Foundations of Deep Learning **Fall 2025**

Lecture 0: Prerequisites

Lecturer: Aurelien Lucchi

1 Notations & Math symbols

1.1 Vectors

A column vector is a $d \times 1$ array that is, an array consisting of a single column of d elements, denoted as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}.$$

Similarly, a row vector is a $1 \times d$ array that is, an array consisting of a single row of d elements, denoted as

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_d \end{bmatrix}.$$

Throughout, boldface is used for the row and column vectors. The transpose (indicated by $\top$) of a row vector is a column vector. We also recall that the inner product between two vectors $\mathbf{u} \in \mathbb{R}^d$ and $\mathbf{v} \in \mathbb{R}^d$ is defined as

$$\mathbf{u}^\top \mathbf{v} = \sum_{i=1}^{d} u_i v_i.$$

The **span** of a set of vectors is the set of all possible linear combinations of those vectors. If we have a set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ in a vector space V, the span of these vectors, denoted by $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)$, is defined as:

$$\text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k) = \{c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k \mid c_1, c_2, \dots, c_k \in \mathbb{R}\}.$$

Consider the following example illustrated in Figure 1 where we show the span of two vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^2$. Taking for instance coefficients $c_1 = 0.5$ and $c_2 = 1$ yields the vector $\begin{pmatrix} 0.5 \\ 4 \end{pmatrix}$ since:

$$\underbrace{\begin{pmatrix} -3 & 2 \\ 2 & 3 \end{pmatrix}}_{\mathbf{V}} \begin{pmatrix} 0.5 \\ 1 \end{pmatrix} = 0.5 \underbrace{\begin{pmatrix} -3 \\ 2 \end{pmatrix}}_{\mathbf{v}_1} + 1 \underbrace{\begin{pmatrix} 2 \\ 3 \end{pmatrix}}_{\mathbf{v}_1} = \begin{pmatrix} 0.5 \\ 4 \end{pmatrix}$$

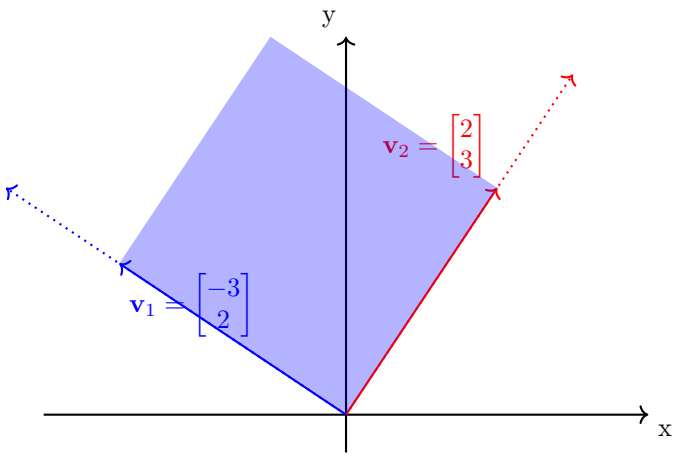


Figure 1: Illustration of the span of two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ in $\mathbb{R}^2$. The blue parallelogram shows the span of these two vectors for coefficients $\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$ where $c_i \in [0, 1]$.

As another example, consider the span of two dependent vectors $\mathbf{v}_2 = 2\mathbf{v}_1$, then one can check that we can only obtain a multiple of the vector $\mathbf{v}_1$, for instance:

$$\underbrace{\begin{pmatrix}1 & 2\\ 2 & 4\end{pmatrix}}_{\mathbf{V}}\begin{pmatrix}1\\ 0.5\end{pmatrix} = 1\underbrace{\begin{pmatrix}1\\ 2\end{pmatrix}}_{\mathbf{v}_1} + 0.5\underbrace{\begin{pmatrix}2\\ 4\end{pmatrix}}_{\mathbf{v}_1} = \begin{pmatrix}2\\ 4\end{pmatrix}$$

If we consider the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ as columns of a matrix $\mathbf{V}$, then the linear combinations of these vectors constitute the column space, as described in the definition below. A similar definition applies to the row space.

Definition 1. *The column space (also called the range or image) of a matrix $\mathbf{A}$ is the span (set of all possible linear combinations) of its column vectors.*

1.2 Matrices

Matrices will be denoted by a boldface capital letter, for instance

$$\mathbf{A} = \begin{pmatrix}A_{11} & \dots & A_{1d}\\ & \dots & \\ A_{p1} & \dots & A_{pd}\end{pmatrix}, \quad \mathbf{A}^\top = \begin{pmatrix}A_{11} & \dots & A_{p1}\\ & \dots & \\ A_{1d} & \dots & A_{pd}\end{pmatrix}$$

Recall that the multiplication of two matrices $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times d}$ is

$$\mathbf{C} = \mathbf{AB}, \quad C_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}.$$

The matrix multiplication operation has the following properties:

- **Distributive property:** Matrix multiplication is distributive over matrix addition. For matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ of compatible dimensions,
$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC} \quad \text{and} \quad (\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}.$$
- **Associative property:** Matrix multiplication is associative. For matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ of compatible dimensions,
$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}).$$
- **Not commutative:** Matrix multiplication is generally not commutative. For matrices $\mathbf{A}$ and $\mathbf{B}$ of compatible dimensions,
$$\mathbf{AB} \neq \mathbf{BA} \quad \text{in general.}$$

The **inverse** of a matrix $\mathbf{A}$ is a matrix, denoted by $\mathbf{A}^{-1}$, such that when it is multiplied by $\mathbf{A}$, it yields the identity matrix. Specifically, for an $n \times n$ matrix $\mathbf{A}$,

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n,$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

1.3 Notation

- Bold small letter $\mathbf{x}$ is a column vector
- Bold capital letter $\mathbf{A}$ is a matrix
- $\mathbf{x}^\top, \mathbf{A}^\top$: transpose of a vector (i.e. a row vector) and a matrix respectively
- $\mathbb{R}$: reals
- $a := b$: a is defined by b
- $\frac{\partial f}{\partial x}$: partial derivative of a function f with respect to x
- $\frac{df}{dx}$: total derivative of a function f with respect to x
- ∇: gradient
- $\|\cdot\|$: norm (by default $\|\mathbf{x}\| = \|\mathbf{x}\|_2$).

Math symbols

- $C(X, Y)$ Space of continuous functions $f : X \to Y$.
- $C(\mathbb{R})$ space of continuous functions $f : \mathbb{R} \to \mathbb{R}$, usually endowed with the uniform norm topology
- $C_c(\mathbb{R})$ continuous functions with compact support
- $C([a, b])$ space of all continuous functions that are defined on a closed interval $[a, b]$
- $C_b(\mathbb{R})$ space of continuous bounded functions
- $B(\mathbb{R})$ bounded functions
- $C_0(\mathbb{R})$ continuous functions which vanish at infinity
- $C^r(\mathbb{R})$ continuous functions that have continuous first r derivatives.
- $C^\infty(\mathbb{R})$ smooth functions
- $C_c^\infty(\mathbb{R})$ smooth functions compact support
- $\mathbb{F}$ Field of either real $\mathbb{R}$ or complex numbers $\mathcal{C}$
- ℓ^p is used to indicate a p-summable *discrete* set of values. For example, $\ell^p(\mathbb{Z}^+)$ is the set of complex-valued sequences $\{(a_n)\}$ such that $\sum_{n\in\mathbb{Z}^+} |a_n|^p < \infty$. For example:
  - ℓ^1, the space of sequences whose series is absolutely convergent
  - ℓ^2, the space of square-summable sequences, which is a Hilbert space
  - ℓ^∞, the space of bounded sequences
- L^p is typically used to indicate p-summable functions (with respect to some measure) on a *non-discrete* measure space, such as the usual $L^p(\mathbb{R})$, the set of functions $f : \mathbb{R} \to \mathbb{C}$ such that $\int_{\mathbb{R}} |f(x)|^p \, dx < \infty$.
- The expectation of a random variable X is denoted by $\mathbb{E}[X]$ or $\mathbb{E}_D[X]$ to make the distribution of X, denoted by D, explicit
- Given a symmetric matrix $\mathbf{A}$, $\mathbf{A} \succcurlyeq 0$ means that $\mathbf{A}$ is positive semidefinite, i.e. $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$.

