Andreas Krause, Fanny Yang

Introduction to Machine Learning

SPRING 2022

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Institute for Machine Learning
Department of Computer Science

Compilation date: **February 22, 2022**

Table of Notations

Abbreviations	
w.r.t.	with respect to
e.g.	for example
i.e.	that is
i.i.d.	independent and identically distributed
SVD	singular value decomposition
CDF	cumulative distribution function
PMF	probability mass function
PDF	probability density function
LLN	law of large numbers
CLT	central limit theorem
Linear Algebra	
$A, B, C, \ldots$	matrices
$\mathbb{R}^{m \times n}$	the set of m-by-n matrices
$\boldsymbol{x}, \boldsymbol{y}, \ldots$	vectors
x_i or $x[i]$	the i-th component of the vector $\boldsymbol{x}$
$\boldsymbol{x}_i$	the i-th vector in a collection of vectors
$\operatorname{span}(\mathcal{S})$	span of vectors in $\mathcal{S}$
$\operatorname{range}(A)$	range of the matrix A
$\ker(A)$	kernel of the matrix A
$\operatorname{rank}(A)$	rank of the matrix A
$\det(A)$	determinant of the square matrix A
$\mathbf{0}$	the zero vector
$\mathbf{0}_{m \times n}$	an $m \times n$ matrix filled with zeros
A^{-1}	inverse of the invertible matrix A
$A^\top$	transpose of A
$\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ or $\boldsymbol{u}^\top \boldsymbol{v}$	inner product of the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$
$\lVert \cdot \rVert_2$	Euclidean norm
$\lVert \cdot \rVert_p$	the p-norm
$\perp$	perpendicular
$\mathcal{S}^\perp$	orthogonal complement of $\mathcal{S}$
δ_{ij}	Kronecker delta function
$\sigma_i(A)$	the i-th largest singular value of A
$\lambda_i(A)$	the i-th largest eigenvalue of A

$x \mapsto f(x)$	description of a function that maps x to $f(x)$
I	Identity matrix
$A^\dagger$	Moore-Penrose pseudo-inverse
$\mathrm{tr}(A)$	trace of the square matrix A
Π_X	projection matrix onto range(X)

Analysis

$Df(\boldsymbol{x})$	derivative of f at the point $\boldsymbol{x}$
$Df(\boldsymbol{x})[\boldsymbol{v}]$	derivative of f at the point $\boldsymbol{x}$ applied to $\boldsymbol{v}$
$\frac{\partial f}{\partial x_i}$	partial derivative of f with respect to x_i
$\nabla f(\boldsymbol{x})$	gradient of f at the point $\boldsymbol{x}$
$D^2 f(\boldsymbol{x})$	Hessian of f at the point $\boldsymbol{x}$
$f \circ g$	composition of f and g
$o(\cdot)$	little-o notation
$\mathcal{O}(\cdot)$	big-O notation

Probability

Ω	sample space
$\mathcal{F}$	family of events
$\mathbb{P}$	probability function
$(\Omega, \mathcal{F}, \mathbb{P})$	probability space
X, Y	random variables
$\mathbb{P}_X$	distribution/law of the random variable X
F_X	cumulative distribution function of the random variable X
p_X	probability mass or probability density function of the random variable X
$\mathbb{E}[X]$	expected value of X
$\mathbb{E}[X \mid Y = y]$	conditional expectation of X given $Y = y$
$\mathbb{E}[X \mid Y]$	conditional expectation of X given Y
$\mathbb{E}_X[f(X, Y)]$	expectation of f w.r.t. the randomness of X
$\mathbb{E}_{X\|y}[f(X, Y)]$	expectation of f w.r.t. the conditional distribution of $X \mid Y = y$
$\mathrm{Var}(X)$	variance of X
$\mathrm{Cov}(X, Y)$	covariance of X and Y
$\mathcal{N}(\mu, \sigma)$	normal distribution with mean μ and variance σ^2
$\mathrm{Unif}(S)$	uniform distribution on the set S
$\mathrm{Exp}(\lambda)$	exponential distribution with parameter λ
$\mathrm{Ber}(q)$	Bernoulli distribution with parameter q
$\mathrm{Cat}(p_1, p_2, \ldots, p_k)$	categorical distribution with parameters $p_1, p_2, \ldots, p_k$
$\mathrm{Binom}(q, n)$	binomial distributions with parameters q and n
$\mathrm{Poisson}(\lambda)$	Poisson distribution with parameter λ

Contributors

This set of notes is the result of the efforts of the following contributors:

- The contents of the *Preliminaries* part is an accumulation of the "Math Recap" tutorials throughout the last years, and is written, expanded, and refined by Riccardo Zuliani, Tobias Wegel, and Mohammad Reza Karimi. All typos or mistakes can be blamed on the latter.

Contents

Part I

Preliminaries

1

Linear Algebra

In this chapter we review the most important concepts, ideas, definitions, and results of linear algebra that are needed in this course. We do not take an abstract standpoint, and only consider the Euclidean space $\mathbb{R}^n$ with its standard basis.

Roadmap

In Section 1.1 we recall the notion of linear subspace, linear independence, and span. In Section 1.2 we review some algebraic facts about matrices. Most importantly, we revisit matrix-vector multiplication, subspaces related to a matrix, and notions of rank and nullity. Section 1.3 starts with the additional structure of the Euclidean space: inner product. This is followed by notions of orthogonality and orthogonal matrices. The pinnacle of this section is the singular value decomposition, which gives a complete geometric characterization of "how a linear map works." After orthogonality, we build intuition on how to project a vector onto a given subspace in Section 1.4. We will first solve this problem for the case where the subspace is explained by a set of linearly independent vectors, and then give a complete solution for a general set of vectors in Section 1.4.1, where we introduce the notion of pseudo-inverse. We then review eigenvalues and eigenvectors in Section 1.5, which are vital to understand the dynamic behavior of a linear map, used in analyzing optimization algorithms. In Section 1.6, we construct the multivariate analogue of quadratic functions and describe the important notion of positive-definiteness. We will describe geometric objects such as ellipses (or hyperellipses) using these functions, and create the foundation for understanding the second order derivatives. Lastly, in Section 1.7 we introduce the notion of trace, which assigns a number to a square matrix, and is a useful tool both in derivations and understanding quadratic functions.

Learning Objectives

After reading this chapter you should know

- how to understand matrix-vector multiplication via linear combinations.
- what is rank, and how is it related to the kernel of a matrix.
- what is a p-norm and how inner product is related to different norms.
- how to check if a matrix is orthogonal.
- what is SVD, and how to interpret it geometrically.
- how to orthogonally project a vector onto a subspace spanned by a set of linearly independent vectors.

- what is pseudo-inverse, how to compute it based on SVD, and how to construct a projection matrix by it.
- what is eigen-decomposition and how it is related to SVD for symmteric matrices.
- what is a p.s.d. matrix, how to interpret positive-definiteness geometrically, and different ways to verify if a matrix is positive.
- how to create (hyper-)ellipses in multiple dimensions via p.s.d. matrices.
- what is trace, and how is it related to eigenvalues.
- the trace trick.

1.1 Vector Space Notions

Definition 1.1 (Linear subspace)**.** A nonempty set $\mathcal{V} \subseteq \mathbb{R}^n$ is a linear subspace of $\mathbb{R}^n$ if for any two vectors $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathcal{V}$ and two scalars $\alpha, \beta \in \mathbb{R}$, it holds that $\alpha\boldsymbol{u}_1 + \beta\boldsymbol{u}_2 \in \mathcal{V}$.

Definition 1.2 (Linear independence, Dimension)**.** A set of vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ in $\mathbb{R}^n$ are linearly independent if no trivial linear combination of them is the zero vector. That is, if

$$\sum_{i=1}^{k} c_i \boldsymbol{v}_i = \boldsymbol{0},$$

then $c_1 = \cdots = c_k = 0$. The *dimension* of a linear subspace $\mathcal{V}$ is the size of a maximal linearly independent subset of $\mathcal{V}$.

Definition 1.3 (Span)**.** The *span* of a set of vectors $\mathcal{S} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k\}$, with $\boldsymbol{a}_i \in \mathbb{R}^n$ for $i = 1, \ldots, n$, is the set of all possible linear combinations of them:

$$\operatorname{span}(\mathcal{S}) = \left\{ \sum_{i=1}^{k} c_i \boldsymbol{a}_i : c_i \in \mathbb{R}, \quad i = 1, \ldots, k \right\}.$$

Notice that the span of a set of vectors is a linear subspace, and, if $\boldsymbol{a}_i$ are linearly independent, their span has dimension k.

1.2 Matrix algebra

A matrix $A \in \mathbb{R}^{m \times n}$ defines a *linear map* between the Euclidean spaces $\mathbb{R}^n$ and $\mathbb{R}^m$, i.e., it maps a column vector $\boldsymbol{x} \in \mathbb{R}^n$ to the column vector $A\boldsymbol{x} \in \mathbb{R}^m$. We write x_i (or sometimes $x[i]$) to denote the i-th component of the vector $\boldsymbol{x}$.

An important observation about matrix-vector multiplication is that the vector $\boldsymbol{b} = A\boldsymbol{x}$ is a linear combination of the columns of A:

$$\boldsymbol{b} = A\boldsymbol{x} = \sum_{i=1}^{n} x_i \boldsymbol{a}_i,$$

where $\boldsymbol{a}_i \in \mathbb{R}^m$ denotes the i-th column of A. Similarly, we can understand matrix-matrix multiplication in this way. Let $B \in$

$\mathbb{R}^{n\times p}$. Then the matrix product $AB \in \mathbb{R}^{m\times p}$, consists of linear combinations of the columns of A:

$$\begin{bmatrix} & A & \end{bmatrix} \underbrace{\begin{bmatrix} | & & | \\ \boldsymbol{b}_1 & \cdots & \boldsymbol{b}_p \\ | & & | \end{bmatrix}}_{B} = \begin{bmatrix} | & & | \\ A\boldsymbol{b}_1 & \cdots & A\boldsymbol{b}_p \\ | & & | \end{bmatrix}.$$

Example 1 (Outer Product). Given two vectors $\boldsymbol{u} \in \mathbb{R}^m$ and $\boldsymbol{v} \in \mathbb{R}^n$, their outer product is the $m \times n$ matrix $\boldsymbol{u}\boldsymbol{v}^\top$, which contains repeated copies of $\boldsymbol{u}$ multiplied by the elements of $\boldsymbol{v}$:

$$\boldsymbol{u}\boldsymbol{v}^\top = \begin{bmatrix} | & & | \\ v_1\boldsymbol{u} & \dots & v_n\boldsymbol{u} \\ | & & | \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

Definition 1.4 (Image, Kernel, Rank). The *image* or *range* of a matrix A, denoted as $\text{range}(A)$, is the span of its columns. The *kernel* or *null space* of a matrix $A \in \mathbb{R}^{m\times n}$ is

$$\ker(A) = \{\boldsymbol{x} \in \mathbb{R}^n : A\boldsymbol{x} = \boldsymbol{0}\}.$$

The *rank* of a matrix is the dimension of its range.

Example 2 (Rank of some small matrices). Let the matrices A and B be defined as

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 \\ 0.5 & 1 \end{bmatrix}.$$

Then $\text{rank}(A) = 2$ and $\text{rank}(B) = 1$. For any two nonzero vectors, one can verify that their outer product is always a matrix of rank one.

The kernel and image are tied together via the following theorem:

Theorem 1.1 (Rank-Nullity). *Let $A \in \mathbb{R}^{m\times n}$, then*

$$\dim(\ker(A)) + \text{rank}(A) = n.$$

Definition 1.5 (Determinant). The determinant of an $n \times n$ matrix A is defined as the (signed) volume of the parallelepiped made out of the columns of A.[1]

If the columns of A are linearly dependent, the parallelepiped is degenerate and has no volume, hence, the determinant is zero.

We state the following facts without proof. The reader who is not familiar with these facts is urged to prove them, or consult any textbook on linear algebra, e.g., [TB97].

- $\text{rank}(A) = \text{rank}(A^\top)$. Hence, for a matrix $A \in \mathbb{R}^{m\times n}$, $\text{rank}(A) \leq \min\{m, n\}$. We say A has *full rank* if $\text{rank}(A) = \min\{m, n\}$.

