considering $x = e_j$, $y = e_k$ we get that $A_{jk} = A^t_{jk}$ for all $j, k = 1, \ldots, n$.

**Complex case**:

$$\langle Tx, y\rangle = \langle x, Ty\rangle \quad \Longleftrightarrow \quad y^*Ax = (Ay)^*x = y^*A^*x.$$

By considering $x = e_j$, $y = e_k$ we get that $A_{jk} = A^*_{jk}$ for all $j, k = 1, \ldots, n$.

### Self-adjoint operators have real spectrum

The eigenvalues of a self-adjoint operator are real, and eigenspaces corresponding to different eigenvalues are orthogonal.

**Proof**

If $T = T^*$, then

$$\langle Tx, x\rangle = \langle x, Tx\rangle = \overline{\langle Tx, x\rangle} \in \mathbb{R} \quad \text{for all } x \in H.$$

Hence, if $\mu, \lambda \in \mathbb{C}$, $\mu \neq \lambda$, are eigenvalues of $T$, with eigenvectors $x, y$, respectively, then

$$\mu\|x\|^2 = \langle \mu x, x\rangle = \langle Tx, x\rangle \in \mathbb{R},$$

so that $\mu \in \mathbb{R}$ (and, similarly, $\lambda \in \mathbb{R}$). Then

$$(\mu - \lambda)\langle x, y\rangle = \langle \mu x, y\rangle - \langle x, \lambda y\rangle = \langle Tx, y\rangle - \langle x, Ty\rangle = 0,$$

since $T$ is self-adjoint. Thus $x \perp y$.

### Unitary operators and orthogonal matrices

- An operator $U \in B(H)$ is called **unitary** if $UU^* = U^*U = \mathrm{Id}$. Similarly, a matrix $A \in M_{n\times n}(\mathbb{K})$ is called **unitary** if the corresponding operator is unitary, i.e., if $A^*A = I$.
- A unitary real matrix $Q \in M_{n\times n}(\mathbb{R})$ is called **orthogonal**.

### Unitary operators preserve inner products

If $U \in B(H)$ is unitary, then

$$\langle Ux, Uy\rangle = \langle x, y\rangle, \quad \text{for all} \quad x, y \in H.$$

In particular, $U$ is an isometry.

**Proof**

$$\langle Ux, Uy\rangle = \langle x, U^*Uy\rangle = \langle x, y\rangle.$$

### Lemma: there are no nontrivial nilpotent self-adjoint operators

If $N = N^*$ satisfies $N^k = 0$ for some $k \in \mathbb{N}$, then $N = 0$.

**Proof**

If $N^2 = 0$, consider

$$\|Nx\|^2 = \langle Nx, Nx\rangle = \langle x, N^2x\rangle = 0,$$

to see that $Nx = 0$ for all $x \in H$.

Else, let $k_0 \geq 2$ be the smallest positive even number such that $N^{2k_0} = 0$, and note that

$$\|N^{k_0}x\|^2 = \langle N^{k_0}x, N^{k_0}x\rangle = \langle x, N^{2k_0}x\rangle = 0.$$

Hence, $N^{k_0} = 0$ and either $k = k_0$ or $k = k_0 + 1$ is a strictly smaller positive even number satisfying $N^k = 0$. This is a contradiction, so there is no such number $k_0$.

**Orthogonal matrices matrices describe orthonormal bases**
The columns $\{Q_1, \dots, Q_n\}$ of an orthogonal (unitary) matrix $Q$ is an orthonormal basis for $\mathbb{R}^n$ $(\mathbb{C}^N)$.

**N.b.** The same is true for the rows of $Q$.

**Proof**

Since the rows of $Q^*$ are columns of $\overline{Q}$, we have

$$Q^*Q = ((Q_i)_i)^*(Q_j)_j = (\overline{Q_i} \cdot Q_j)_{ij} = I$$

if and only if $Q_i \perp Q_j$ for $i \neq j$, and $|Q_i| = 1$.

## The spectral theorem

Let $A \in M_{n\times n}(\mathbb{K})$ be symmetric (hermitian). Then there exists an orthonormal basis $\{Q_j\}_{j=1}^n$ for $\mathbb{R}^n$ $(\mathbb{C}^n)$ of eigenvectors of $A$, such that

$$A = QDQ^*,$$

where $Q$ is the orthogonal (unitary) matrix with columns $(Q_j)_j$ and $D$ is a diagonal matrix with the eigenvalues of $A$ as its diagonal elements.