2 Vector spaces

Vector space Recall that a vector space (also called a linear space) is a collection of objects called vectors, which may be added together and multiplied by scalars. Formally, a vector space over a field F (e.g. the field of real numbers) is a set V together with two operations that satisfy several axioms (associativity, commutativity, identity, inverse).

Subspace A subspace of a vector space V is a subset U of V that is itself a vector space under the same operations of vector addition and scalar multiplication as those defined on V. One simple example is a line through the origin in $\mathbb{R}^2$.

Norm A normed vector space is a space equipped with a norm, i.e. a function from V to $\mathbb{R}$ (we give a formal definition of a norm below). Informally this means that we need to be able to add vectors and scale them with a scalar.

Definition 2. *Given a vector space V over a subfield F of the complex numbers, a norm on V is a nonnegative-valued scalar function $p : V \to [0, +\infty)$ with the following properties:*
For all $a \in F$ and all $u, v \in V$,

- $p(v) \geq 0$ *(non-negativity)*
- *If* $p(v) = 0$ *then* $v = 0$ *(positive definite)*
- $p(av) = |a|p(v)$ *(absolutely homogeneous)*
- $p(u + v) \leq p(u) + p(v)$ *(triangle inequality).*

Example 1: vector norm A vector norm is a function that assigns a non-negative scalar value to a vector in a vector space, which intuitively represents the length or size of the vector. More formally, if $\mathbf{v}$ is a vector in a vector space V, a norm $\|\mathbf{v}\|$ satisfies the following properties:

i) **Non-negativity (or Positivity)**:

$$\|\mathbf{v}\| \geq 0 \quad \text{and} \quad \|\mathbf{v}\| = 0 \iff \mathbf{v} = \mathbf{0}$$

This means the norm of any vector is always non-negative, and it is zero if and only if the vector itself is the zero vector.

ii) **Scalar Multiplication**:

$$\|\alpha\mathbf{v}\| = |\alpha|\|\mathbf{v}\|$$

for any scalar α. This means that scaling a vector by a scalar α scales the norm of the vector by the absolute value of α.

iii) **Triangle Inequality**:

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$$

for any vectors $\mathbf{u}$ and $\mathbf{v}$. This means that the norm of the sum of two vectors is less than or equal to the sum of the norms of the two vectors.

Common examples of vector norms include:

- **Euclidean norm (or 2-norm)**:

$$\|\mathbf{v}\|_2 = \left(\sum_{i=1}^{n} |v_i|^2\right)^{1/2}$$

where $\mathbf{v} = (v_1, v_2, \ldots, v_n)$.

- **1-norm (or Manhattan norm)**:

$$\|\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i|$$

- **Infinity norm (or maximum norm)**:

$$\|\mathbf{v}\|_\infty = \max_{1 \leq i \leq n} |v_i|$$

Example 2: matrix norm Suppose a vector norm $\|\cdot\|$ on $\mathbb{R}^m$ is given (e.g. the Euclidean norm). Any $m \times n$ matrix $\mathbf{A}$ induces a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$ with respect to the standard basis, and one defines the corresponding operator norm on the space $\mathbb{R}^{m \times n}$ of all $m \times n$ matrices as follows:

$$\begin{aligned} \|\mathbf{A}\| &= \sup\{\|\mathbf{A}\mathbf{x}\| : \mathbf{x} \in \mathbb{R}^n \text{ with } \|\mathbf{x}\| = 1\} \\ &= \sup\left\{\frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} : \mathbf{x} \in \mathbb{R}^n \text{ with } \mathbf{x} \neq 0\right\}. \end{aligned} \tag{1}$$

In particular, if the p-norm for vectors is used for both spaces $\mathbb{R}^n$ and $\mathbb{R}^m$, the corresponding induced operator norm is:

$$\|\mathbf{A}\|_p = \sup\left\{\frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} : \mathbf{x} \in \mathbb{R}^n \text{ with } \mathbf{x} \neq 0\right\}. \tag{2}$$

In practice, one is often interested in the case $p = 2$, for which

$$\begin{aligned} \|\mathbf{A}\|_2 &= \sup\left\{\frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} : \mathbf{x} \in \mathbb{R}^n \text{ with } \mathbf{x} \neq 0\right\} \\ &= \sigma_1(\mathbf{A}), \end{aligned} \tag{3}$$

where $\sigma_1(\mathbf{A})$ is the largest singular value of $\mathbf{A}$.

Norm inequalities We introduce the Cauchy-Schwarz inequality, which is a pillar inequality in the field of mathematics and will be a recurrent inequality in many proofs discussed in this course.

Proposition 1. *The Cauchy-Schwarz inequality states that for all vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *of an inner product space, the following holds:*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|. \tag{4}$$

Writing $\mathbf{x} = [x_1 \dots x_n]$ *and* $\mathbf{y} = [y_1 \dots y_n]$*, this can also be written as*

$$\left(\sum_{i=1}^{n} x_i y_i \right)^2 \leq \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2. \tag{5}$$

Equality occurs if and only if there exists a constant c *such that* $x_i = c y_i \; \forall i = 1, \dots n$*.*

Proof. We will give two different proofs (there are many more):

Proof I) A very simple proof can be obtained using the definition of an inner product:

$$\mathbf{x}^\top \mathbf{y} = \|\mathbf{x}\| \|\mathbf{y}\| \cos(\theta), \tag{6}$$

with the fact that $\cos(\theta) \leq 1$.

Proof II) Another proof is as follows. Define the following non-negative function:

$$\begin{aligned} f(z) &= \sum_{i=1}^{n} (x_i - z y_i)^2 \geq 0 \\ &= \sum_{i=1}^{n} (x_i^2 - 2 x_i y_i z + z^2 y_i^2) \\ &= \left(\sum_{i=1}^{n} y_i^2 \right) z^2 - 2 \left(\sum_{i=1}^{n} x_i y_i \right) z + \sum_{i=1}^{n} x_i^2. \end{aligned}$$

Note that this is a quadratic function in z of the form:

$$f(z) = A z^2 - B z + C,$$

and whose minimum value occurs when $z = \frac{B}{2A}$. Since the whole expression has to be non-negative, we need

$$\Delta = B^2 - 4AC \leq 0 \implies C \geq \frac{B^2}{4A} \implies B^2 \leq 4AC.$$

One can verify that $B^2 \leq 4AC$ implies the inequality we are looking for:

$$\left(\sum_{i=1}^{n} x_i y_i \right)^2 \leq 4 \left(\sum_{i=1}^{n} y_i^2 \right) \left(\sum_{i=1}^{n} x_i^2 \right). \tag{7}$$

□

3 Functions

We here review the definition of a gradient and then discuss two fundamental properties of functions.

3.1 Gradients

Consider a real-valued (univariate) function $f : \mathbb{R} \to \mathbb{R}$. Its derivative is defined by

$$f'(x) := \lim_{\epsilon \to 0} \frac{f(x + \epsilon) - f(x)}{\epsilon}. \tag{8}$$

The function f is *differentiable* at x if the limit in Eq. (8) exists.