[1] A computational definition of the determinant of a matrix is not needed in this course. However, it should be clear by this definition that

$$\det\left(\begin{bmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{bmatrix}\right) = a_1 a_2 \cdots a_n.$$

- The linear map defined by $A \in \mathbb{R}^{m\times n}$ is *injective* (i.e., $A\boldsymbol{x} = A\boldsymbol{y}$ only if $\boldsymbol{x} = \boldsymbol{y}$) if and only if $\ker(A) = \{\mathbf{0}\}$.
- The linear map defined by $A \in \mathbb{R}^{m\times n}$ is *surjective* (i.e., $\text{range}(A) = \mathbb{R}^m$) if and only if $\text{rank}(A) = m$.
- The linear map defined by A is *bijective* (i.e., it is surjective and injective) if and only if one of the following equivalent conditions are true:
  - A is square and $\ker(A) = \{\mathbf{0}\}$;
  - A is square and $\text{rank}(A) = m$;
  - A is square and $\det(A) \neq 0$.

Definition 1.6 (Invertible matrix)**.** A square matrix $A \in \mathbb{R}^{n\times n}$ is *invertible* (or *nonsingular*) if there exists a square matrix $B \in \mathbb{R}^{n\times n}$ such that

$$AB = BA = I.$$

Here, I is the identity matrix. If this is the case, then B is the (unique) *inverse* of A, and it is denoted by $B = A^{-1}$. Equivalently, A square matrix is invertible if and only if the linear map defined by it is bijective. A non-invertible matrix is said to be *singular*.

Bonus Material

Let $A \in \mathbb{R}^{n\times n}$ be a square invertible matrix and let $\boldsymbol{b} = (b_1, \ldots, b_n) \in \mathbb{R}^n$ be arbitrary. As A is invertible, $\boldsymbol{b}$ is in the range of A, and hence, can be expressed as a linear combination of columns of A:

$$\boldsymbol{b} = \sum_{i=1}^{n} c_i \boldsymbol{a}_i.$$

Hence, we can numerically represent the same vector by either b_i (in the Euclidean basis) or c_i (in the basis formed by columns of A). The coordinates b_i and c_i are related to each other via A and A^{-1}. That is, if b_i are given, we can obtain c_i by a matrix multiplication by A^{-1}:

$$\begin{bmatrix} b_1 \\ \vdots \\ c_n \end{bmatrix} = A^{-1} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix},$$

and, if c_i are given, we can get b_i by multiplying by A. Thus, A^{-1} can be viewed as a change of basis operator, converting the representation of a vector in Euclidean basis into the representation in the basis of columns of A.

1.3 Geometric Constructs

Definition 1.7 (Euclidean inner product and norm)**.** The *Euclidean*

inner product of two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ is defined as

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^\top \boldsymbol{v} = \sum_{i=1}^{n} u_i v_i,$$

where $u_i, v_i \in \mathbb{R}$ are the i-th component of $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively. The *Euclidean norm* of a vector $\boldsymbol{v} \in \mathbb{R}^n$ is defined as

$$\|\boldsymbol{v}\|_2 = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle}.$$

We will someties drop the subscript 2 when referring to the Euclidean norm and write $\|\cdot\|$ instead of $\|\cdot\|_2$.

The Euclidean inner product can be used to define the angle between vectors:

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \|\boldsymbol{v}\| \, \|\boldsymbol{u}\| \cos \angle(\boldsymbol{u}, \boldsymbol{v}),$$

where $\angle(\boldsymbol{u}, \boldsymbol{v})$ is the angle between $\boldsymbol{u}$ and $\boldsymbol{v}$.

Remark. Recall that in Section 1.2, we observed that matrix-vector multiplication can be seen as taking a linear combination. Here we present another point of view. Let $A \in \mathbb{R}^{m \times n}$ and $\boldsymbol{x} \in \mathbb{R}^n$, and denote by $\boldsymbol{a}^i \in \mathbb{R}^n$ the ith row of A. Then

$$A\boldsymbol{x} = \begin{bmatrix} \langle \boldsymbol{a}^1, \boldsymbol{x} \rangle \\ \vdots \\ \langle \boldsymbol{a}^m, \boldsymbol{x} \rangle \end{bmatrix}.$$

When facing with a matrix-vector product, sometimes this understanding is useful and sometimes the former.

There are norms other than the Euclidean norm,[2] that have different geometric meanings and are interesting for us in different occasions. A rich class of norms, that include the Euclidean norm as a special case, is the class of p-norms.

[2] In general, a *norm* is any function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ that satisfies these properties:

- $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0$ iff $\boldsymbol{x} = \boldsymbol{0}$.
- $\|\alpha \boldsymbol{x}\| = |\alpha| \, \|\boldsymbol{x}\|$ for all $\alpha \in \mathbb{R}$.
- $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

Definition 1.8 (p-norm)**.** Let $p \in [1, \infty)$ be a real number. The p-norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

where x_i is the i-th entry of $\boldsymbol{x}$. The *infinity norm* is also defined as

$$\|\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,n} |x_i|.$$

Figure 1.1 shows the *unit-norm balls* in $\mathbb{R}^2$, i.e., the sets $\{\boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x}\|_p \leq 1\}$, for different values of p.

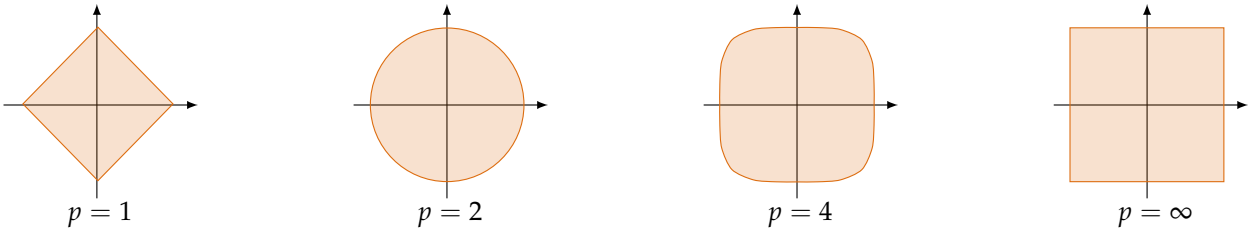

Figure 1.1: Unit-norm balls in $\mathbb{R}^2$ for different p-norms.

Theorem 1.2 (Hölder and Cauchy-Schwarz inequalities). *Let $p, q \in [1, \infty]$ with $1/p + 1/q = 1$. Then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have that*

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q . \tag{1.1}$$

For $q = p = 2$, this inequality is known as Cauchy-Schwarz inequality.

We now move back to the Euclidean space with its usual norm and inner product. First, we deal with the notion of orthogonality, which is central to the discussions that follows.

Definition 1.9 (Orthogonality). Two vectors $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^n$ are *orthogonal*, denoted by $\boldsymbol{v} \perp \boldsymbol{u}$, if $\langle \boldsymbol{v}, \boldsymbol{u} \rangle = 0$.

If two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ are orthogonal, then

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

The reader can confirm that this is indeed the *Pythagoras theorem*.

A subspace $\mathcal{S} \subset \mathbb{R}^n$ is orthogonal to subspace $\mathcal{R} \subset \mathbb{R}^n$ if every vector in $\mathcal{S}$ is orthogonal to every vector in $\mathcal{R}$.

Definition 1.10 (Orthogonal complement). Let $\mathcal{S}$ be a subspace of $\mathbb{R}^n$. The orthogonal complement of $\mathcal{S}$ is the set $\mathcal{S}^\perp$ whose elements are orthogonal to every element of $\mathcal{S}$:

$$\mathcal{S}^\perp = \{\boldsymbol{x} \in \mathbb{R}^n \mid \langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0, \text{ for all } \boldsymbol{y} \in \mathcal{S}\}.$$

Exercise 1.1. Prove that the orthogonal complement of any set is a linear subspace.

Definition 1.11 (Orthonormal basis). A set of vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m$ in $\mathbb{R}^n$ are *orthonormal* if they are pairwise orthogonal, i.e., $\langle \boldsymbol{q}_i, \boldsymbol{q}_j \rangle = 0$ for $i \neq j$, and have unit norm, i.e., $\|\boldsymbol{q}_i\| = 1$ for $i = 1, \ldots, m$. In a

more compact form,

$$\langle \boldsymbol{q}_i, \boldsymbol{q}_j \rangle = \delta_{ij} := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

A set of n orthonormal vectors in $\mathbb{R}^n$ is called an *orthonormal basis*.

If $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\}$ form an orthonormal basis of $\mathbb{R}^n$, then any vector $\boldsymbol{v} \in \mathbb{R}^n$ can be expressed as

$$\boldsymbol{v} = \sum_{i=1}^{n} \left(\boldsymbol{v}^\top \boldsymbol{q}_i \right) \boldsymbol{q}_i.$$

Remark. By changing the parenthesis, we see that the equation above can also be written as

$$\boldsymbol{v} = \sum_{i=1}^{n} \left(\boldsymbol{q}_i \boldsymbol{q}_i^\top \right) \boldsymbol{v}.$$

This new equation is very different from the previous one: each term in the sum is the application of the *matrix* $\boldsymbol{q}_i \boldsymbol{q}_i^\top$ to the vector $\boldsymbol{v}$. Recall that this matrix is the outer product of $\boldsymbol{q}_i$ with itself, and hence is a rank 1 matrix. We will see later that this matrix is the orthogonal projection matrix (see Definition 1.14) on the direction $\boldsymbol{q}_i$.

Definition 1.12 (Orthogonal matrix)**.** An *orthogonal matrix* U is a real square matrix whose columns are orthonormal. Equivalently, U is orthogonal iff $U^\top U = I$.

Exercise 1.2. Let U be an orthogonal matrix. Argue why $UU^\top = I$. Moreover, prove that an orthogonal matrix, preserves the Euclidean inner product and norm:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle U\boldsymbol{x}, U\boldsymbol{y} \rangle .$$

and

$$\|U\boldsymbol{x}\| = \|\boldsymbol{x}\| .$$

Exercise 1.3. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be vectors in $\mathbb{R}^m$. Consider the matrix $X \in \mathbb{R}^{m \times n}$ whose columns are $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

1. Represent the elements of the matrix $X^\top X$ using the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.
2. Let $n \leq m$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be orthonormal. What does the linear map $\boldsymbol{b} \mapsto XX^\top \boldsymbol{b}$ do?

Solution. 1. The entries of the square matrix $X^\top X \in \mathbb{R}^{n \times n}$ correspond to the inner products of all pairs of $\boldsymbol{x}_i$:

$$\left(X^\top X \right)_{ij} = \left\langle \boldsymbol{x}_i, \boldsymbol{x}_j \right\rangle .$$

2. The linear operator $XX^\top$ is an orthogonal projection matrix (see Definition 1.14) on the range of X:

$$XX^\top \boldsymbol{b} = X \begin{bmatrix} \langle \boldsymbol{x}_1, \boldsymbol{b} \rangle \\ \vdots \\ \langle \boldsymbol{x}_n, \boldsymbol{b} \rangle \end{bmatrix} = \sum_{i=1}^{n} \langle \boldsymbol{x}_i, \boldsymbol{b} \rangle \boldsymbol{x}_i.$$

If $m = n$ then $XX^\top \boldsymbol{b} = \boldsymbol{b}$. Figure 1.2 shows an example of an orthogonal projection from $\mathbb{R}^3$ to a 2-dimensional plane.

□

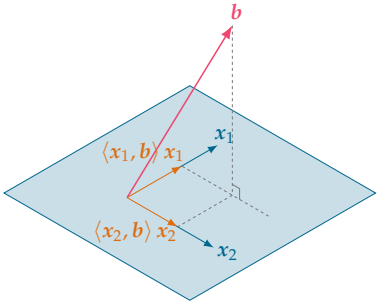

Figure 1.2: Projection of a vector $\boldsymbol{b} \in \mathbb{R}^3$ on the plane spanned by $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^3$.

Understanding what a matrix does as a linear map can be complicated. It is oftentimes useful to *decompose* the matrix into simpler building blocks. The singular value decomposition is such a tool that reveals a lot of geometric facts about the linear map.

Theorem 1.3 (Singular Value Decomposition). *Let $A \in \mathbb{R}^{m \times n}$. We can decompose A as the product $A = U\Sigma V^\top$, where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, and Σ is an $m \times n$ diagonal matrix with non-negative diagonal entries. This factorization is called a* singular value decomposition (SVD) *of A.*

The diagonal entries of Σ, denoted by $\sigma_i(A) := \Sigma_{ii}$, are called the *singular values* of A, columns of V are called the *right singular vectors* of A, and columns of U are called the *left singular vectors* of A. Often, we assume that the singular values are sorted decreasingly, that is, $\sigma_1 \geq \sigma_2 \geq \cdots$. See Figure 1.3 for a schematic description of SVD. Also, it is sometimes easier to consider a *reduced* SVD, see Figure 1.4.