**N.b.** It is possible to extend the spectral theorem to all **normal** matrices, characterized by $AA^* = A^*A$.[1]

**Proof**

**Each eigenspace is maximal**: Let $\lambda$ an eigenvalue of $A$ and pick $x \in \ker((A-\lambda I)^2)$. Since $A$ is self-adjoint with real eigenvalues we have

$$0 = \langle (A-\lambda I)^2 x, x\rangle = \|(A-\lambda I)x\|^2 \quad \Longrightarrow \quad x \in \ker(A-\lambda I).$$

Hence the Riesz index of $\lambda$ is 1, and all eigenvalues are semi-simple, meaning that $\dim(\ker(A-\lambda I)) = \text{mult}(\lambda)$.

**Applying the spectral decomposition**: The statement now follows from the spectral (or Jordan) decomposition: The maximal generalized eigenspaces coincide with the eigenspaces, these are mutually orthogonal, and we may pick an orthonormal basis for each of them. Together, these form an orthonormal basis for $K^n$, described by $Q$.

**Diagonalization by direct computation**: With $D + N$ representing the diagonal and nilpotent part of the spectral representation of $A$, one can see directly that

$$Q(D+N)Q^{-1} = A = A^* = (Q(D+N)Q^{-1})^* = (Q^{-1})^*(D+N)^*Q^* = Q(D+N^*)Q^{-1},$$

in view of that $D = D^*$ is a diagonal real matrix. By applying $Q^{-1} = Q^*$ from the left and $Q = (Q^*)^{-1}$ from the right, we obtain that $N = N^*$, whence $N = 0$ by the above lemma. Thus, $A$ is diagonalized by the spectral decomposition given by the orthonormal basis $Q$.

## Positive definiteness

Let $A = A^t \in M_{n\times n}(\mathbb{R})$ be a symmetric matrix.

[1] In fact, the class of normal matrices is the biggest class of matrices for which the spectral theorem holds.

- $A$ is said to be **positive definite** if

$$\langle Ax, x\rangle = x^t Ax > 0 \quad \text{for} \quad x \neq 0.$$

- $A$ is said to be **positive semi-definite** if $\langle Ax, x\rangle = x^t Ax \geq 0$ for all $x \in \mathbb{R}^n$.

### Characterization of positive definite matrices

A symmetric matrix $A = A^t \in M_{n\times n}(\mathbb{R})$ is positive definite exactly if one (and hence all) of the following conditions hold:

- $\langle A\cdot, \cdot\rangle$ defines an inner product.
- All the eigenvalues of $A$ are strictly positive.
- $A = R^t R$ for some invertible matrix $R$.

**Proof (contains important methods)**

**Inner-product property**. Assume that $A$ is positive definite. Since $x \mapsto Ax$, $\mathbb{R}^n \to \mathbb{R}^n$, is linear, and the usual inner product is sesqui-linear (linear in its first argument, anti-linear in its second), the form

$$(x, y) \mapsto \langle Ax, y\rangle, \qquad \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R},$$

is also sesqui-linear. Furthermore,

$$\langle Ax, y\rangle = \langle x, Ay\rangle = \langle Ay, x\rangle,$$

by the symmetry of $A$ and of the standard inner product, and

$$\langle Ax, x\rangle > 0 \quad \text{for} \quad x \neq 0,$$

by the definition of positive definiteness, so the inner product $\langle A\cdot, \cdot\rangle$ is non-degenerate symmetric.

Considering the same arguments, one also sees that $\langle A\cdot, \cdot\rangle$ cannot be an inner product unless $A$ is positive definite.

**Eigenvalue property**: Since $A$ is symmetric, there exists an orthonormal basis of eigenvectors $\{v_1, \ldots, v_n\}$ with $Av_j = \lambda_j v_j$. Let $x_j$ be the coordinates of a vector $x$ in this basis. Then

$$\langle Ax, x\rangle = \Big\langle A\sum_{j=1}^n x_j v_j, \sum_{k=1}^n x_k v_k\Big\rangle = \Big\langle \sum_{j=1}^n x_j(Av_j), \sum_{k=1}^n x_k v_k\Big\rangle = \sum_{j,k=1}^n \lambda_j\langle x_j v_j, x_k v_k\rangle = \lambda_j x_j^2 > 0$$

for all $x \neq 0$ if and only if $\lambda_j > 0$ for all $j = 1, \ldots, n$.