Let's generalize this definition to the case where the function f is multivariate, i.e. $f : \mathbb{R}^d \to \mathbb{R}$. In this case, we have one (partial) derivative for each dimension, i.e. $\frac{\partial f}{\partial x_i}$ defined as

$$\frac{\partial f}{\partial x_i} := \lim_{\epsilon \to 0} \frac{f(x_1, \ldots, x_i + \epsilon, \ldots, x_d) - f(x_1, \ldots, x_i, \ldots, x_d))}{\epsilon}. \tag{9}$$

The gradient denoted by $\nabla f(\mathbf{x})$ gives us a way to pack all partial derivatives into one vector. Denoting by $\mathbf{x} = (x_1, x_2, \ldots x_d)$, we then define the gradient as

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d}, \end{pmatrix} \tag{10}$$

where we can also define

$$\frac{\partial f}{\partial x_i} := \lim_{\epsilon \to 0} \frac{f(\mathbf{x} + \epsilon \mathbf{e}_i) - f(\mathbf{x})}{\epsilon}, \tag{11}$$

with $\mathbf{e}_i$ being a standard basis vector (composed of zeros and a single one at the i-th coordinate). Note that $\frac{\partial f}{\partial x_i}$ is also the directional derivative of f along the direction $\mathbf{e}_i$, and can be expressed as $\nabla f(\mathbf{x}) \cdot \mathbf{e}_i$ (where $\cdot$ is the standard inner product in $\mathbb{R}^d$.

3.2 Chain rule

The chain rule is a formula to compute the derivative of a composite function $f(g(x))$. We start by describing the case where the functions involved in the composition are single variable functions, i.e. $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$. In this case, the chain rule is

$$\frac{d}{dx} f(g(x)) = f'(g(x)) g'(x)$$

Alternatively, it can also be written as

$$\frac{d}{dx} f(g(x)) = \frac{df}{dg} \cdot \frac{dg}{dx},$$

where dg is an abbreviation for $dg(x)$.

Proof idea. First note that by definition of the derivative as a limit:

$$(f \circ g)'(a) = \lim_{x \to a} \frac{f(g(x)) - f(g(a))}{x - a}.$$

Assuming that $g(x) \neq g(a)$ any x near a, then the previous expression is equal to the product of two factors:

$$\lim_{x \to a} \frac{f(g(x)) - f(g(a))}{g(x) - g(a)} \cdot \frac{g(x) - g(a)}{x - a} = \lim_{x \to a} \frac{f(g(x)) - f(g(a))}{g(x) - g(a)} \cdot \lim_{x \to a} \frac{g(x) - g(a)}{x - a},$$

where the equality uses the fact that the limit of a product is equal to the product of the limits (limit law).

We have therefore reached the desired expression. The case $g(x) = g(a)$ has to be handled with more care, see e.g. Rudin et al. (1976) for a complete proof.

□

Let's consider the general case where $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^d$ and $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^m$. Then the composed function is $f(g(\mathbf{x})) : \mathbb{R}^k \to \mathbb{R}^d$. In this case, the chain rule is written as

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{f}(\mathbf{g}(\mathbf{x})) = \frac{\partial \mathbf{f}}{\partial \mathbf{g}}(g(\mathbf{x})) \cdot \frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}) \in \mathbb{R}^{d \times k},$$

where $\frac{\partial \mathbf{f}}{\partial \mathbf{g}}(g(\mathbf{x})) \in \mathbb{R}^{d \times m}$ is the Jacobian matrix of $\mathbf{f}$ with respect to $\mathbf{g}$ and $\frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}) \in \mathbb{R}^{m \times k}$ is the Jacobian matrix of $\mathbf{g}$ with respect to $\mathbf{x}$.

The Jacobian $\frac{\partial \mathbf{f}}{\partial \mathbf{g}}(g(\mathbf{x})) \in \mathbb{R}^{d \times m}$ contains all first-order partial derivatives of the components of $\mathbf{f}$ w.r.t. the components of $\mathbf{g}$. Specifically, if $\mathbf{f} = [f_1, f_2, \ldots, f_d]^\top$ and $\mathbf{g} = [g_1, g_2, \ldots, g_m]^\top$, then the (i, j)-th entry of this Jacobian matrix is $\frac{\partial f_i}{\partial g_j}$.

3.3 Taylor's Theorem

Taylor's theorem gives an approximation of a k-times differentiable function f around a given point.

Scalar functions We first state the scalar version where the function f is defined over $\mathbb{R}$.

Theorem 2 (Taylor's theorem)**.** *Let $k \geq 1$ be an integer and let the function $f : \mathbb{R} \to \mathbb{R}$ be k times differentiable at the point $a \in \mathbb{R}$. Then there exists a function $h : \mathbb{R} \to \mathbb{R}$ such that*

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + h_k(x)(x-a)^k, \tag{12}$$

and

$$\lim_{x \to a} h_k(x) = 0. \tag{13}$$

Proof. We prove it for $k = 3$ as the generalization follows easily.

Note that we assumed that the function f is k-times differentiable. We can therefore repeatedly apply the fundamental theorem of calculus as follows:

$$\begin{aligned}
f(x) &= f(a) + (f(x) - f(a)) = f(a) + \int_a^x f'(x_1)dx_1 \\
&= f(a) + \int_a^x f'(a) + (f'(x_1) - f'(a))dx_1 \\
&= f(a) + f'(a)(x-a) + \int_a^x (f'(x_1) - f'(a))dx_1 \\
&= f(a) + f'(a)(x-a) + \int_a^x \int_a^{x_1} f^{(2)}(x_2)dx_2dx_1 \\
&= \dots \\
&= f(a) + f'(a)(x-a) + \int_a^x \int_a^{x_1} f^{(2)}(a) + (f^{(2)}(x_2) - f^{(2)}(a))dx_2dx_1 \\
&= f(a) + f'(a)(x-a) + f^{(2)}(a)(x-a)^2 + \int_a^x \int_a^{x_1} \int_a^{x_2} f^{(3)}(x_3)dx_3dx_2dx_1.
\end{aligned}$$

Recall $f^{(2)}$ is continuous, therefore $|f^{(3)}(x)| \leq M \ \forall x$. The remainder can therefore be bounded as follows:

$$\begin{aligned}
\left| \int_a^x \int_a^{x_1} \int_a^{x_2} f^{(3)}(x_3)\ dx_3dx_2dx_1 \right| &\leq \int_a^x \int_a^{x_1} \int_a^{x_2} |f^{(3)}(x_3)|\ dx_3dx_2dx_1 \\
&\leq M \int_a^x \int_a^{x_1} \int_a^{x_2} 1\ dx_3dx_2dx_1 \\
&= M\frac{(x-a)^3}{3!}
\end{aligned}$$

□

By choosing $k = 1$ in Theorem 2, we recover the mean value theorem stated below.

Theorem 3 (Mean value theorem (scalar version))**.** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function on the closed interval $[a, b]$, and differentiable on the open interval (a, b), where $a < b$. Then there exists some $c \in (a, b)$ such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a} \tag{14}$$

Multivariable functions An extension of the previous result to a multivariate function $f : \mathbb{R}^d \to \mathbb{R}$ can be stated in two ways, either as

$$f(\mathbf{x} + \mathbf{y}) - f(\mathbf{x}) = \nabla f(\mathbf{x} + c\mathbf{y})^\top \mathbf{y}, \text{ for some } c \in (0, 1), \tag{15}$$

or in an integral form, as stated in the theorem below.

Theorem 4 (Mean value theorem (multivariable version))**.** *Given a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and any vector $\mathbf{y} \in \mathbb{R}^d$, we have that*

$$f(\mathbf{x} + \mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f(\mathbf{x} + t\mathbf{y})^\top \mathbf{y}\ dt. \tag{16}$$

Proof. Define $g : [0, 1] \to \mathbb{R}$ as $g(t) = f(\mathbf{x} + t\mathbf{y})$. Then by applying the fundamental theorem of calculus to g, we get

$$f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}) = g(1) - g(0) = \int_0^1 g'(t)dt.$$

Let $\mathbf{u}(t) = \mathbf{x} + t\mathbf{y}$ with entries $u_i(t)$. By the multivariate chain rule,

$$g'(t) = \sum_{i=1}^{d} \frac{\partial f(\mathbf{x}+t\mathbf{y})}{\partial u_i} \cdot \frac{\partial u_i}{\partial t} = \sum_{i=1}^{d} \frac{\partial f(\mathbf{x}+t\mathbf{y})}{\partial u_i} \cdot y_i = \nabla f(\mathbf{x}+t\mathbf{y})^\top \mathbf{y}.$$

Therefore

$$f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f(\mathbf{x}+t\mathbf{y})^\top \mathbf{y}\ dt.$$

□

Exercise Try to prove Eq. (15) by modifying the proof of Theorem 4.