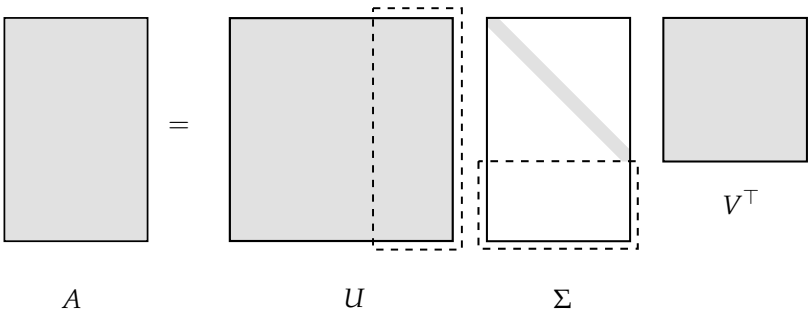

Figure 1.3: SVD of the matrix A, as defined in Theorem 1.3. Sometimes this is called the *full SVD*. The white parts are zeros, and shaded areas are the parts that can have nonzero numbers. The part of U that is marked with dashed line has no information, as it correspond to the dashed part of Σ, which is all zeros.

The singular value decomposition has the following properties:

- The rank of A is equal to the number of nonzero singular values.
- The left singular vectors of A that correspond to nonzero singular values form an orthonormal basis (see Definition 1.11) for range(A).
- The right singular vectors of A that correspond to zero singular values form an orthonormal basis for ker(A).

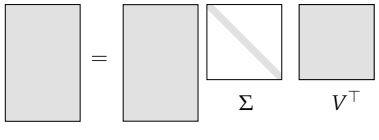

Figure 1.4: Another way to write down the SVD. In this way, Σ becomes a square matrix, but U will not be square. However, it still holds that the columns of U are orthonormal; $U^\top U = I$. This is called the *reduced SVD*.

The relation between left and right singular vectors is simple: $Av_i = \sigma_i u_i$. Moreover, it left as an exercise to see that the SVD of A can also be written as the sum

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top, \tag{1.2}$$

where r is the number of nonzero singular values of A ($= \text{rank}(A)$).

Bonus Material

Equation (1.2) shows how we can decompose X into a sum of r rank-1 matrices (recall that the outer product $u_i v_i^\top$ is of rank 1). Each matrix is determined by a pair of directions (i.e., u_i and v_i) and a magnitude σ_i. If we decide to truncate the sum at a value $k < r$, then we obtain a *rank-k approximation* of the matrix A

$$A \approx \sum_{i=1}^{k} \sigma_i u_i v_i^\top.$$

The error of this approximation is related to the magnitude of the remaining singular values σ_i for $i = k+1, \ldots, r$.

Remark. The SVD theorem also has a geometric interpretation:

> The image of the unit sphere under a linear transformation is always a hyperellipse.

A *hyperellipse* is generalization of a sphere, which can be obtained by stretching the unit sphere in $\mathbb{R}^m$ by some factors $\sigma_1, \ldots, \sigma_m$ is some orthogonal direction $u_1, \ldots, u_m \in \mathbb{R}^m$. If we choose $\|u_i\| = 1$, then the vectors $\sigma_i u_i$ are the *principal semiaxes* of the hyperellipse.

If we apply the linear mapping defined by $A \in \mathbb{R}^{m \times n}$ to the unit sphere of $\mathbb{R}^n$, the SVD theorem tells us that the result is a hyperellipse. Moreover, the left singular vectors u_i of A are the directions of the principal semiaxes of the hyperellipse and the associated singular values σ_i are the magnitudes of the corresponding semiaxes (see Figure 1.5 for an example where $n = m = 2$).

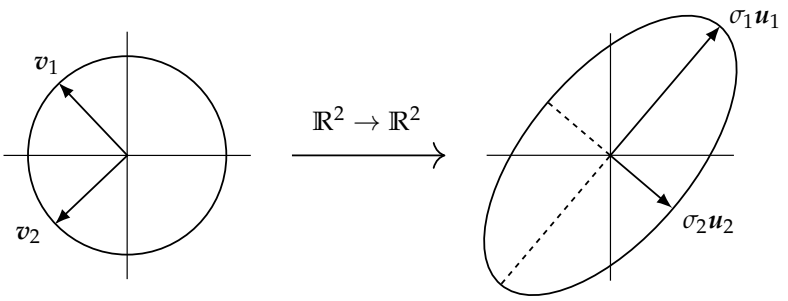

Figure 1.5: Image of the unit sphere under a linear transformation from $\mathbb{R}^2$ to $\mathbb{R}^2$.

1.4 Projection Matrices

Definition 1.13 (Projection matrix)**.** A *projection matrix* is a square matrix P that satisfies $P^2 = P$.

One can think about a general projection as the shadow of an object on a plane, when the sun (or a parallel light source) is shining with an angle. Notice that the definition does not necessarily imply that P projects points orthogonally. The direction of the projection can be easily deduced by drawing the line that connects $\boldsymbol{v}$ to $P\boldsymbol{v}$, i.e., the vector $P\boldsymbol{v} - \boldsymbol{v}$ (see Figure 1.6). By applying the projection matrix P to this vector, we get

$$P(P\boldsymbol{v} - \boldsymbol{v}) = P^2\boldsymbol{v} - P\boldsymbol{v} = \boldsymbol{0},$$

which means that $P\boldsymbol{v} - \boldsymbol{v} \in \ker(P)$. This shows that the direction of the projection is always described by a vector in $\ker(P)$.

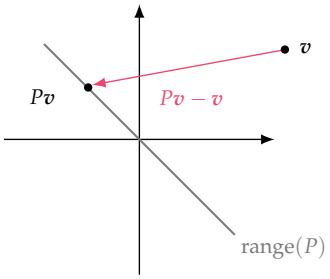

Figure 1.6: Projection of the vector $\boldsymbol{v}$ using the projection matrix P.

Exercise 1.4. Let $P \in \mathbb{R}^{n\times n}$ be a projection matrix. Then the *complementary projection* $Q \in \mathbb{R}^{n\times n}$ of P is given by

$$Q = I - P. \tag{1.3}$$

First, prove that Q is a projection matrix. Then, show that $\text{range}(Q) = \ker(P)$. With a similar argument, show that $\ker(Q) = \text{range}(P)$.

Solution. Consider a vector $\boldsymbol{u} \in \ker(P)$. Then

$$(I - P)\boldsymbol{u} = \boldsymbol{u} - P\boldsymbol{u} = \boldsymbol{u};$$

therefore, any vector in $\ker(P)$ is also in $\text{range}(I - P)$ and

$$\ker(P) \subseteq \text{range}(I - P). \tag{1.4}$$

Next, let $\boldsymbol{x} \in \text{range}(I - P)$, that is

$$\boldsymbol{x} = (I - P)\boldsymbol{v} = \boldsymbol{v} - P\boldsymbol{v}. \tag{1.5}$$

for some $\boldsymbol{v}$. Applying the projection matrix operator to $\boldsymbol{x}$ and using the equivalence in (1.5) we have

$$P\boldsymbol{x} = P\boldsymbol{v} - P^2\boldsymbol{v} = P\boldsymbol{v} - P\boldsymbol{v} = 0,$$

where we have used the fact that $P^2 = P$. We can conclude that $\boldsymbol{x} \in \ker(P)$ and that

$$\text{range}(I - P) \subseteq \ker(P). \tag{1.6}$$

Combining (1.4) with (1.6) we obtain

$$\text{range}(I - P) = \ker(P). \qquad \square$$

Definition 1.14 (Orthogonal projection)**.** A projection matrix is an *orthogonal projection* if its range and kernel are orthogonal (see Figure 1.7). We usually write Π for orthogonal projection matrices.

Let $\mathcal{V}$ be a linear subspace of $\mathbb{R}^n$. Then we know that any $\boldsymbol{x} \in \mathbb{R}^n$ can be uniquely decomposed as $\boldsymbol{x} = \boldsymbol{x}_{\mathcal{V}} + \boldsymbol{x}_{\mathcal{V}^\perp}$, where $\boldsymbol{x}_{\mathcal{V}} \in \mathcal{V}$, $\boldsymbol{x}_{\mathcal{V}^\perp} \in \mathcal{V}^\perp$.[3] The (linear) map that takes $\boldsymbol{x}$ and outputs $\boldsymbol{x}_{\mathcal{V}}$ is indeed an orthogonal projection (why?), and we denote it by $\Pi_{\mathcal{V}}$.

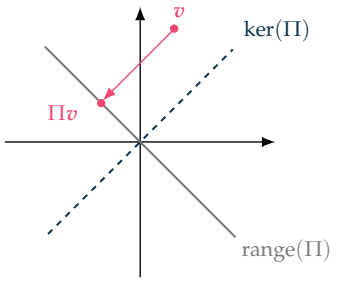

Figure 1.7: Projection of a vector $\boldsymbol{v}$ using the orthogonal projection Π.

[3] This is the so-called *orthogonal decomposition* of $\boldsymbol{x}$.

Definition 1.15 (Orthogonal projection onto a subspace)**.** Let $\mathcal{V}$ be a linear subspace of $\mathbb{R}^n$. The orthogonal projection $\Pi_{\mathcal{V}}$ whose range is $\mathcal{V}$ is called the *orthogonal projection onto* $\mathcal{V}$, and satisfies

$$\Pi_{\mathcal{V}}(\boldsymbol{x}) = \boldsymbol{x}_{\mathcal{V}}.$$

For a matrix X, we abuse the notation and write Π_X to denote $\Pi_{\text{range}(X)}$. This is indeed the orthogonal projection matrix on the span of columns of X.

Exercise 1.5. Show that a projection matrix Π is orthogonal if and only if Π is symmetric, that is, $\Pi = \Pi^\top$.

Solution. We only prove the "$\Leftarrow$" part, and leave the other direction to the reader.

Assume Π is a projection matrix with $\Pi = \Pi^\top$. Let us take two arbitrary vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ in the range and kernel of Π respectively. Our goal is to prove that $\boldsymbol{x} \perp \boldsymbol{y}$. We can write $\boldsymbol{x} = \Pi \boldsymbol{v}$ and $\boldsymbol{y} = (I - \Pi)\boldsymbol{w}$ for some $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^n$ (see Exercise 1.4). We then have:

$$\begin{aligned}
\langle \boldsymbol{x}, \boldsymbol{y} \rangle &= \langle \Pi \boldsymbol{v}, (I - \Pi)\boldsymbol{w} \rangle \\
&= \boldsymbol{v}^\top \Pi^\top (I - \Pi)\boldsymbol{w} \\
&= \boldsymbol{v}^\top \Pi (I - \Pi)\boldsymbol{w} && \text{as } \Pi \text{ is symmetric} \\
&= \boldsymbol{v}^\top (\Pi - \Pi^2)\boldsymbol{w} \\
&= 0 && \text{as } \Pi \text{ is a projection.}
\end{aligned}$$

We conclude that $\ker(\Pi) \perp \text{range}(\Pi)$ and hence, Π is an orthogonal projection. □

Exercise 1.6. Consider n linearly independent vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^m$ ($n \leq m$). What is the orthogonal projection matrix onto $\text{span}(\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\})$?

Solution. Consider the matrix $A \in \mathbb{R}^{m \times n}$ whose columns are given by $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$. It is clear that $\text{range}(A) = \text{span}(\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\})$, and as $\boldsymbol{a}_i$ are linearly independent, A is full-rank. Assume that $\boldsymbol{y} \in \mathbb{R}^m$ is the orthogonal projection of $\boldsymbol{v}$ on $\text{range}(A)$. Of course $\boldsymbol{y} \in \text{range}(A)$, thus, we can express $\boldsymbol{y}$ as $\boldsymbol{y} = A\boldsymbol{x}$ for some $\boldsymbol{x} \in \mathbb{R}^n$. As the projection is orthogonal,

$$\boldsymbol{y} - \boldsymbol{v} \perp \text{range}(A),$$

or equivalently

$$\begin{aligned}
\boldsymbol{y} - \boldsymbol{v} \perp \boldsymbol{a}_i\,, \quad \forall\, i \quad &\Leftrightarrow \quad \boldsymbol{a}_i^\top (A\boldsymbol{x} - \boldsymbol{v}) = 0\,, \quad \forall\, i \\
&\Leftrightarrow \quad \underbrace{\begin{bmatrix} — & \boldsymbol{a}_1^\top & — \\ & \vdots & \\ — & \boldsymbol{a}_n^\top & — \end{bmatrix}}_{A^\top} (A\boldsymbol{x} - \boldsymbol{v}) = 0 \\
&\Leftrightarrow \quad A^\top (A\boldsymbol{x} - \boldsymbol{v}) = 0 \\
&\Leftrightarrow \quad A^\top A\boldsymbol{x} = A^\top \boldsymbol{v}
\end{aligned}$$

It is easy to see that $A^\top A$ is invertible.[4] Using this fact we can conclude

[4] Let $A = U\Sigma V^\top$ be the SVD of A. Notice that since $n \leq m$,

$$\Sigma = \left[\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ \hline & \boldsymbol{0}_{(m-n)\times n} & \end{array}\right] \in \mathbb{R}^{m \times n}$$

is a tall matrix, where $\boldsymbol{0}_{(m-n)\times n} \in \mathbb{R}^{(m-n)\times n}$ is a matrix full of zeros. Moreover, since A is full-rank, $\sigma_i > 0$. We therefore have that

$$\begin{aligned}
A^\top A &= V\Sigma^\top U^\top U\Sigma V^\top \\
&= V\Sigma^\top \Sigma V^\top \\
&= V \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix} V^\top,
\end{aligned}$$

which is an invertible matrix.

that

$$x = (A^\top A)^{-1} A^\top \boldsymbol{v} \implies \boldsymbol{y} = A\boldsymbol{x} = A(A^\top A)^{-1} A^\top \boldsymbol{v}.$$

Thus, the projection matrix is $A(A^\top A)^{-1} A^\top$. □

To summarize, we showed in Exercise 1.6 that if A is full-rank,

$$\Pi_A = A(A^\top A)^{-1} A^\top.$$

Notice that if the columns of A were orthonormal, i.e., $A^\top A = I$, then $\Pi_A = AA^\top$. This is the same result that we had in Exercise 1.3.