**Factorization property**: If $A = R^t R$ with $R$ invertible, then

$$\langle Ax, x\rangle = \langle R^t Rx, x\rangle = \|Rx\|^2 > 0,$$

unless $Rx = 0$, which happens only if $x = 0$ (as $R$ is invertible).

Contrariwise, if $A$ is symmetric and positive definite, we can write

$$A = Q^t DQ = Q^t\sqrt{D}\sqrt{D}Q = Q^t(\sqrt{D})^t\sqrt{D}Q = (\sqrt{D}Q)^t(\sqrt{D}Q),$$

where $Q$ is an orthogonal matrix of eigenvectors, $D = (\lambda_j)_j$ is a diagonal matrix with positive eigenvalues, and $\sqrt{D} = (\sqrt{\lambda_j})_j$ has $\sqrt{\lambda_j}$ as diagonal elements. Thus, $A = R^T R$.

### The Cholesky decomposition

An alternative to the spectral decomposition used in the proof of that $A = R^tR$ for positive definite matrices is the **Cholesky decomposition**: If one divides out the main pivots (diagonal elements) in the $LU$-factorization of $A$, one gets an $LDU$-**decomposition**, $D$ being a diagonal matrix with the main pivots as diagonal elements, and $L$ *and* $U$ *having only unit elements on their main diagonals.* This factorization is unique. We thus obtain

$$LDU = A = A^t = (LDU)^t = U^tD^tL^t = U^tDL^t,$$

where $U^t$ is lower triangular, and $L^t$ is upper triangular. By uniqueness of the $LDU$-factorization, we must have $L = U^t$ and $U = L^t$. Consequently,

$$A = LDL^t = L\sqrt{D}\sqrt{D}L^t = L\sqrt{D}(L\sqrt{D})^t,$$

with $L\sqrt{D}$ being invertible ($\sqrt{D}$ has only positive elements along the diagonal and $L$ is lower triangular with units along the diagonal).

## Singular values

### The singular value theorem

Let $A \in M_{m\times n}(\mathbb{R})$ be the realization of a linear transformation $\mathbb{R}^n \to \mathbb{R}^m$, such that $\mathrm{rank}(A) = r$. Then $A^tA$ is a positive semi-definite matrix of rank $r$, and there exists an orthonormal basis of eigenvectors of $A^tA$,

$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^n, \qquad \text{with eigenvalues} \quad \sigma_1^2 \geq \ldots \geq \sigma_r^2 > 0, \quad \sigma_{r+1}^2 = \ldots = \sigma_n^2 = 0,$$

and a corresponding orthonormal basis

$$\{u_1, \ldots, u_r, u_{r+1}, \ldots, u_m\} := \{\frac{1}{\sigma_1}Av_1, \ldots, \frac{1}{\sigma_r}Av_r, u_{r+1}, \ldots, u_m\}$$

for $\mathbb{R}^m$ ($u_{r+1}, \ldots, u_m$ arbitrary to fit the orthonormal basis), such that

$$Av_j = \begin{cases} \sigma_j u_j, & j = 1, \ldots, r, \\ 0, & j = r+1, \ldots, n. \end{cases}$$

The unique scalars $\sigma_1, \ldots, \sigma_r, 0, \ldots, 0$ (extended to a total of $\min(m, n)$) are called **singular values** of $A$. If

$$(\Sigma_{ij})_{ij} := (\delta_{ij}\sigma_j)_{ij} \in M_{m\times n}(\mathbb{R})$$

is the diagonal matrix with $\sigma_1, \ldots, \sigma_r, 0, \ldots, 0$ on its main diagonal, $U = (u_j)_j \in M_{m\times m}(\mathbb{R})$ is the orthogonal matrix with $u_j$ as columns, and $V = (v_j)_j \in M_{n\times n}(\mathbb{R})$ is the orthogonal matrix with $v_j$ as columns, it follows that

$$A = U\Sigma V^{-1} = U\Sigma V^t.$$

This is the **singular value decomposition** of $A$.

**N.b.** The singular values for $A^t$ equals those of $A$. For $A^t$, the orthonormal bases $V$ and $U$ are simply exchanged (in comparison to $A$).