High-order variant A similar expression to Eq. (16) can be stated for twice differentiable functions:

$$f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{y} + \frac{1}{2}\int_0^1 \mathbf{y}^\top \nabla^2 f(\mathbf{x}+t\mathbf{y})\mathbf{y}\ dt. \tag{17}$$

Gradient Note that the mean value theorem also applies to the gradient ∇f for functions f that are twice differentiable. It of course also applies to higher-order derivatives if they exist. For the first derivative, we simply get:

$$\nabla f(\mathbf{x}+\mathbf{y}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x}+t\mathbf{y})^\top \mathbf{y}\ dt. \tag{18}$$

3.4 Calculating gradients

Example: Gradient of a Multivariable Function

Consider the function $f(\mathbf{x}) = \mathbf{A}\mathbf{x} \in \mathbb{R}^p$, where $\mathbf{x} \in \mathbb{R}^d, \mathbf{A} \in \mathbb{R}^{p\times d}$. Our goal is to compute the gradient $\frac{df}{d\mathbf{x}}$.
First, note that the dimension of the gradient $\frac{df}{d\mathbf{x}}$ is $\mathbb{R}^{p\times d}$. Let's compute the partial derivative of f w.r.t. a single x_j. We have

$$f_i(\mathbf{x}) = \sum_{j=1}^{d} A_{ij}x_j \implies \frac{\partial f_i}{\partial x_j} = A_{ij}.$$

Collecting all the partial derivatives in the Jacobian, we obtain the following expression for the gradient:

$$\frac{df}{d\mathbf{x}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_d} \\ & \dots & \\ \frac{\partial f_p}{\partial x_1} & \dots & \frac{\partial f_p}{\partial x_d} \end{pmatrix} = \begin{pmatrix} A_{11} & \dots & A_{1d} \\ & \dots & \\ A_{p1} & \dots & A_{pd} \end{pmatrix} = \mathbf{A} \in \mathbb{R}^{p\times d}.$$

Composition of functions and chain rule

When considering compositions of functions of vectors or matrices, one often requires the use of the chain rule to calculate gradients. For instance, assume we want to calculate a loss $L(\mathbf{g}(\mathbf{x})) := \|\mathbf{g}(\mathbf{x})\|_2^2$, where $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^n$. By the chain rule, we have

$$\frac{\partial L}{\partial \mathbf{x}} = \frac{\partial L}{\partial \mathbf{g}} \cdot \frac{\partial \mathbf{g}}{\partial \mathbf{x}}. \tag{19}$$

As an example, let's consider the minimization of the least-square loss (with a linear model), which is defined as

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left[L(\mathbf{A}\mathbf{x}) := \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2\right], \tag{20}$$

where $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^n$ (thus $L : \mathbb{R}^n \to \mathbb{R}$). Sometimes, we will simply write $L(\mathbf{x}) := L(\mathbf{A}\mathbf{x})$.

In this case, we have $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{y} \in \mathbb{R}^{n\times 1}$ (column vector) and $L(\mathbf{g}) = \|\mathbf{g}\|_2^2$, thus $\frac{\partial L}{\partial \mathbf{g}} = 2\mathbf{g}^\top$. The other derivative is simply $\frac{\partial \mathbf{g}}{\partial \mathbf{x}} = \mathbf{A}$, therefore

$$\frac{\partial L}{\partial \mathbf{x}} = \frac{\partial L}{\partial \mathbf{g}} \cdot \frac{\partial \mathbf{g}}{\partial \mathbf{x}} = 2(\mathbf{A}\mathbf{x} - \mathbf{y})^\top \mathbf{A}. \tag{21}$$

Note that with the above notation, $\frac{\partial L}{\partial \mathbf{x}}$ is a row vector since $\frac{\partial L}{\partial \mathbf{x}} \in \mathbb{R}^{1\times d}$, but by convention, one might want a column vector, in which case we should transpose the result to obtain $2\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{y})$.

One alternative to the chain rule for this loss can be obtained by noting that

$$L(\mathbf{x}) := \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 = (\mathbf{A}\mathbf{x}-\mathbf{y})^\top(\mathbf{A}\mathbf{x}-\mathbf{y}), \tag{22}$$

and taking the derivative of the inner product (exercise: try to figure this out yourself).

3.5 Lipschitz property

Intuitively, a Lipschitz continuous function is limited in how fast it can change. See the formal definition below.

Definition 3. *The function $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz continuous if:*

$$|f(\mathbf{x}) - f(\mathbf{y})| \le L\,\|\mathbf{x}-\mathbf{y}\| \quad \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^d. \tag{23}$$

The following lemma gives a characterization of Lipschitz continuity using the gradient.

Lemma 5. *A function $f(\mathbf{x})$ is L-Lipschitz continuous if its gradient is bounded by L.*

Proof. By the mean value theorem, we have :

$$f(\mathbf{x}) - f(\mathbf{y}) = \nabla f(\mathbf{y} + c(\mathbf{x}-\mathbf{y}))^\top(\mathbf{x}-\mathbf{y})$$

Using the Cauchy-Schwarz inequality (see Proposition 1), we conclude that

$$|f(\mathbf{x}) - f(\mathbf{y})| \le \sup_{\mathbf{z}\in\mathbb{R}^d} \|\nabla f(\mathbf{z})\|\|\mathbf{x}-\mathbf{y}\| \le L\|\mathbf{x}-\mathbf{y}\|.$$

□

3.6 Smoothness

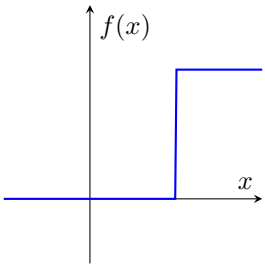
We say that a differentiable function is smooth if its gradient is Lipschitz continuous. We formalize this below.

Definition 4. *The function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be smooth if it is differentiable and it has L-Lipschitz-continuous gradient, i.e. if*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\,\|\mathbf{x}-\mathbf{y}\|\,\forall \mathbf{x},\mathbf{y} \in \mathbb{R}^d. \tag{24}$$

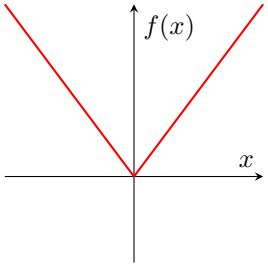
Figure 2 shows some examples of continuous and differentiable functions. For instance, the function $f(x) = x^2$ is smooth with constant $L = 2$ since:

$$|f'(x) - f'(y)| = 2|x-y|.$$



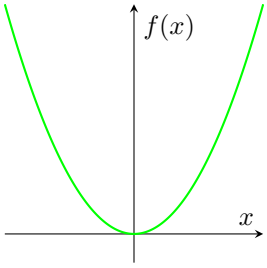
(a) $f(x) = \begin{cases} 0 & x < 1 \\ 1 & x \ge 1 \end{cases}$

Continuous: No
Differentiable: No



(b) $f(x) = |x|$

Continuous: Yes
Differentiable: No



(c) $f(x) = x^2$

Continuous: Yes
Differentiable: Yes

Figure 2: Examples of functions illustrating their continuity and differentiability properties.