1.4.1 *Pseudo-Inverse*

A useful construct related to orthogonal projections is the Moore-Penrose pseudo-inverse. Conceptually, the pseudo-inverse can be seen as a generalization of *inverse* for arbitrary matrices. But we will see that the pseudo-inverse of a matrix is closely related to the orthogonal projection on the range of that matrix. We first bring a formal definition of what properties should the pseudo-inverse satisfy, and then we give a computationally friendly way to describe it.

Definition 1.16 (Pseudo-inverse)**.** Let A be a real matrix. The pseudo-inverse of A is a matrix $A^\dagger$ satisfying the following properties:

$$\begin{aligned}
&(i) && AA^\dagger A = A,\\
&(ii) && A^\dagger AA^\dagger = A^\dagger,\\
&(iii) && (AA^\dagger)^\top = AA^\dagger,\\
&(iii) && (A^\dagger A)^\top = A^\dagger A.
\end{aligned}$$

It turns out that pseudo-inverse always exists and is unique.

Theorem 1.4. *The pseudo-inverse $A^\dagger$ of A also satisfies:*

$$\begin{aligned}
&(iv) && (A^\top)^\dagger = (A^\dagger)^\top,\\
&(v) && (AA^\top)^\dagger = (A^\dagger)^\top A^\dagger,\\
&(vi) && A^\dagger \boldsymbol{x} = 0 \Leftrightarrow \boldsymbol{x}^\top A = 0 \Leftrightarrow A^\top \boldsymbol{x} = 0.
\end{aligned}$$

The pseudo-inverse can be computed via SVD easily:

Theorem 1.5. *The pseudo-inverse of a real matrix $A \in \mathbb{R}^{m\times n}$ with SVD $A = U\Sigma V^\top$ is the matrix*

$$A^\dagger = V\Sigma^\dagger U^\top \in \mathbb{R}^{n\times m},$$

where $\Sigma^\dagger$ *is given by*

$$\Sigma^\dagger = \left[\begin{array}{ccc|c} \sigma_1^{-1} & & & \\ & \ddots & & \mathbf{0}_{r\times(m-r)} \\ & & \sigma_r^{-1} & \\ \hline & \mathbf{0}_{(n-r)\times r} & & \mathbf{0}_{(n-r)\times(m-r)} \end{array}\right] \in \mathbb{R}^{n\times m},$$

and $\sigma_1, \sigma_2, \ldots, \sigma_r$ *are the nonzero singular values of* A.[5]

[5] If $r = n$, matrix $\Sigma^\dagger$ does not contain the terms $\mathbf{0}_{(n-r)\times r}$ and $\mathbf{0}_{(n-r)\times(m-r)}$; similarly, if $r = m$ $\Sigma^\dagger$ does not contain $\mathbf{0}_{r\times(m-r)}$ and $\mathbf{0}_{(n-r)\times(m-r)}$.

We now bring the main theorem of this section, that the pseudo-inverse of a matrix is closely related to the orthogonal projection on the range.

Theorem 1.6 (Projection by pseudo-inversion)**.** *Let* $A \in \mathbb{R}^{m\times n}$ *be a real matrix. Then* $\Pi = AA^\dagger$ *is the orthogonal projection onto* range(A). *Consequently, the matrix* $Q = I - AA^\dagger$ *is the orthogonal projection onto* $\ker(A)$.

Proof. Any vector $\boldsymbol{x} \in \mathbb{R}^m$ can be orthogonally decomposed as $\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{u}$, where $\boldsymbol{u} \in$ range(A) and $\boldsymbol{v} \in$ range$(A)^\perp$. Since $\boldsymbol{u} = A\boldsymbol{r}$ for some $\boldsymbol{r} \in \mathbb{R}^n$, we have (by property (i) of Definition 1.16) that

$$AA^\dagger \boldsymbol{u} = AA^\dagger A\boldsymbol{r} = A\boldsymbol{r} = \boldsymbol{u}.$$

Since $\boldsymbol{v} \in$ range$(A)^\perp$, we have (by property (vi) of Theorem 1.5) that:

$$\begin{aligned} & \langle \boldsymbol{v}, A\boldsymbol{s}\rangle = 0 \ \ \forall \boldsymbol{s} \in \mathbb{R}^n, \\ \iff & \boldsymbol{s}^\top A^\top \boldsymbol{v} = 0 \ \ \forall \boldsymbol{s} \in \mathbb{R}^n, \\ \iff & A^\top \boldsymbol{v} = 0, \\ \iff & A^\dagger \boldsymbol{v} = 0. \end{aligned}$$

As a result, $AA^\dagger \boldsymbol{x} = AA^\dagger \boldsymbol{v} + AA^\dagger \boldsymbol{u} = \boldsymbol{u}$, and $\Pi = AA^\dagger$ is the orthogonal projection to range(A). Using Exercise 1.4, we can prove that $Q = I - AA^\dagger$ is the projection onto $\ker(A)$. □

1.5 Eigenvectors and Eigenvalues

Definition 1.17 (Eigenvector and Eigenvalue)**.** An n-dimensional nonzero vector $\boldsymbol{v}$ is an *eigenvector* of the $n \times n$ matrix A if it satisfies

$$A\boldsymbol{v} = \lambda \boldsymbol{v}, \tag{1.7}$$

for some scalar λ, called the *eigenvalue* associated to $\boldsymbol{v}$. In other words, the eigenvector $\boldsymbol{v}$ is a direction in the Euclidean space which is merely scaled by the linear mapping defined by A. The eigenvalue λ represents the scaling factor.

Notice that (1.7) has a nonzero solution in $\boldsymbol{v}$ if and only if there exists some λ such that

$$(A - \lambda I)\boldsymbol{v} = \mathbf{0},$$

or equivalently, if

$$\det(A - \lambda I) = 0. \tag{1.8}$$

Equation (1.8) is referred to as the *characteristic equation* of A, and $\det(A - \lambda I)$ is the *characteristic polynomial*.

While the characteristic equation (which is a polynomial) might have complex roots, if A is symmetric, all its roots become real:

Theorem 1.7. *If $A \in \mathbb{R}^{n\times n}$ is a real symmetric matrix then all of its eigenvalues are real and its eigenvectors can be chosen to form an orthonormal basis of the Eucledean space $\mathbb{R}^n$.*

We saw that *any* matrix has an SVD. If the study of eigenvectors and eigenvalues is of interest, there is another useful decomposition that shows this information. However, this decomposition only exists under specific assumptions. For example, if A is a square $n \times n$ matrix with n linearly independent eigenvectors, then A has a so-called *eigen-decomposition* or *spectral decomposition*, which is written as

$$A = Q\Lambda Q^{-1}.$$

Here, Q is a square $n \times n$ matrix whose columns are the n eigenvectors of A and Λ is a diagonal matrix whose diagonal elements are the corresponding eigenvalues. If A is symmetric, according to Theorem 1.7, Q will be orthonormal, and hence the decomposition looks like

$$A = Q\Lambda Q^{\top}.$$

The relation between singular values and eigenvalues of a matrix $A \in \mathbb{R}^{n\times n}$ is given by[6]

$$\sigma_i(A)^2 = \lambda_i(AA^{\top}) = \lambda_i(A^{\top}A).$$

[6] Here, $\lambda_i(B)$ is the ith eigenvalue of B.

1.6 Quadratic forms

Definition 1.18 (Quadratic form)**.** Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. The *quadratic form* induced by A is defined as the function $f : \mathbb{R}^n \to \mathbb{R}$ with $f(\boldsymbol{x}) = \boldsymbol{x}^{\top}A\boldsymbol{x}$.

Notice that the quadratic form $f(\boldsymbol{x})$ can be expressed as a polynomial of degree 2 in terms of the components of $\boldsymbol{x}$:

$$f(\boldsymbol{x}) = \boldsymbol{x}^{\top}A\boldsymbol{x} = \langle \boldsymbol{x}, A\boldsymbol{x}\rangle = \sum_{i,j} a_{i,j}x_ix_j\ , \quad i,j = 1,\dots,n\ ,$$

where $x_i \in \mathbb{R}$ is the i-th component of the vector $\boldsymbol{x}$ and $a_{i,j} \in \mathbb{R}$ are the entries of the matrix A.

Definition 1.19. A symmetric matrix with real entries $A \in \mathbb{R}^{n\times n}$ is *positive (semi-) definite*, or p.d. (p.s.d.), if and only if its induced quadratic form is positive (non-negative) for every nonzero $\boldsymbol{x} \in \mathbb{R}^n$, i.e.,

$$\begin{aligned} A \in \mathbb{R}^{n\times n} \text{ is p.s.d} &\iff \boldsymbol{x}^\top A\boldsymbol{x} \geq 0 \;\; \forall \boldsymbol{x} \in \mathbb{R}^n, \\ A \in \mathbb{R}^{n\times n} \text{ is p.d} &\iff \boldsymbol{x}^\top A\boldsymbol{x} > 0 \;\; \forall \boldsymbol{x} \in \mathbb{R}^n, \; \boldsymbol{x} \neq 0 \, . \end{aligned}$$

Example 3. Let

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \, , \quad A_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \, , \quad A_3 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \, . \tag{1.9}$$

It is easy to verify that A_1 is positive definite; A_2 is positive semi-definite and A_3 is nondefinite (i.e., neither positive semi-definite or negative semi-definite). The quadratic forms induced by A_1, A_2, A_3 in (1.9), are

$$\begin{aligned} f_1(\boldsymbol{x}) &= x_1^2 + x_2^2 \, , \\ f_2(\boldsymbol{x}) &= x_2^2 \, , \\ f_3(\boldsymbol{x}) &= x_2^2 - x_1^2 \, , \end{aligned} \tag{1.10}$$

where $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top$ and $x_1, x_2 \in \mathbb{R}$. Figure 1.8 shows the plots of the quadratic functions defined in (1.10).

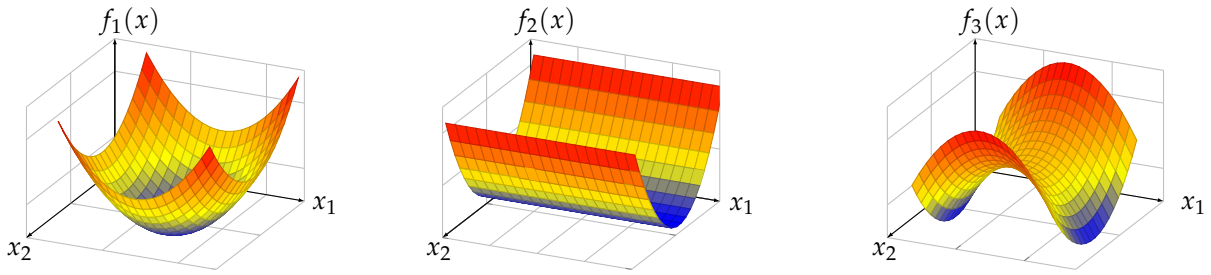

Figure 1.8: Plots of the quadratic forms described in (1.10).

Positivity of a matrix can be read-off from its eigenvalues:

Theorem 1.8. *A symmetric matrix with real entries is positive (semi-) definite if and only if all its eigenvalues are positive (nonnegative).*

Similar to the real numbers, where for any real $a \in \mathbb{R}$, we have $a^2 \geq 0$, the same argument holds for matrices, with some tweaks.