### The pseudoinverse

A finite-dimensional linear transformation can always be inverted on its range. Let $A \in M_{m\times n}(\mathbb{R})$ be the realization of a linear transformation $\mathbb{R}^n \to \mathbb{R}^m$, and let

$$A|_{\ker(A)^\perp} \colon \ker(A)^\perp \to \mathrm{ran}(A)$$

be the restriction of $A$ to the orthogonal complement of its null space. Then (according to the rank–nullity theorem) $A|_{\ker(A)}$ is bijective. Its inverse, $A^\dagger$, is called the **(Moore-Penrose) pseudoinverse**:

$$A^\dagger \colon \mathrm{ran}(A) \to \ker(A)^\perp, \qquad A^\dagger(Ax) = x.$$

### The pseudoinverse from the singular value decomposition

If $A = U\Sigma V^t$ is the singular value decomposition of $A \in M_{m\times n}(\mathbb{R})$, using the bases $\{u_j\}_j$ and $\{v_j\}_j$, we only have to invert $Av_j = \sigma_j u_j$ for $j = 1, \ldots, r$.

More precisely, if

$$(\Sigma^\dagger_{ij})_{ij} = (\delta_{ij}\frac{1}{\sigma_j})_{ij} \in M_{n\times m}(\mathbb{R})$$

is the diagonal matrix with the reciprocals of the singular values along its main diagonal, then

$$A^\dagger = V\Sigma^\dagger U^t$$

is the singular value decomposition of the pseudoinverse $A^\dagger$.

## The Gram–Schmidt orthogonalization and QR-decompositions

### Gram–Schmidt

A set $\{v_1, \ldots, v_n\} \subset H$ of linearly independent vectors can transform into an orthonormal system using the **Gram–Schmidt orthogonalization** algorithm.[1] Define

$$e_1 := \frac{v_1}{\|v_1\|},$$

$$\tilde{e}_2 := v_2 - \langle v_2, e_1\rangle e_1, \qquad e_2 := \frac{\tilde{e}_2}{\|\tilde{e}_2\|},$$

and, recursively,

$$\tilde{e}_{k+1} := v_{k+1} - \sum_{j=1}^{k}\langle v_{k+1}, e_j\rangle e_j, \qquad e_{k+1} := \frac{\tilde{e}_{k+1}}{\|\tilde{e}_{k+1}\|}, \qquad k = 1, \ldots, n-1.$$

Then $\{e_1, \ldots, e_n\}$ is an orthonormal system.

### QR-decompositions

If, in the Gram–Schmidt orthoghonalization, we express the vectors $\{v_1, \ldots, v_n\}$ in terms of $\{e_1, \ldots, e_n\}$, we obtain

$$v_1 = \langle v_1, e_1\rangle e_1, \qquad v_2 = \langle v_2, e_1\rangle e_1 + \langle v_2, e_2\rangle e_2, \qquad v_k = \sum_{j=1}^{k}\langle v_k, e_j\rangle e_j,$$

which can be seen either by direct calculation or from the 'closest point'-corollary (since $\operatorname{span}\{v_1, \ldots v_k\} = \operatorname{span}\{e_1, \ldots e_k\}$ for $k = 1, \ldots, n$). This gives us the $QR$-decomposition of a full-rank matrix $A$:

If $A = [v_1, \ldots, v_n] \in M_{n\times n}(\mathbb{R})$ is matrix of full rank (so that its column vectors span $\mathbb{R}^n$), the Gram–Schmidt orthogonalization applied to the columns vectors $v_1, \ldots, v_n$ yields the QR-decomposition of $A$:

$$A = QR = \begin{bmatrix}[e_1] & [e_2] & \cdots & [e_n]\end{bmatrix}\begin{bmatrix}\langle v_1, e_1\rangle & \langle v_2, e_1\rangle & \ldots & \langle v_n, e_1\rangle \\ 0 & \langle v_2, e_2\rangle & \ldots & \langle v_n, e_2\rangle \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \langle v_n, e_n\rangle\end{bmatrix},$$

where $Q$ is orthogonal ($Q^t = Q^{-1}$), and $R$ is upper (right) triangular. Consequently, one can calculate $R = Q^t A$ by finding the orthonormal basis given by $Q$.

**N.b.**

[1] The Gram–Schmidt orthogonalization is equally valid for linearly independent sequences.

- It is possible to extend the QR-decomposition to general rectangular matrices.
- Just like the LU-decomposition, the QR-decomposition can help in solving linear systems, since

  $$Ax = b \quad \Longleftrightarrow \quad QRx = b \quad \Longleftrightarrow \quad Rx = Q^t b,$$

  where $R$ is uppper triangular.

# Index