4 Linear Algebra

4.1 Eigenvalues

We recall the definition of eigenvalues and eigenvectors in a real-vector space (this definition can be generalized to the complex domain but we will mostly deal with reals in this lecture). Consider a square matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ (recall that eigenvalues are not defined for rectangular matrices, for that we need the concept of singular values). Then $\lambda \in \mathbb{R}$ is an eigenvalue of $\mathbf{A}$ and $\mathbf{v} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$ is the corresponding eigenvector of $\mathbf{A}$ if the following holds:

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}. \tag{25}$$

Note that this relation is also equivalent to $(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = 0$ or $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$. The polynomial $\det(\mathbf{A} - \lambda \mathbf{I})$ plays an important role in linear algebra. For instance, the number of distinct eigenvalues of $\mathbf{A}$ is equal to the multiplicity of λ as a root of this polynomial.

In the following, we will denote the eigenvalues of $\mathbf{A}$ by λ_i and assume there are sorted as follows:

$$\lambda_1(\mathbf{A}) \geq \ldots \geq \lambda_n(\mathbf{A}).$$

Trace of a matrix The trace of a square matrix $\mathbf{A}$, denoted as $\mathrm{tr}(\mathbf{A})$, is the sum of its diagonal elements. If $\mathbf{A}$ is an $n \times n$ matrix given by:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

then the trace of $\mathbf{A}$ is defined as:

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

The trace of a matrix is also equal to the sum of its eigenvalues. If $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{A}$, then:

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i.$$

Below are some important properties of the trace we will be using later on:

- **Linearity**: For any two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$, and any scalars α and β,
$$\mathrm{tr}(\alpha\mathbf{A} + \beta\mathbf{B}) = \alpha\,\mathrm{tr}(\mathbf{A}) + \beta\,\mathrm{tr}(\mathbf{B})$$
- **Cyclic property**: For any $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$,
$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$$
- **Trace of a transpose**: For any $n \times n$ matrix $\mathbf{A}$,
$$\mathrm{tr}(\mathbf{A}^\top) = \mathrm{tr}(\mathbf{A})$$

Loewner order Recall that a matrix $\mathbf{A}$ is positive semi-definite (PSD) if $\mathbf{u}^\top \mathbf{A}\mathbf{u} \geq 0$ for all $\mathbf{u} \in \mathbb{R}^d$, or equivalently all the eigenvalues λ_i of $\mathbf{A}$ are non-negative, i.e. $\lambda_i(\mathbf{A}) \geq 0 \;\; \forall\, i = 1, \ldots d$. We will use the notation $\mathbf{A} \succcurlyeq 0$ to denote that $\mathbf{A}$ is PSD.

Definition 5 (Loewner order)**.** *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two symmetric matrices. We say that* $\mathbf{A} \succcurlyeq \mathbf{B}$ *if* $\mathbf{A} - \mathbf{B}$ *is positive semi-definite.*

Note that $\succcurlyeq$ is only a partial order (instead of a total order).

4.2 Singular values

The main object of interest will now be a rectangular $m \times n$ matrix $\mathbf{A}$ with real entries. We denote by $\sigma_i(\mathbf{A})$ the i-th singular value of $\mathbf{A}$ which is equal to

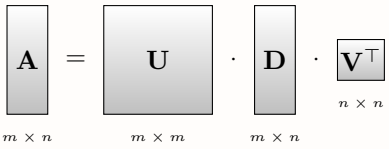
$$\sigma_i(\mathbf{A}) = \sqrt{\lambda_i(\mathbf{A}\mathbf{A}^\top)} = \sqrt{\lambda_i(\mathbf{A}^\top\mathbf{A})}. \tag{26}$$

The number of nonzero singular values of $\mathbf{A}$ equals to $\mathrm{rank}(\mathbf{A}) \leq \min(m, n)$, where we recall that the rank of a matrix is the maximum number of linearly independent rows or columns in the matrix.

We also note that given a matrix $\mathbf{A}$, the largest singular value $\sigma_1(\mathbf{A})$ is equal to the operator norm of $\mathbf{A}$.

Singular Value Decomposition (SVD) The Singular Value Decomposition (SVD) is a widely-used technique to decompose a matrix $\mathbf{A}$, and to expose some of its properties.

Theorem 6 (SVD)**.** *Every real (or complex) matrix* $\mathbf{A}$ *can be decomposed into*



where $\mathbf{U}$*,* $\mathbf{V}$ *orthogonal (or unitary),* $\mathbf{D}$ *diagonal. More precisely:*

- $\mathbf{U}$ *is an* m *by* m *orthogonal matrix,* $\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$*,*
- $\mathbf{D}$ *is an* m *by* n *diagonal matrix, padded with* $\max\{m,n\} - \min\{m,n\}$ *zero rows or columns,*
- $\mathbf{V}$ *is an* n *by* n *orthogonal matrix,* $\mathbf{V}^\top\mathbf{V} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}$*.*

Proof. Omitted. See standard linear algebra textbook, e.g. (Golub and Van Loan, 2013). □

We here recall the definition of orthogonal vectors and matrices for the convenience of the reader.

Definition 6 (Orthogonal vectors)**.** *A set of vectors* $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$ *in an inner product space are orthogonal if all pairwise inner products are zero, i.e.*

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0, \quad (\forall i \neq j).$$

Definition 7 (Orthogonal matrix)**.** *An orthogonal matrix* $\mathbf{A}$ *is a square matrix with real entries whose columns and rows are orthogonal unit vectors (i.e. orthonormal vectors):*

$$\langle \mathbf{a}_i, \mathbf{a}_j \rangle = \delta_{ij} \quad (\forall i, j).$$

Equivalently, $\mathbf{A}\mathbf{A}^\top = \mathbf{A}^\top\mathbf{A} = \mathbf{I}$*.*

Theorem 7. *The inverse of an orthogonal matrix is its transpose.*

Proof. It follows directly from the above equations as

$$\mathbf{A}^\top\mathbf{A} = \mathbf{I} \quad \Longrightarrow \quad \mathbf{A}^\top = \mathbf{A}^{-1}.$$

□

SVD: Singular Values The elements on the diagonal of $\mathbf{D}$ are the singular values and will be denoted by $\sigma_i := d_{ii}$ $(i = 1, \ldots, \min\{m,n\})$, i.e.

$$\mathbf{D} = \operatorname{diag}\left(\sigma_1, \ldots, \sigma_{\min\{m,n\}}\right).$$

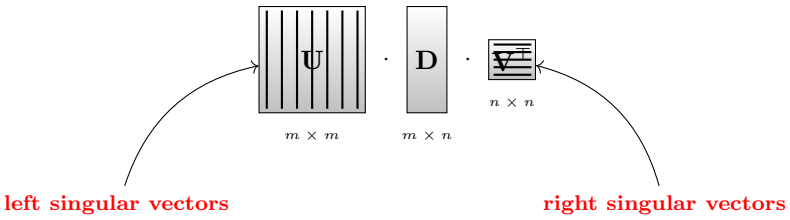
By convention, we typically order the singular values in decreasing order:

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min\{m,n\}} \geq 0$$

Also note the non-negativity of singular values. Finally, an important notion related to singular values is the rank of the matrix $\mathbf{A}$ defined as

$$\sigma_i = d_{ii} = 0, \quad (\forall i > \operatorname{rank}(\mathbf{A})).$$

SVD: Singular Vectors The columns of $\mathbf{U}$ (denoted by $\mathbf{u}_i \in \mathbb{R}^m$) are called the left singular vectors and they form an orthonormal basis for columns space of $\mathbf{A}$. Similarly, the rows of $\mathbf{V}^\top$ = columns of $\mathbf{V}$ (denoted by $\mathbf{v}_i \in \mathbb{R}^n$) are called the right singular vectors and form an orthonormal basis for the row space of $\mathbf{A}$.



The left and right singular vectors $\mathbf{u}_i \in \mathbb{R}^m$ and $\mathbf{v}_i \in \mathbb{R}^n$ are also the eigenvectors of $\mathbf{A}\mathbf{A}^\top$ and $\mathbf{A}^\top\mathbf{A}$ respectively.

We note that the SVD decomposition is only unique up to rotation and degeneracies (i.e. σ_i with two or more linearly independent left (or right) singular vectors).