Theorem 1.9. *Let $A \in \mathbb{R}^{m\times n}$. Then the matrices $AA^\top$ and $A^\top A$ are symmetric and positive semi-definite. Moreover, if* $\operatorname{rank}(A) = m$ *then $AA^\top$ is also positive definite.*

Proof. Consider the quadratic form induced by $AA^\top$,

$$\boldsymbol{x}^\top AA^\top \boldsymbol{x} = (A^\top \boldsymbol{x})^\top (A^\top \boldsymbol{x}) \, , \quad \boldsymbol{x} \in \mathbb{R}^m \, . \tag{1.11}$$

Let $\boldsymbol{y} = A^\top \boldsymbol{x}$. Then the quadratic form in (1.11) can be rewritten as:

$$\boldsymbol{x}^\top A A^\top \boldsymbol{x} = \boldsymbol{y}^\top \boldsymbol{y} = \|\boldsymbol{y}\|^2 \geq 0.$$

This proves the first statement of Theorem 1.9 for $AA^\top$, the proof for $A^\top A$ is analogous.

To prove the second statement of the theorem, we need to show that if $\text{rank}(A) = m$ and $\boldsymbol{x} \neq \boldsymbol{0}$, then $\boldsymbol{y} \neq \boldsymbol{0}$. This is always true except if $\boldsymbol{x} \in \ker(A^\top)$; therefore, we need to prove that if $\text{rank}(A^\top) = m$ then $\ker(A^\top)$ is trivial ($= \{\boldsymbol{0}\}$). By the rank-nullity theorem (Theorem 1.1), we have that

$$\text{rank}(A^\top) + \dim(\ker(A^\top)) = m$$

therefore

$$\ker(A^\top) = \{\boldsymbol{0}\} \iff \text{rank}(A^\top) = m \iff \text{rank}(A) = m,$$

which is satisfied by assumption. This proves the second statement of the theorem. □

In real numbers, every non-negative number has a square root; similar argument holds for p.s.d. matrices. However, notice that the square root is not unique in general, but there exists a specific square root which is unique if the matrix is p.d.

Theorem 1.10 (Cholesky Decomposition)**.** *Every positive (semi-) definite matrix $A \in \mathbb{R}^{n\times n}$ can be factored as*

$$A = LL^\top, \tag{1.12}$$

where $L \in \mathbb{R}^{n\times n}$ is a lower triangular matrix with positive (nonnegative) entries on its diagonal. The factorization given in (1.12) is referred to as the Cholesky decomposition *of A and it is unique if A is positive definite.*

In the following example we use the Cholesky decomposition to characterize a set described by a quadratic inequality.

Example 4. Suppose $A \in \mathbb{R}^{n\times n}$ is a positive definite matrix and let $\boldsymbol{c} \in \mathbb{R}^n$. We will investigate the shape of the set

$$\mathcal{E} = \left\{\boldsymbol{x} \in \mathbb{R}^n : (\boldsymbol{x} - \boldsymbol{c})^\top A^{-1}(\boldsymbol{x} - \boldsymbol{c}) \leq 1\right\}.$$

To obtain a better characterization of $\mathcal{E}$, consider the Cholesky decomposition of A,

$$A = LL^\top.$$

Since A is positive definite, the determinant of L is positive[7] and L is invertible. As a result, we can write:

$$A^{-1} = (L^{-1})^\top L^{-1}.$$

Let $\boldsymbol{y} = L^{-1}(\boldsymbol{x} - \boldsymbol{c})$. Using $\boldsymbol{y}$, we can write:

$$(\boldsymbol{x} - \boldsymbol{c})^\top A^{-1}(\boldsymbol{x} - \boldsymbol{c}) = (\boldsymbol{x} - \boldsymbol{c})(L^{-1})^\top L^{-1}(\boldsymbol{x} - \boldsymbol{c}) = \boldsymbol{y}^\top \boldsymbol{y} = \|\boldsymbol{y}\|^2.$$

[7] We used the fact that the determinant of a triangular matrix is the product of the entries on its diagonal.

Notice that $\boldsymbol{x} = L\boldsymbol{y} + \boldsymbol{c}$; therefore, the set $\mathcal{E}$ can be equivalently described by:

$$\mathcal{E} = \{L\boldsymbol{y} + \boldsymbol{c} : \boldsymbol{y} \in \mathbb{R}^n, \ \|\boldsymbol{y}\| \leq 1\}. \tag{1.13}$$

Using (1.13) and the remark after the SVD theorem (Theorem 1.3) we conclude that $\mathcal{E}$ is a hyperellipse with center c.

1.7 Trace

Definition 1.20 (Trace)**.** The *trace* of a square matrix is the sum of its diagonal entries.

We now state the main properties of the trace.

1. The trace is a *linear functional*, i.e., for all square matrices $A, B \in \mathbb{R}^{n\times n}$ and all $c \in \mathbb{R}$ we have that

$$\operatorname{tr}(A + B) = \operatorname{tr} A + \operatorname{tr} B, \quad \operatorname{tr}(cA) = c \cdot \operatorname{tr} A.$$

2. For all $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times m}$ we have that

$$\operatorname{tr}(AB) = \operatorname{tr}(BA).$$

3. The trace is invariant under *cyclic permutations*, i.e.,

$$\operatorname{tr}(ABCD) = \operatorname{tr}(BCDA) = \operatorname{tr}(CDAB) = \operatorname{tr}(DABC).$$

Notice that arbitrary permuations are not allowed.

4. For any square matrix A and any invertible matrix P of appropriate dimensions we have that

$$\operatorname{tr} A = \operatorname{tr}(P^{-1}AP) = \operatorname{tr}(APP^{-1}).$$

5. Let Π_A be the orthogonal projection on the range(A), then $\operatorname{tr}(\Pi_A) = \operatorname{rank}(A)$.

6. The trace of a square matrix is the sum of its eigenvalues

$$\operatorname{tr} A = \sum_{i=1}^{n} \lambda_i,$$

where $A \in \mathbb{R}^{n\times n}$ and λ_i is the ith eigenvalue of A. (Remember that the determinant of a square matrix is the product of its eigenvalues $\det(A) = \prod_{i=1}^{n} \lambda_i$.)

Using the trace we can define an inner product on $\mathbb{R}^{m\times n}$ as

$$\langle A, B\rangle = \operatorname{tr}(A^\top B) = \sum_{i,j} a_{ij}b_{ij},$$

and a matrix norm as

$$\|A\|_F = \sqrt{\langle A, A\rangle} = \sqrt{\sum_{i,j} a_{ij}^2}. \tag{1.14}$$

The norm in (1.14) is referred to as the *Frobenius norm* of A.

The following exercise needs knowledge of probability theory. In case you are not comfortable with probability, leave the exercise here, and after reading the rest of the notes, come back and try to solve it.

Exercise 1.7. Let $\varepsilon \in \mathbb{R}^n$ be a random vector with mean μ and covariance Σ. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix and consider the quadratic form $\varepsilon^\top A\varepsilon$. What is the expected value $\mathbb{E}\left[\varepsilon^\top A\varepsilon\right]$ in terms of μ, Σ, and A?

Solution. Using the properties of the trace and the fact that $\varepsilon^\top A\varepsilon \in \mathbb{R}$, we have

$$\varepsilon^\top A\varepsilon = \operatorname{tr}\left(\varepsilon^\top A\varepsilon\right) = \operatorname{tr}\left(A\varepsilon\varepsilon^\top\right). \tag{1.15}$$

The expected value can now be computed as:

$$\begin{aligned}
\mathbb{E}\left[\varepsilon^\top A\varepsilon\right] &= \mathbb{E}\left[\operatorname{tr}\left(A\varepsilon\varepsilon^\top\right)\right], && \triangleright \text{From (1.15)}\\
&= \operatorname{tr}\left(\mathbb{E}\left[A\varepsilon\varepsilon^\top\right]\right), && \triangleright \operatorname{tr} \text{ and } \mathbb{E} \text{ are linear}\\
&= \operatorname{tr}\left(A\mathbb{E}\left[\varepsilon\varepsilon^\top\right]\right), && \triangleright A \text{ is deterministic, } \mathbb{E} \text{ linear}\\
&= \operatorname{tr}\left(A(\Sigma+\mu\mu^\top)\right), && \triangleright \mathbb{E}\left[\varepsilon\varepsilon^\top\right] = \Sigma+\mu\mu^\top\\
&= \operatorname{tr}(A\Sigma) + \operatorname{tr}\left(A\mu\mu^\top\right), && \triangleright \operatorname{tr} \text{ is linear}\\
&= \operatorname{tr}(A\Sigma) + \mu^\top A\mu. && \triangleright \text{Cyclic permuation}
\end{aligned}$$

We conclude that $\mathbb{E}\left[\varepsilon^\top A\varepsilon\right] = \operatorname{tr}(A\Sigma) + \mu^\top A\mu$. □

2

Analysis

This chapter will briefly recap the most important concepts from multivariate analysis, and recall the asymptotic notation.

Roadmap

We start by recalling what is a derivative of a multivariate function in Section 2.1. There, we introduce the gradient of a real-valued function, its relation to the derivative, and its geometric meaning. In Section 2.2, we review the chain rule, and use it to derive the directional derivative and derivative of the inverse function. This rule is computationally important for us later when learning about neural networks. Section 2.3 reviews the first-order necessary conditions of optimality. This is vital to understand how optimization algorithms work. In Section 2.4 and Section 2.5 we introduce the idea of approximating a function via its first and second order derivatives. This includes introducing the Hessian, which relates to quadratic form reviewed before in Section 1.6. Lastly, we recall the asymptotic notation in Section 2.6, which we will use to express how good an approximation is, and also use it to specify how our algorithms scale with input size later in the course.

Learning Objectives

After reading this chapter you should know

- what is the derivative and Jacobian, and how to compute the Jacobian for a differentiable function.
- what is the gradient of a function, its relation to derivative, and its geometric meaning that points to the steepest ascent direction.
- how to use chain rule to compute the derivative of the inverse of a function.
- what are critical points and what is their relation to optimization.
- how to construct a first- and second-order approximation to a function.
- how to interpret $f(x) = \mathcal{O}(g(x))$ and $f(x) = o(g(x))$ as $x \to a$.

2.1 *Multivariate Derivatives*

Let $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ be a map between Euclidean spaces, where Ω is some open subset of $\mathbb{R}^n$. Recall that the *derivative* of f at a point $\boldsymbol{x}_0 \in \Omega$ is the *best linear approximation* of f at $\boldsymbol{x}_0$ (assuming it exists).

Formally, if there exists a unique linear map $Df(\boldsymbol{x}_0) : \mathbb{R}^n \to \mathbb{R}^m$ that fulfills

$$\lim_{\|\boldsymbol{h}\|\to 0} \frac{\|f(\boldsymbol{x}_0+\boldsymbol{h}) - f(\boldsymbol{x}_0) - Df(\boldsymbol{x}_0)[\boldsymbol{h}]\|}{\|\boldsymbol{h}\|} = \boldsymbol{0}, \tag{2.1}$$

we say that f is differentiable at $\boldsymbol{x}_0$ and call $Df(\boldsymbol{x}_0)$ the derivative of f at $\boldsymbol{x}_0$. If $Df(\boldsymbol{x})$ exists for all $\boldsymbol{x} \in \Omega$, we call f differentiable. If additionally $Df(x)$ is continuous in $\boldsymbol{x}$, we call f continuously differentiable. From now on, we will only consider continuously differentiable maps f and omit any further case distinctions.

Note that because $Df(\boldsymbol{x}_0)$ is a linear map, we can interpret it as a matrix which we call the *Jacobian* of f. The entries of the Jacobian are $[Df(\boldsymbol{x}_0)]_{i,j} = \frac{\partial f_i}{\partial x_j}(\boldsymbol{x}_0)$, i.e.,

$$Df(\boldsymbol{x}_0) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}_0) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{x}_0) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{x}_0) \end{bmatrix}.$$

Specifically, if f is real-valued, the Jacobian $Df(\boldsymbol{x}_0)$ becomes a row vector:

$$Df(\boldsymbol{x}_0) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}_0) & \cdots & \frac{\partial f}{\partial x_n}(\boldsymbol{x}_0) \end{bmatrix}.$$

The transpose of this row vector, considered as a vector, is called the *gradient* of f at $\boldsymbol{x}_0$ and is often denoted as $\nabla f(\boldsymbol{x}_0)$. Note that it follows that

$$Df(\boldsymbol{x}_0)[\boldsymbol{h}] = \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{h} \rangle . \tag{2.2}$$

Among various important properties of the gradient, the following are particularly relevant to optimization, in which context they will repeatedly appear:

1. The gradient $\nabla f(\boldsymbol{x}_0)$ of f at $\boldsymbol{x}_0$ is a vector that points towards the direction in which f increases the most locally around $\boldsymbol{x}_0$. Likewise, $-\nabla f(\boldsymbol{x}_0)$ points towards the direction in which f decreases the most locally around $\boldsymbol{x}_0$. That is why the gradient is sometimes called the *steepest ascent* direction.
2. The above property also implies that $\nabla f(\boldsymbol{x}_0)$ is orthogonal to the level set of f at $\boldsymbol{x}_0$, defined as $\{\boldsymbol{x} \in \Omega \mid f(\boldsymbol{x}) = f(\boldsymbol{x}_0)\}$. In two dimensions, this corresponds to lines called contour lines, as Figure 2.1 shows.