SVD as sum of rank-1 matrices Let the columns of $\mathbf{U}$ be denoted by $\mathbf{u}_i$ and the columns of $\mathbf{V}$ be denoted by $\mathbf{v}_i$. Then by multiplying the SVD equation by $\mathbf{V}$ one gets

$$\mathbf{AV} = \mathbf{UD} \quad \Longleftrightarrow \quad \mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i \quad (\forall i \leq \min\{m, n\})$$

Similarly

$$\mathbf{A}^\top\mathbf{U} = \mathbf{V}\mathbf{D}^\top \iff \mathbf{A}^\top\mathbf{u}_i = \sigma_i \mathbf{v}_i \quad (\forall i \leq \min\{m, n\})$$

Note that for the special case of $m = n$ and $\mathbf{U} = \mathbf{V}$ (i.e. $\mathbf{A}$ symmetric) we retrieve the eigendecomposition. In this case, the vectors $\mathbf{u}_i = \mathbf{v}_i$ are just the eigenvectors.

Based on the above equations, we observe that we can write the SVD decomposition of a matrix as a sum of rank-1 matrices:

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \text{where } r = \text{rank}(\mathbf{A}), \tag{27}$$

where $\mathbf{u}_i\mathbf{v}_i^\top$ is an outer product between two vectors and therefore has rank 1 (exercise: try to prove this claim yourself).

Interpretation As shown in Figure 3, the SVD gives us a way to decompose a linear map as three subsequent operations: rotation, axis scaling, and another rotation

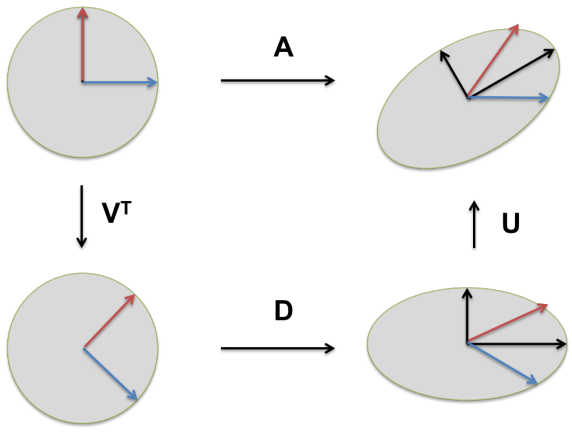


Figure 3: Illustration of the SVD transformation where: **Top:** The effect of $\mathbf{A}$ on the unit disc D and the canonical unit vectors $\mathbf{e}_1$ and $\mathbf{e}_2$. **Left:** The effect of $\mathbf{V}^\top$ is a rotation on D, $\mathbf{e}_1$, and $\mathbf{e}_2$. **Bottom:** The effect of $\mathbf{D}$, scaling horizontally by the singular value σ_1 and vertically by σ_2. **Right:** The effect of $\mathbf{U}$, another rotation. Source: Wikipedia.

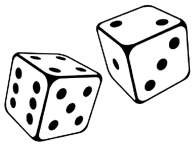
Relation between singular values and eigenvalues For symmetric matrices, the eigenvalues and singular values are closely related. A nonnegative eigenvalue, $\lambda \geq 0$, is also a singular value, $\sigma = \lambda$. The corresponding vectors are equal to each other, $\mathbf{u} = \mathbf{v}$. A negative eigenvalue, $\lambda < 0$, must reverse its sign to become a singular value, $\sigma = |\lambda|$. One of the corresponding singular vectors is the negative of the other, $\mathbf{u} = -\mathbf{v}$. So in general, if $\mathbf{A}$ is a symmetric matrix then the singular values of $\mathbf{A}$ are the absolute values of the eigenvalues λ_i of $\mathbf{A}$: $\sigma_i(\mathbf{A}) = |\lambda_i(\mathbf{A})|$.

5 Probability Theory

We first recall the definition of a probability space.

Definition 8 (Probability space)**.** *A probability space W is a unique triple $W = \{\Omega, \mathcal{F}, \Pr\}$, where Ω is its sample space , $\mathcal{F}$ its σ-algebra of events, and* Pr *its probability measure.*

Sample space The sample space Ω is the set of all possible samples or elementary events ω. Take for example the case where we throw a die once and define the random Variable X (we will later give a formal definition of this concept) as "the score shown on the top face". This random variable X can take the values 1, 2, 3, 4, 5 or 6. Therefore, the sample space is $\Omega := \{1, 2, 3, 4, 5, 6\}$. Let us list a few more examples of sample spaces:



- Toss of a coin (with head H and tail T): $\Omega = \{H, T\}$
- Two tosses of a coin: $\Omega = \{HH, HT, TH, TT\}$
- The positive integers: $\Omega = 1, 2, 3, \ldots$.

σ-algebra A σ-algebra is a collection of subsets of the sample space that satisfies certain properties. The σ-algebra represents the collection of events for which we can assign probabilities. It provides a structure to define and manipulate sets of outcomes or events.

Formally, the σ-algebra $\mathcal{F}$ is the set of all of the *considered* events A, i.e., subsets of Ω: $\mathcal{F} = \{A | A \subseteq \Omega, A \in \mathcal{F}\}$. It also has to satisfy the following properties:

1. $\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow \Omega \backslash A \in \mathcal{F}$ (closed under complementation)
3. $A_1, A_2, \cdots \in \mathcal{F} \Rightarrow \cup_n A_n \in \mathcal{F}$ (closed under countable unions)

Below are several examples of σ-algebra:

- **The trivial σ-algebra:**
  The smallest σ-algebra on any set X is the trivial σ-algebra, which contains only the empty set and the set itself.
  $$\mathcal{A} = \{\emptyset, X\}$$
- **The Power set σ-algebra:**
  The largest σ-algebra on a set X is the power set of X, which contains all subsets of X.
  $$\mathcal{A} = \mathcal{P}(X)$$
- **The σ-Algebra generated by a partition:**
  If $X = \{1, 2, 3, 4\}$ and we consider the partition $\{\{1, 2\}, \{3, 4\}\}$, the σ-algebra generated by this partition is:
  $$\mathcal{A} = \{\emptyset, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$$
- **The Borel σ-algebra on $\mathbb{R}$:**
  The Borel σ-algebra on the real numbers $\mathbb{R}$, denoted by $\mathcal{B}(\mathbb{R})$, is generated by all open intervals $(a, b) \subset \mathbb{R}$. This σ-algebra contains all Borel sets, which are the sets that can be formed from open intervals using countable unions, countable intersections, and relative complements.
- **σ-Algebra on a finite set:**
  If $X = \{a, b, c\}$, one possible σ-algebra on X could be:
  $$\mathcal{A} = \{\emptyset, \{a\}, \{b, c\}, \{a, b, c\}\}$$
- **The σ-Algebra generated by a single set:**
  Let X be a set and $A \subseteq X$. The σ-algebra generated by A is:
  $$\mathcal{A} = \{\emptyset, A, A^c, X\}$$

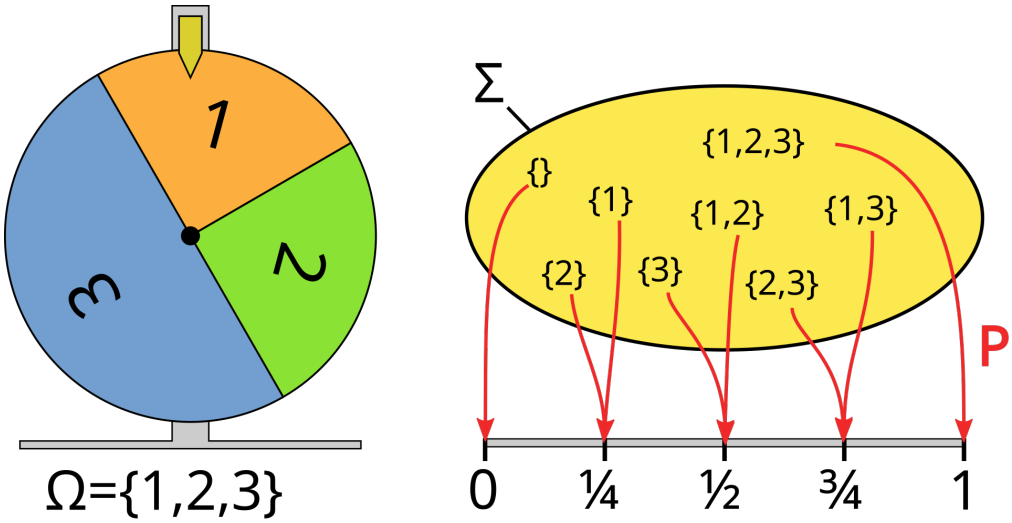


Figure 4: A probability measure mapping the σ-algebra for 2^3 events to the unit interval. Source: Wikipedia.

Probability measure A function $\Pr : \mathcal{F} \to [0, 1]$ is called a probability measure if it satisfies the following three properties:

1. Non-negativity: For every $A \in \mathcal{F}$, $\Pr(A) \geq 0$.
2. Normalization: $\Pr(\Omega) = 1$, where Ω is the entire sample space.
3. σ-additivity: For any countable collection $\{A_n\}$ of pairwise disjoint sets in $\mathcal{F}$ (i.e., $A_i \cap A_j = \varnothing$ for $i \neq j$), the probability of the union is equal to the sum of the probabilities of the individual sets:

$$\Pr\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \Pr(A_n).$$

An illustration of a probability measure is shown in Figure 4.

Random variable A random variable is similar to a variable in mathematics, but instead of taking on just one value, it can take on a range of possible values, each with a certain probability. Below, we state a more formal definition of a random variable.

Definition 9 (Random variable)**.** *Given a probability space* $(\Omega, \mathcal{F}, \Pr)$*, a random variable is a function from* Ω *to a measurable space* E [a]*. This function must satisfy the constraint* $\{\omega : X(\omega) \in B\} \in \mathcal{F}$ *for any Borel set* $B \subset E$[b]*, i.e. it is a valid event* $\forall B \in E$*. Each random variable has a distribution* $F_X(x) = \Pr(X \leq x)$*.*

[a] A measurable space is a set equipped with a collection of subsets that are designated as measurable.

[b] In simple terms, a Borel set is any set that you can create starting from open sets and applying these operations a finite or countable number of times

More formal definition of a Borel set A Borel set is a special type of set that is defined using the concept of open sets in a topological space. The collection of Borel sets includes all open sets and is closed under operations like countable union, countable intersection, and relative complement. For example, consider the real number line. An open interval like $(0, 1)$ is a Borel set because it is an open set. If you take the union of open intervals $(0, 1)$ and $(2, 3)$, you get another Borel set, $(0, 1) \cup (2, 3)$. Even if you take more complex combinations like countable unions or intersections of such intervals, the resulting sets are still Borel sets. So, any set that can be built from open intervals on the real line using operations like union, intersection, and complement will be a Borel set.

Given a set $S \subseteq E$, the probability of a random variable is defined as

$$\Pr(X \in S) = \Pr(\{\omega \in \Omega | X(\omega) \in S\}). \tag{28}$$

If X maps onto a finite or countable set, it is *discrete* and has a probability mass function (PMF) where $p_X(x) = \Pr(X = x)$.
If $dF_X(x)/dx$ exists and is finite for all x, then X is continuous and has a density $f_X(x) = dF_X(x)/dx$.

Example of random variables:
a) Throw two dices and take $\Omega = \{1,2,3,4,5,6\}^2$ and $\mathcal{F} = P(\Omega)$ where P is the power set, i.e. the set of all subsets of Ω (thus $\mathcal{F} = \{(1,1),(1,2),\ldots\}$). Then $X : \Omega \to \mathbb{R}$ defined as $(\omega_1, \omega_2) \to \omega_1 + \omega_2$ (i.e. sum the numbers on each die) is indeed a random variable.
b) Throw two dices and take $\Omega = \{1,2,3,4,5,6\}^2$ and $\mathcal{F} = \{\emptyset, \Omega\}$. Then X as defined in a) is not a random variable. For instance $X^{-1}(\{2\}) = \{(1,1)\} \notin \mathcal{F}$.

Continuous vs discrete probability spaces The following table compares the case where the probability space is continuous and discrete (note that one can also mix them but we won't be discussing this case in these notes).

Characteristic	**Discrete Probability Space**	**Continuous Probability Space**
Definition	Consists of a sample space Ω, a σ-algebra $\mathcal{F}$, and a probability mass function (PMF) $\Pr(X = x)$	Consists of a sample space Ω, a σ-algebra $\mathcal{F}$, and a probability density function (PDF) $f(x)$
Sample Space	Countable and finite or countably infinite, e.g. $\Omega = \{1,2,3,4,5,6\}$ (for a six-sided die)	Uncountably infinite, e.g. $\Omega = [0,1]$
σ-algebra	Typically the power set $\mathcal{F} = P(\Omega)$	Typically the Borel set formed by Ω
Probability Function	PMF $\Pr(X = x)$ for each x in Ω	PDF $f(x)$ such that $\Pr(a \le X \le b) = \int_a^b f(x)\,dx$
Probability Assignment	$\Pr(X = x) \ge 0$ for all x and $\sum_{x\in\Omega} \Pr(X = x) = 1$	$f(x) \ge 0$ for all x and $\int_{-\infty}^{\infty} f(x)\,dx = 1$
Probability at a Point	$\Pr(X = x) > 0$ for some x	$\Pr(X = x) = 0$ for all x
Sum/Integral of Probabilities	$\sum_{x\in\Omega} \Pr(X = x) = 1$	$\int_{-\infty}^{\infty} f(x)\,dx = 1$
Random Variable	X takes values in Ω with probabilities assigned by the PMF	X takes values in Ω with probabilities assigned by the PDF
Example	Throwing a six-sided die: $\Pr(X = 1) = \frac{1}{6}, \Pr(X = 2) = \frac{1}{6}, \ldots, \Pr(X = 6) = \frac{1}{6}$	Choosing a point in $[0,1]$: $f(x) = 1$ for $0 \le x \le 1$, and $f(x) = 0$ elsewhere

Table 1: Differences Between Discrete and Continuous Probability Spaces

Probability density function (PDF) and probability mass function (PMF) A probability density function is most commonly associated with absolutely continuous univariate distributions. A random variable X has density f_X, where f_X is a non-negative Lebesgue-integrable function, if:

$$\Pr[a \le X \le b] = \int_a^b f_X(x)\,dx. \tag{29}$$

Hence, if F_X is the cumulative distribution function of X, then:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du, \tag{30}$$

and (if f_X is continuous at x)

$$f_X(x) = \frac{d}{dx} F_X(x). \tag{31}$$

Intuitively, one can think of $f_X(x)\,dx$ as being the probability of X falling within the infinitesimal interval$[x, x + dx]$.

Similar properties hold for discrete probability spaces where the density function $f_X(x)$ is replaced by a discrete function $p : \mathcal{F} \to [0,1]$ defined by $p_X(x) = \Pr(X = x)$. Since the function p is defined over a discrete set, the integral is replaced by a sum, therefore for a given set A, we have

$$\Pr(A) = \sum_{\omega\in A} p_X(\omega).$$

Expectation If a random variable X has a continuous density $f_X(x)$, then its expectation is given by

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

For a discrete random variable with a finite list $x_1, \ldots, x_k$ of possible outcomes each of which (respectively) has probability $p_1, \ldots, p_k$ of occurring, the expectation can be written as

$$\mathbb{E}[X] = \sum_{i=1}^{k} x_i\, p_i.$$

Finally, the expected value of a non-negative random variable can be written as the integral of its tail distribution. This is because if X is a nonnegative real number, then

$$X = \int_0^{+\infty} [X \geq t] \mathrm{d}t,$$

where $[\cdot]$ is the indicator function.