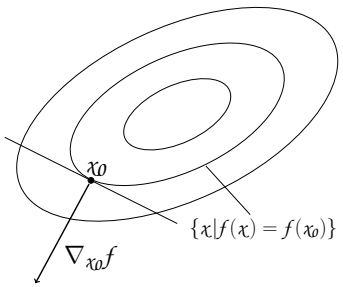

Figure 2.1: The gradient of f is orthogonal to its level sets.

2.2 Chain Rule

The chain rule is one of the fundamental laws of multivariate analysis and a handy tool to prove some further properties of the derivative as well as the basis for deep learning.

Theorem 2.1 (Chain Rule). *Let $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ and $g : \Omega' \subset \mathbb{R}^k \to \mathbb{R}^n$ be differentiable functions, where Ω and Ω' are open and* range$(g) \subset \Omega$. *Then it holds for $\boldsymbol{x}_0 \in \Omega'$ that*[1]

[1] The notation $\circ$ means the composition of two maps. So $f \circ g(x) = f(g(x))$.

$$D(f \circ g)(x_0) = (Df)(g(x_0)) \circ Dg(x_0).$$

Note that because $(Df)(g(x_0))$ and $Dg(x_0)$ are linear, their composition is also linear. Composition of linear maps in matrix form becomes matrix multiplication, and thus, the formula can be written in terms of Jacobians as

$$D(f \circ g)(x_0) = (Df)(g(x_0))Dg(x_0).$$

Also note that $(Df)(g(x_0))$ is *not* equal to the differential of $f \circ g$ evaluated at x_0, $D(f \circ g)(x_0)$, but rather the differential of f evaluated at $g(x_0)$.

We will omit the proof of Theorem 2.1, note however that this follows directly from the definition (2.1) of the derivative.

Example 5 (Directional Derivative)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ and consider the function $\phi_v : \mathbb{R} \to \mathbb{R}^n$ with $\phi_v(t) = x_0 + tv$. This function traces a line in the direction v passing through x_0. Recall that the *directional derivative* of f at x_0 in direction $v \in \mathbb{R}^n$ is defined as

$$\frac{\partial f}{\partial v}(x_0) := D(f \circ \phi_v)(0) = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t} \in \mathbb{R}.$$

Using the chain rule, we can show that the directional derivative in direction of $v \in \mathbb{R}^n$ corresponds to $Df(x_0)[v]$:

$$\frac{\partial f}{\partial v}(x_0) = D(f \circ \phi_v)(0) = (Df)(\phi_v(0)) \underbrace{D\phi_v(0)}_{=v} = Df(x_0)[v].$$

In particular, by the definition of the gradient,

$$\frac{\partial f}{\partial v}(x_0) = \langle \nabla f(x_0), v \rangle . \tag{2.3}$$

The chain rule also directly implies what the Jacobian of the inverse of a bijective $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ looks like.

Corollary 2.2 (Jacobian of the Inverse)**.** *Assume that f and f^{-1} are differentiable. Then it holds for any $y_0 \in \operatorname{range}(f)$ that*

$$Df^{-1}(y_0) = (Df)^{-1}(f^{-1}(y_0)).$$

Proof. Applying the chain rule from Theorem 2.1 to

$$I = D(\mathrm{id})(y_0) = D(f \circ f^{-1})(y_0) = (Df)(f^{-1}(y_0))Df^{-1}(y_0)$$

yields the result. □

2.3 Extremal Points

Recall that in one-dimensional analysis, first and higher order derivatives of a function can be used to find necessary and sufficient conditions for extremal points. The same is true for multivariate calculus, i.e., for functions $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$. Remember that $x_0 \in \Omega$ is a local minimum of f, if and only if there exists an $\varepsilon > 0$, such that for all x in an open neighbourhood of x_0 defined

as $\{\boldsymbol{x} \in \Omega \mid \|\boldsymbol{x} - \boldsymbol{x}_0\| < \varepsilon\}$, we have $f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0)$. Local maxima are defined analogously, except that $f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)$ needs to hold. The following theorem gives a necessary condition for being a local extremum.

Theorem 2.3 (Necessary Condition for Local Optimality)**.** *Let $\Omega \subset \mathbb{R}^n$ be open and $f : \Omega \to \mathbb{R}$ a continuously differentiable function. Assume $\boldsymbol{x}_0 \in \Omega$ is a local extremum. Then it holds that*

$$\nabla f(\boldsymbol{x}_0) = \mathbf{0}.$$

Therefore, when searching for local extreme points, we only have to search within the set of critical points *defined as $\{\boldsymbol{x} \in \Omega \mid \nabla f(\boldsymbol{x}) = \mathbf{0}\}$.*

Proof. Note that if $\boldsymbol{x}_0$ is a local extremum of f, it also must be a local extremum for any of the functions $\psi_v(t) = f \circ \phi_v(t) = f(\boldsymbol{x}_0 + t\boldsymbol{v})$. Because they are univariate functions $\psi_v : \mathbb{R} \to \mathbb{R}$, we know that $\frac{d\psi_v}{dt}(0) = 0$ for a local extremum $\boldsymbol{x}_0$. Note that $\frac{d\psi_v}{dt}(0) = \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{x}_0)$. As seen in equation (2.3), this implies that for all $\boldsymbol{v} \in \mathbb{R}^n$

$$(\nabla f(\boldsymbol{x}_0))^\top \boldsymbol{v} = \frac{d\psi_v}{dt}(0) = 0.$$

This in turn implies that $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$. □

Note that the converse is not true. If the gradient of a function is zero, the function does not necessarily need to have a local minimum.

Theorem 2.3, together with the property of the gradient that it always points towards the direction of the steepest ascent, motivates the *gradient descent* algorithm - one of the most popular optimization methods used in machine learning.

2.4 Taylor Expansions

It was already mentioned that the derivative is the best linear approximation of a function. This intuitive concept can be formulated rigorously by using the first-order Taylor expansion, as the following theorem shows.

Theorem 2.4 (First-order Taylor Expansion)**.** *Let $\Omega \subset \mathbb{R}^n$ be open and $f : \Omega \to \mathbb{R}$ be a continuously differentiable function. Let $\boldsymbol{x}_0 \in \Omega$. Then, it holds for all $\boldsymbol{x}$ in an open neighbourhood of $\boldsymbol{x}_0$ that*

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0 \rangle + o\left(\|\boldsymbol{x} - \boldsymbol{x}_0\|\right)$$

as $\boldsymbol{x} \to \boldsymbol{x}_0$.[2] *In other words, the approximation error $R(\boldsymbol{x}) = f(\boldsymbol{x}) - (f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0 \rangle)$ satisfies $\lim_{x \to x_0} \frac{R(\boldsymbol{x})}{\|\boldsymbol{x} - \boldsymbol{x}_0\|} = 0$.*

[2] See Section 2.6 for the definition of the little-o notation.

2.5 Second-order Derivatives, Hessian

Just like the first-order derivative is the best linear approximation, the second-order derivative is the best quadratic approximation. In higher dimensions, quadratic forms (see Section 1.6) are used to express quadratic functions.

Let $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ be a real-valued function. We learned that Df is the best linear approximation for f at any point. We say f is twice differentiable at $\boldsymbol{x}_0 \in \Omega$ if there exists a quadratic form $D^2 f(\boldsymbol{x}_0)$ such that

$$\begin{aligned} f(\boldsymbol{x}) = f(\boldsymbol{x}_0) & \\ & + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0 \rangle \\ & + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^\top D^2 f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \\ & + o(\|\boldsymbol{x} - \boldsymbol{x}_0\|^2) \quad \text{as} \quad \boldsymbol{x} \to \boldsymbol{x}_0. \end{aligned}$$

This quadratic form is called the second derivative of f at $\boldsymbol{x}_0$. As any quadratic form can be represented by a symmetric matrix, it turns out that the matrix for $D^2 f(\boldsymbol{x}_0)$ is

$$D^2 f(\boldsymbol{x}_0) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{x}_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\boldsymbol{x}_0) \end{bmatrix}$$

which we call the *Hessian* of f at $\boldsymbol{x}_0$. There is a fundamental link between the local curvature of a function f and its Hessian, as the following exercise suggests.

Exercise 2.1. Let $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top A \boldsymbol{x}$ be a quadratic form for a symmetric matrix $A \in \mathbb{R}^{n \times n}$. Show that the Hessian of f is exactly A, i.e., $D^2 f(\boldsymbol{x}) = A$ for all $\boldsymbol{x}$. Further, show that f has a unique minimum at $\mathbf{0}$ if A is positive definite, and has a unique maximum if A is negative definite. Moreover, if A is indefinite (i.e., has both positive and negative eigenvalues), prove that there are directions $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ such that $f(\alpha \boldsymbol{u}) \to +\infty$ and $f(\alpha \boldsymbol{v}) \to -\infty$ as $\alpha \to \infty$.

2.6 Asymptotic Notation

Asymptotic notation describes the limiting behavior of a function when the argument tends towards a particular value or infinity. It is used in computer science to classify algorithms according to how their run time or space requirements grow as the input size grows. In analysis, it is used to provide growth bounds via easier-to-understand functions, such as polynomials, logarithms, or exponentials [Wik22].

2.6.1 Big-O notation

The *big-O notation* is used to provide *upper bounds* on the growth of a function.

Definition 2.1 (Big-O notation)**.** Let f and g be real valued functions defined on some unbounded subset of the positive real numbers. We write

$$f(x) = \mathcal{O}(g(x)) \quad \text{as} \quad x \to \infty$$

if there exists a positive scalar M and a real number $x_0 > 0$ such that

$$|f(x)| \leq M\,|g(x)| \quad \text{for all } x \geq x_0.$$

In this case, we say that $f(x)$ is of order $\mathcal{O}(g(x))$ asymptotically. Similarly, for a fixed number $a \in \mathbb{R}$ we write

$$f(x) = \mathcal{O}(g(x)) \quad \text{as} \quad x \to a,$$

if there exists some $\delta > 0$ and $M > 0$ such that $|f(x)| \leq M|g(x)|$ for all $0 < |x - a| < \delta$.

In the definition above, it is usually the case that g is an easy-to-understand function and the growth behavior of f is controlled by g. Moreover, one should always mention the limiting argument (if the limit is taken as $x \to \infty$ or $x \to a$).

Example 6 (Approximation error)**.** Let f be the second order Taylor expansion of the exponential function around the point 0:

$$f(x) = 1 + x + \frac{x^2}{2}.$$

Obviously, f is only an approximation and to get an idea of how well it approximates e^x around 0, we can use the big-O notation:

$$e^x - f(x) = \mathcal{O}(x^3) \quad \text{as} \quad x \to 0,$$

that is, the error is smaller than some constant times x^3 if x is close enough to 0. This result is due to the Taylor's theorem.

We now list two important rules for manipulating $\mathcal{O}$ terms:

- If $g(x) = \mathcal{O}(f(x))$ then $cg(x) = \mathcal{O}(f(x))$ for any constant c.
- If $g_1(x)$ and $g_2(x)$ are both $\mathcal{O}(f(x))$ then so is $g_1(x) + g_2(x)$.

2.6.2 *Little-o notation*

The *little-o* notation can be used to express that a function grows slower than some other function. For example, $f(x) = o(g(x))$ signifies that f grows much slower than g and is insignificant in comparison.

Definition 2.2 (Little-o notation)**.** Let f and g be two real valued functions defined on some unbounded subset of the positive real numbers, and let a be a fixed real number or infinity. Provided that g is nonzero in proximity of a, we write

$$f(x) = o(g(x)) \quad \text{as} \quad x \to a$$

if

$$\lim_{x \to a} \frac{f(x)}{g(x)} = 0.$$

Example 7. We have $x^n = o(e^x)$ as $x \to \infty$ for any $n > 0$.

While Big-O notation can be intuitively interpreted as "grows ultimately as fast as", the little-o notation can be understood as "ultimately grows slower than".

3

Probability Theory

In this chapter, we will recap the fundamentals of probability theory. It is a challenge to introduce probability theory while keeping a good level of rigor. We will take a middle ground and keep ourselves away from a fully rigorous treatment of the topic.

Roadmap

We start from scratch, by introducing what is a probability space in Section 3.1. In Section 3.2 we introduce one of the cornerstones of this chapter: random variables. We define what is the distribution of a random variable and how to characterize it in general via the CDF. We then focus on two classes of random variables: discrete and continuous. Discrete random variables can take a discrete set of values and can be described easily via the probability mass function. This is the content of Section 3.2.1, where we also introduce some famous distributions. In Section 3.2.2 and Section 3.2.3, we deal with continuous random variables, which are trickier that discrete ones. We only focus on those that have a so-called density, and introduce some famous continuous distributions at the end. Afterwards, in Section 3.2.4 we introduce another cornerstone of our exposition: the expected value. We will also recall variance of random variables in Section 3.2.5.

After understanding a single random variable, we move on to the situation of multiple random variables. This is the subject of Section 3.3. There, we introduce notions such as joint distribution, independence, marginals and conditional distribution. We finish by mentioning three important theorems about conditional distributions in Section 3.3.9. Section 3.4 is one of the most important parts of our exposition. There we recall how to find the expected value of a function of several random variables, and understand the important notion of conditional expectation. We finish our exposition with the important Gaussian (or normal) distribution in Section 3.5, and bring an exemplar of theorems that show why Gaussians are important.

Learning Objectives

After reading this chapter you should know

- what is a probability space.
- what is a random variable, its distribution, CDF, and what does it mean for a random variable to be discrete or continuous.
- what is PMF and PDF.
- what does expected value represent and how to compute it for a discrete or continuous random variable.

- what is the law of unconscious statistician.
- joint distributions, joint PMF and PDF.
- how to marginalize on a subset of random variables.
- how to verify if events are independent, notion of conditional probability for random variables, law of total probability, chain rule, and Bayes rule.
- how to verify if random variables are independent, notion of conditional distribution of random variables.
- how to compute the expected value of a function of several random variable, linearity of expectation.
- how to interpret $\mathbb{E}_X[f(X,Y)]$.
- the covariance matrix.
- normal distribution.
- the law of large numbers.

3.1 *Probability Spaces*

Let Ω be any set. We will call Ω the *sample space*, and it will be interpreted as the set of all possible outcomes of an experiment. Choosing a suitable Ω is a part of modelling the real world experiment and thus is not necessarily unique. For example, think about possible outcomes of throwing a dart to a dartboard.

- $\Omega = \mathbb{R}^2$ could model the exact place of landing,
- $\Omega = \{0,1\}$ could model a hit or miss, and
- $\Omega = \{0,10,20,\dots,100\}$ could model the score.

Most of the time, especially when Ω is a complicated set (like in the first example), we do not care too much about *what* exactly happened. Rather than asking whether a certain element of Ω has happened or not, we want to ask harder questions. For example, we want to know whether the score is higher than 60 or not, or the point is at least 1 cm away from the center of the dartboard. Notice that these types of questions naturally correspond to specific subsets of Ω. For example, the former question corresponds to the subset $\{80,100\}$. We call each of these subsets an *event*. We also say that "event A occurs" if the outcome of the experiment is in A.

Bonus Material

We call a family $\mathcal{F}$ of subsets of Ω that encode our questions of interest, the *family of events*. Notice how logical "and" is translated into intersection of events, "or" into union, and "not" into complement. Thus, we desire our family of events to be closed under these operations, that is, $\mathcal{F}$ needs to be a so-called *σ-algebra*:

- $\Omega \in \mathcal{F}$,
- if $A \in \mathcal{F}$, then also $A^c \in \mathcal{F}$,

- for $A_1, A_2, \cdots \in \mathcal{F}$, their union must also be contained in $\mathcal{F}$.

Note that these properties also imply that $\mathcal{F}$ is closed under intersection. Examples of $\mathcal{F}$ satisfying these properties are $\mathcal{F} = \{\Omega, \emptyset\}$ or $\mathcal{F} = \mathcal{P}(\Omega)$, the power set of Ω.

After identifying the sample space (Ω) and the set of all events ($\mathcal{F}$), we can go ahead and assign probabilities to the events.[1] We describe this assignment via a function $\mathbb{P} : \mathcal{F} \to [0, 1]$, called the *probability function*. To have a consistent theory, this assignment has to obey certain axioms:

1. Total probability equals one: $\mathbb{P}(\Omega) = 1$, and
2. Probability is additive for disjoint events: for all at most countably many, pairwise disjoint events $A_i \in \mathcal{F}$ it should hold that

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i).$$

We call the triple $(\Omega, \mathcal{F}, \mathbb{P})$ a *probability space*.

[1] The question of *how* to assign probabilities to events is a philosophical one. There are several proposals, and we deal with *frequentist* and *Bayesian* approach in this course. In later sections, these conceptions will be introduced.

Example 8. Consider the experiment of throwing a die and looking at the top face. We model this problem by setting the sample space to be $\Omega = \{⚀, ⚁, \ldots, ⚅\}$. As any subset of Ω can be written as a union of singletons (sets of size 1), we only have to define the probability function for singletons. For example, let $\mathbb{P}(\{⚀\}) = p_1, \ldots, \mathbb{P}(\{⚅\}) = p_6$, where $p_1, \ldots, p_6$ are some numbers between 0 and 1 adding up to 1. Observe that, e.g., $\mathbb{P}(\{⚁, ⚄\}) = p_2 + p_5$.

The following two exercises remind you of two important properties of a probability function.

Exercise 3.1 (Monotonicity)**.** Let $A, B \in \mathcal{F}$ be such that $B \subseteq A$. Show that $\mathbb{P}(B) \leq \mathbb{P}(A)$.

Exercise 3.2 (Union Bound)**.** Suppose we have at most countably many not necessarily disjoint events $A_i \in \mathcal{F}$. Prove

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i).$$

This inequality comes in quite handy in many applications and proofs.

3.2 Random Variables

Random variables are a quantitative way of looking at an experiment. Their job is to take the outcome of an experiment and assign a number to it.

Definition 3.1 (Random Variable)**.** A random variable X is a function that assigns a real number to every outcome of an experiment, $X : \Omega \to \mathbb{R}$.[2] Notice that the function is not random, while the "randomness" is in the input to the function.

[2] As this course is an introductory course, we do not concern ourselves with measure theory.

We bring this quote from [Wil91]:

Once Tyche, Goddess of Chance, decides the outcome $\omega \in \Omega$, the values of all random variables lock into place.

We can naturally use random variables to create events. A notion that is useful in this regard is the *inverse image*. Given a subset of $\mathbb{R}$, we ask what elements of Ω are mapped into that subset. Formally, for a random variable X and a subset $A \subset \mathbb{R}$, we define

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}.$$

We sometimes write this set as $\{X \in A\}$. For example, for an interval $[a, b]$, the event $\{a \leq X \leq b\}$ is the set of all outcomes whose X-value is between a and b. It turns out that all interesting events that one might consider are intervals and their unions and intersections. We denote by $\mathcal{B}(\mathbb{R})$ the family of all these subsets. The *distribution* of a random variable tells us the probabilities of all these events, and is sufficient for a full characterization of X.

Definition 3.2 (Distribution of a Random Variable)**.** For any $A \in \mathcal{B}(\mathbb{R})$, we define

$$\mathbb{P}_X(A) := \mathbb{P}\left(X^{-1}(A)\right) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right),$$

We call $\mathbb{P}_X$ the distribution (or *law*) of X.

Notice that $\mathbb{P}_X$ acts on subsets of $\mathbb{R}$ and tells us the probability that X takes a value in that subset, while $\mathbb{P}$ acts on events and tells us how probable that event is. The nice thing about $\mathbb{P}_X$ is that it is defined over subsets of a fixed set ($\mathbb{R}$), and to characterize it, we do not need to think in terms of Ω, which can sometimes be a difficult set to reason about. Indeed, $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ is a probability space! The proof of this claim is straightforward, and we leave it as an exercise.

It turns out that one can describe any probability function on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by its values on the intervals of the form $(-\infty, x]$ with $x \in \mathbb{R}$. Hence, to describe the law of a random variable, it suffices to identify its value on these subsets only.

Definition 3.3 (Cumulative Distribution Function)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable. Then, the cumulative distribution function (CDF) of X is defined as

$$\begin{aligned} F_X(x) &:= \mathbb{P}_X((-\infty, x]) \\ &= \mathbb{P}\left(X^{-1}((-\infty, x])\right) \\ &= \mathbb{P}\left(\{\omega \in \Omega \mid X(\omega) \leq x\}\right) \\ &=: \mathbb{P}\left(X \leq x\right). \end{aligned}$$

The CDF of a random variable X has the following properties:

- $\lim_{x\to\infty} F_X(x) = 1$ and $\lim_{x\to-\infty} F_X(x) = 0$,

- it is monotonically increasing,
- it is right-continuous,[3]
- and $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$.

[3] Meaning that for every $x \in \mathbb{R}$, the function F_X agrees with its right limit: $\lim_{y \to x^+} F_X(y) = F_X(x)$.

In general, F_X does not need to be continuous. Assume, for example, that for some $x_0 \in \mathbb{R}$, $\mathbb{P}(X = x_0) = \alpha > 0$. Then $F_X(x_0) - \lim_{x \to x_0^-} F_X(x) = \alpha > 0$ and F_X cannot be continuous, see Figure 3.1.

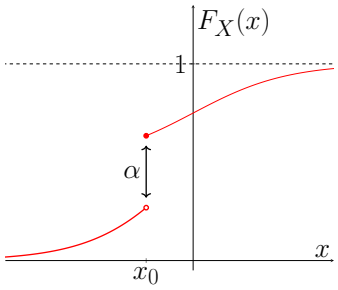

Figure 3.1: The CDF of a random variable with a discontinuity at x_0.

In the following sections, we focus on *discrete* and *continuous* random variables. For each of these cases, one can often describe the law of the random variable in an easier way.

3.2.1 Discrete Random Variables

A random variable is *discrete* if the set of values it can output, denoted by $\mathcal{X}$, is a discrete set (a finite or countable set).[4]

[4] In the die example, $\mathcal{X} = \{100, 200, \ldots, 600\}$.

Example 9 (Dice, continued)**.** In the experiment of throwing a die, suppose we will multiply the face number by 100 and donate that amount to charity. This naturally corresponds to creating a function that translates any outcome into a number (the amount of money we donate). We can then perform the experiment (i.e., throw the die) and apply the function to the outcome and donate the value.

Recall for each possible value $x \in \mathcal{X}$, the notion of the event $\{X = x\}$, which is a shorthand for $\{\omega \in \Omega : X(\omega) = x\}$. Knowing the probability of these events completely identifies $\mathbb{P}_X$, as all other events related to X can be written as a union of sets of the form $\{X = x\}$.

Definition 3.4 (Probability Mass Function)**.** Let X be a discrete random variable and $\mathcal{X}$ be the set of all of its values. We define for each $x \in \mathcal{X}$

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}_X(\{x\}).$$

We call p_X the *probability mass function* (PMF) of X. If X is clear from the context, we just write $p(x)$ instead of $p_X(x)$.

Example 10 (Dice, continued)**.** Let X be the amount of money we donate. Then, the probability mass function of X is $p_X(100) = p_1, \ldots, p_X(600) = p_6$. Moreover, the probability that we donate less than 350 is

$$\mathbb{P}(X < 350) = \mathbb{P}(\{X = 100\} \cup \{X = 200\} \cup \{X = 300\}) = p_1 + p_2 + p_3.$$

Remark. In our running example, one observes that the actual mechanism that resulted in a donation value is not important (whether it was a die that we threw and we multiplied its face by 100, or we asked a random person on the street to choose a real number in $[0, 1]$, and we multiply it by 600 and round it up to a number in $\{100, \ldots, 600\}$). The only thing that matters is "what values can X take" and "with what probability each value is produced". This information is encoded in p_X. That is why we mostly forget about $\Omega, \mathcal{F}, \mathbb{P}$ and directly talk about a random variable X with a probability mass function p_X.

Below is a list of discrete random variables with their associated PMF. We also bring a model probability space on which these random variables are defined.

- **Bernoulli.** We write $X \sim \text{Ber}(q)$ if $\mathcal{X} = \{0,1\}$ and $p_X(1) = q$ (and hence, $p_X(0) = 1 - q$). A model for this random variable is the throw of a (biased) coin, whose probability of landing Heads is q. Defining $X(\text{Heads}) = 1$ and $X(\text{Tails}) = 0$ will result in the desired random variable.
- **Categorical.** We write $X \sim \text{Cat}(p_1, \dots, p_k)$ if $\mathcal{X} = \{1, \dots, k\}$ and $p_X(i) = p_i$ for $i = 1, \dots, k$ (we require $\sum p_i = 1$ and $p_i > 0$). A model for this random variable is selecting among k different choices, each with some probability, and assigning $1, \dots, k$ to the choices.
- **Binomial.** We write $X \sim \text{Binom}(q, n)$, with $q > 0$ and $n \in \mathbb{N} \setminus \{0\}$, if $\mathcal{X} = \{0, 1, \dots, n\}$ and

$$p_X(k) = \binom{n}{k} q^k (1-q)^{n-k}, \quad k = 0, 1, \dots, n.$$

  A model for this random variable is the number of heads in a sequence of n independent tosses of the same coin, where q is the probability of a head.