Then one integrates both sides of the relevant identity with respect to the distribution $\Pr_X$ of X and one uses Fubini's theorem to change the order of the summation/integral and of the expectation:

$$\mathbb{E}[X] = \int_\Omega X \ \mathrm{d}\Pr = \int_\Omega \int_0^{+\infty} [X > t] \mathrm{d}t \ \mathrm{d}\Pr = \int_0^{+\infty} \int_\Omega [X > t] \mathrm{d}\mathbb{P} \ \mathrm{d}t$$

that is,

$$\mathbb{E}[X] = \int_0^{+\infty} \Pr(X \geq t) \ \mathrm{d}t.$$

6 Basic Topological Concepts

In this section, we explained some important concepts in topology that are needed for the course. We will only need a basic understanding of these concepts but we refer the interested reader to (Munkres, 2000) for more detailed explanations.

Let S be a subset of a topological space X, i.e. a set whose collection of open subsets satisfies certain conditions. We recall that topological spaces are abstractions of other spaces such as metric spaces, see illustration in Figure 5.

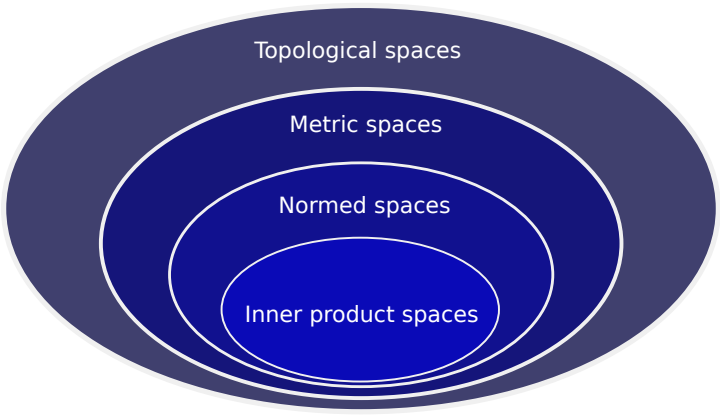


Figure 5: Source: Wikimedia Commons

Open set A set is open if it does *not* include any of its boundary points. Formally, the concept of open sets can be formalized with various degrees of generality. For example, in the case of metric spaces, we have the following definition.

Definition 10 (Open set)**.** *A subset U of a metric space (M, d) is called open if, given any point $x \in U$, there exists a real number $\epsilon > 0$ such that, given any point $y \in M$ with $d(x, y) < \epsilon$, y also belongs to U. Equivalently, U is open if every point in U has a neighborhood contained in U.*

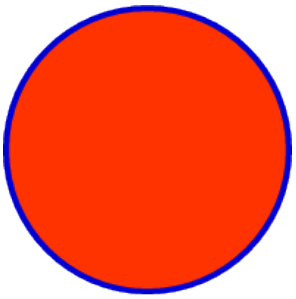


Figure 6: Example: The blue circle represents the set of points (x, y) satisfying $x^2 + y^2 = r^2$. The red disk represents the set of points (x, y) satisfying $x^2 + y^2 < r^2$. The red set is an open set, the blue set is its boundary set, and the union of the red and blue sets is a closed set.

Closed sets, Limit points and closure We start with the definition of a closed set.

Definition 11 (Closed set)**.** *A set is closed if its complement is open.*

Note that closed balls are closed sets (proof: show by contradiction that the complement of a closed ball is an open set, i.e. show it violates the triangle inequality).

Definition 12 (Limit point)**.** *We say that $p \in X$ is a limit point of S if every open neighborhood of p contains one point in S (other than p).*

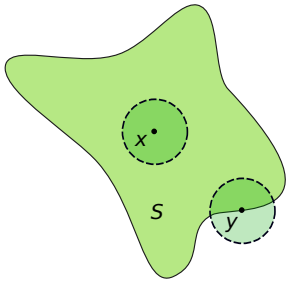


Figure 7: Example of limit points: Both x (interior point) and y (boundary point) are limit points of the set S since any neighborhood of these points contain a point in S.

One can give a characterization of a closed set using limit points, see the next proposition.

Proposition 8. *A closed set is a set that contains all of its limit points.*

Proof. Proof by contradiction: Assume a closed set S does not contain all its limit points, i.e. $\exists y \in S^{\complement}$ such that y is a limit point of S. Since $S^{\complement}$ is open and y is an interior point, then there exists a neighborhood of y, denoted by $\mathcal{B}(y, \epsilon)$ s.t. $\mathcal{B}(y, \epsilon) \subseteq S^{\complement}$ but this contradicts the fact that y is a limit point of S.
For the converse, show that any $y \in S^{\complement}$ is not a limit point of S therefore proving $S^{\complement}$ is open. □

The last proposition can also be stated as follows.

Proposition 9. *A set $A \subseteq X$ is closed if and only if for every convergent sequence $(a_n)_{n \in N} \subseteq A$, we have $\lim_{n \to \infty} a_n \in A$.*

Definition 13 (Closure)**.** *The closure of a set A is the union of all its limit points. It is usually denoted by $\bar{A}$ or $cl(A) = A \cup A'$, where A' is the set of all limit points.*

Definition 14 (Dense set)**.** *A subset A of a topological space X is called dense (in X) if every point $x \in X$ either belongs to A or is a limit point of A. Alternatively, A is dense if it has **non-empty intersection** with an arbitrary non-empty open subset $B \subset X$.*

A well-known example is the fact that the rationals are dense in the set of reals, which we formalize in the next theorem.

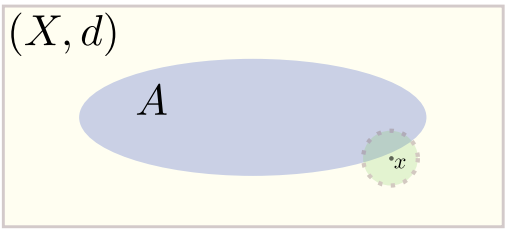


Figure 8: The set A shown in blue is dense in X if every $x \in X$ is a limit point of A.

Theorem 10 (Density theorem, $\mathbb{Q}$ is dense on $\mathbb{R}$)**.**

$$\forall a < b \in \mathbb{R}, \exists x \in \mathbb{Q} \text{ } s.t. \text{ } x \in (a, b), \text{ } i.e. \text{ } a < x < b$$

Compact sets. There are typically two characterizations of compact spaces, one in terms of open sets and another one in terms of convergent sequences. We start with the definition in terms of open sets.

Definition 15 (Compact set, definition 1)**.** *A topological space X is called compact if each of its open covers [a] has a finite subcover.*

[a] A cover of a set X is a collection of sets whose union includes X as a subset.

Explicitly, this means that for every arbitrary collection $\{U_\alpha\}_{\alpha \in A}$ of open subsets of X such that $X = \bigcup_{\alpha \in A} U_\alpha$, there is a **finite** subset J of A such that $X = \bigcup_{i \in J} U_i$. See illustration in Figure 9.

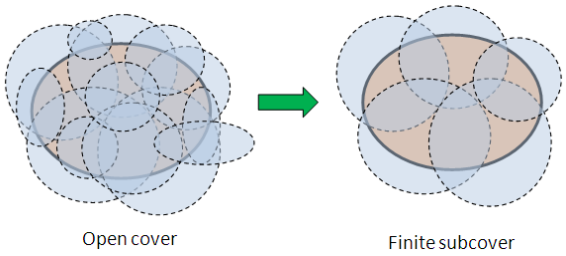


Figure 9: Source: `https://mathstrek.blog/`

Definition 16 (Compact set, definition 2)**.** *We call X a compact set if all sequences $(f_n)_{n \geq 1} \subset X$ have a convergent subsequence $(f_{n(k)})$ with limit point in X.*

Finally, we note that the following characterization of compact sets is often used in the literature: a subset of $\mathbb{R}^d$ is compact if it is closed and bounded (Heine-Borel theorem).

References

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

James R Munkres. Topology, 2000.

Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.