Bonus Material

- **Poisson.** We write $X \sim \text{Poisson}(\lambda)$, with $\lambda > 0$, if $\mathcal{X} = \mathbb{N}$ and

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2 \dots$$

  The Possion distribution with parameter λ is a good approximation of the Binomial distribution with parameters n and q if n is large, q is small and $\lambda = nq$. In fact, provided that $\lambda = nq$, then

$$\lim_{n \to \infty} p_{X,\text{Binom}}(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \tag{3.1}$$

  where with $p_{X,\text{Binom}}(k)$ we denote the Binomial PMF with parameters n and q. As a result, we can use the Possion random variable to model Binomial random variables where n is large and q is small (e.g. the number of car accidents in a city on a given day, since the number of cars is large and the probability of accident for a single car is very small).

Exercise 3.3. Prove (3.1). Hint: if $\lambda = nq$, then the Binomial PMF with parameters n and q can be rewritten as

$$\begin{aligned} p_X(k) &= \frac{n!}{(n-k)!k!} q^k (1-q)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k}. \end{aligned}$$

3.2.2 Continuous Random Variables

In this section, we introduce *continuous* random variables. Our goal is to give a somewhat rigorous, but simultaneously intuitive, short intro for the sake of completeness. A truly rigorous exposition can be found in probability theory books (see, e.g., [Dur19]).

A random variable is *continuous*, if, loosely speaking, the set of values it can produce is uncountably infinite and the probability of attaining a single value is zero, that is $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$.[5]

[5] In these notes, we do not differentiate between continuous and absolutely continuous random variables.

Example 11. Take the experiment of choosing a "random" point on a disk. That is, let $\Omega = \{x \in \mathbb{R}^2 \mid \|x\| \leq 1\}$ be the two-dimensional disk, and $\mathbb{P}$ be the uniform measure, i.e., for $A \subseteq \Omega$ we have $\mathbb{P}(A) = \text{area}(A)/\text{area}(\Omega)$.[6] Let

$$X : \Omega \to \mathbb{R}, \quad X(\omega) = \|\omega\|$$

be the distance of the point ω to the center of the disk. Then X is an example of a continuous random variable. Notice that $\mathbb{P}(X = a) = 0$ for any $a \in [0, 1]$ (why? try to prove this fact).

[6] One has to be careful here, as there are subsets of the disk that one cannot assign any area to them. To resolve this issue, the Borel σ-algebra and measurability should be rigorously defined, which we do not discuss here.

Remark. It turns out that all random variables are either discrete, continuous, or "a mixture of the two".[7] This implies that we only have to treat two cases: discrete and continuous.

[7] This rather technical statement is called the Lebesgue decomposition theorem.

3.2.3 Probability Density

To understand a continuous random variable, knowing its CDF is sufficient. However, many random variables that we are considering in this course can be characterized even easier, in terms of a density function. Before we go into details, we first bring an intuitive explanation of what a density is.

Let M be a non-homogeneous physical object, for example, a rock. We define $\text{diam}(M)$ to be the diameter of M, $\text{vol}(M)$ to be its volume, and $m(M)$ to be its mass. Take a point $x \in M$ and consider small balls around x. For each of these neighborhoods, compute the mass and volume. By physical intuition, we know that if we make the neighborhood smaller and smaller, these two quantities tend to zero, i.e., if I is a neighborhood around x, we have

$$\lim_{\substack{x \in I \\ \text{diam}(I) \to 0}} \text{vol}(I) = 0, \qquad \lim_{\substack{x \in I \\ \text{diam}(I) \to 0}} m(I) = 0.$$

However, if we divide these two numbers, the ratio converges to a number which we call *density of M at x*,

$$\lim_{\substack{x \in I \\ \text{diam}(I) \to 0}} \frac{m(I)}{\text{vol}(I)} = \rho(x).$$

The relation between density and mass becomes clear with the following formula: for any subset A of M,

$$m(A) = \int_A \rho(x)\, dx.$$

That is, the density is there to be integrated!

Remember that if you ask "What is the mass of a single point $m(\{x\})$?" the answer would be 0, but one can assign a density to a single point, which could be nonzero.

We can do the same thing with continuous random variables by replacing "mass" with "probability". We can then introduce the density at the point $a \in \mathbb{R}$ (if it exists) as

$$p_X(a) := \lim_{\substack{a \in I \\ |I| \to 0}} \frac{\mathbb{P}(X \in I)}{|I|}.$$

Here, I is an interval containing x and $|I|$ is its length.

As in our physical example, for a subset $A \subseteq \mathbb{R}$ we have

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \int_A p_X(x)\, dx.$$

Definition 3.5 (Probability Density Function)**.** Let X be a random variable. If there exists a (measurable) function $p_X : \mathbb{R} \to [0, \infty)$, such that

$$\mathbb{P}_X(I) = \mathbb{P}(X \in I) = \int_I p_X(x)\, dx$$

for all intervals I in $\mathbb{R}$, we call it the *probability density function* (PDF) of X.

Notice that $\int_{\mathbb{R}} p_X(x)\, dx = 1$, since $\mathbb{P}_X(\mathbb{R}) = 1$. The relation between CDF and density is as you might guess:

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} p_X(t)\, dt.$$

Consequently, if F_X is differentiable, it holds that

$$\frac{dF_X}{dx}(x_0) = p_X(x_0). \tag{3.2}$$

Here, we bring some famous continuous random variables and their densities:

- **Uniform** We write $X \sim \text{Unif}([a, b])$, with $a, b \in \mathbb{R}$ and $a < b$, if $X \in [a, b]$ and

  $$p_X(x) = \begin{cases} c & \text{if } x \in [a, b] \\ 0 & \text{otherwise,} \end{cases}$$

  where c can be determined from a and b using

  $$\int_a^b c\, dx = c(b - a) = 1,$$

  hence

  $$c = \frac{1}{b - a}.$$

  The uniform random variable can be used to model events where intervals of the same length are equally likely.

- **Exponential** We write $X \sim \text{Exp}(\lambda)$, with $\lambda > 0$, if $X \in \mathbb{R}$ and

$$p_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

  An exponential random variable can be used, for example, to model the amount of time before a piece of equipment breaks down. Notice that the probability that X exceeds a certain value $x \geq 0$ decreases exponentially with growing values of x:

$$\Pr(X \geq a) = \int_a^\infty \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_a^\infty = e^{-\lambda a}.$$

- **Normal** We write $X \sim \mathcal{N}(\mu, \sigma)$, with $\sigma > 0$ and $\mu \in \mathbb{R}$, if $X \in \mathbb{R}$ and

$$p_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

  X is a *standard normal random variable* if $X \sim \mathcal{N}(0,1)$. The normal random variable plays a fundamental role in signal processing, since it generally models well the additive effect of many independent factors. Mathematically, this is captured by the *central limit theorem*, which states that the sum of a large number of independent random variables drawn from the same distribution tends to a normal distribution, even if the original variables were not normally distributed.

3.2.4 Expected Value

If we redo the same experiment many times and look at the average of the observed values of a random variable X, it starts to "converge" to a number, which we call the *expected value* of X. We will restate this fact in a rigorous way later. If we want to guess what this value would be for a discrete random variable, the following definition would make sense:

Definition 3.6 (Expected Value, Discrete)**.** Let X be a discrete random variable with values in $\mathcal{X}$. We define the *expected value of* X (or its *first moment*) as

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}(X = x) = \sum_{x \in \mathcal{X}} x\, p_X(x).$$

If the above sum does not converge, we say X has no expected value.

For continuous random variables that have a PDF, the definition of expected value is as follows:

Definition 3.7 (Expected Value, Continuous)**.** Let X be a continuous random variable with density p_X. We define the expected value of X as[8]

$$\mathbb{E}[X] := \int_{\mathbb{R}} x\, p_X(x)\, dx.$$

[8] The integral in Definition 3.7 can be infinite. An example is the random variable X with the PDF

$$p_X(x) = \frac{1}{\pi(1+x^2)}.$$

This distribution is called the Cauchy distribution, and looks surprisingly much like the normal distribution (see Figure 3.2). However, it has so-called "heavy tails", which means that the density does not become small fast enough as $|x| \to \infty$.

Exercise 3.4 (St. Petersburg Problem)**.** Suppose you enter the following game: you flip a fair coin until it comes up heads. Let X denote the round that the coin turns up heads.

(a) What is the PMF of X and its expected value? This random variable follows the so-called *Geometric distribution*.

(b) Suppose you get a reward of 2^n Francs if the game stops at round n. What is the expected value of your reward?

The following theorem is sometimes called the "law of the unconscious statistician," as one thinks that it is trivial, while it is not. It is a nice exercise to prove it for the discrete case:

Theorem 3.1. *Let X be a random variable, and $g : \mathbb{R} \to \mathbb{R}$ be a function.*

- *If X is discrete,*
$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x)\, p_X(x).$$
- *If X is continuous having a PDF p_X, then*
$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\, p_X(x)\, dx.$$

3.2.5 Variance

The *variance* of a random variable, in contrast to its expectation, does not measure the location, but rather the spread.

Definition 3.8 (Variance)**.** Let X be a random variable. The variance of X is defined as
$$\operatorname{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

Note that the variance of X is finite, if and only if $\mathbb{E}\left[X^2\right]$ (often called *the second moment of X*) is finite. Also, note that $\operatorname{Var}(X) \geq 0$ by Jensen's inequality (Lemma 3.16).

3.3 Jointly Distributed Random Variables

In many occasions, there are several random variables defined over the same probability space, and we wish to study events that involve more than one of the variables. First notice that there can be "dependency" among the random variables. Take the following extreme example: in a die tossing experiment, let X be the top face and Y be the bottom. Clearly, if the die is fair, X and Y have the same distribution (both uniform on $\{1, \ldots, 6\}$). However, it is always the case that $X + Y = 7$, regardless of what happens. Hence, knowing X determines Y.

3.3.1 Joint Distributions

In the general case, to understand a set of random variables, there is no other way than asking for the probabilities of all possible events that concern all random variables.

When looking at two random variables X and Y at the same time, it is useful to pack them together and look at them as a function that maps every outcome of the experiment to a *vector* in $\mathbb{R}^2$, that is, $\omega \in \Omega$ is mapped to $(X(\omega), Y(\omega)) \in \mathbb{R}^2$. The events (questions) that are related to both X and Y can be described as subsets of $\mathbb{R}^2$.

This viewpoint is the key to the definition of the joint distribution, which is very similar to Definition 3.2.

Definition 3.9 (Joint Distribution)**.** Let X, Y be two random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The joint distribution of X, Y is the map $\mathbb{P}_{X,Y}$ that takes a subset A of $\mathbb{R}^2$ as input,[9] and gives the probability

$$\mathbb{P}_{X,Y}(A) = \mathbb{P}\left(\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A\}\right)$$

[9] Just to be careful here, not all subsets of $\mathbb{R}^2$ are allowed. However, almost all of the subsets that are interesting for us are valid inputs to $\mathbb{P}_{X,Y}$ and we call these $\mathcal{B}(\mathbb{R}^2)$.

Example 12. Suppose we choose a person uniformly at random from a population Ω. We measure the person's weight in kilograms and call that X, and measure the person's height in meters and call it Y. The event "the person has a body-mass-index lower than 25" can be described using X and Y:

$$\left\{\omega \in \Omega : X(\omega)/Y(\omega)^2 < 25\right\} \subset \Omega.$$

The same event can be described as the set of all points (x, y) in $\mathbb{R}^2$ such that $\{x/y^2 < 25\}$. The probability function $\mathbb{P}$ gets the former set as input and gives a probability, while the probability distribution function $\mathbb{P}_{X,Y}$ gets the latter set as input and give out the same number, that is,

$$\mathbb{P}\left(\left\{X/Y^2 < 25\right\}\right) = \mathbb{P}_{X,Y}\left(\left\{(x, y) \in \mathbb{R}^2 : x/y^2 < 25\right\}\right).$$

It turns out that the *joint cumulative distribution function* is sufficient to identify the joint distribution.

Definition 3.10 (Joint CDF)**.** Given a collection of random variables $X_1, \ldots, X_n$ defined on the same probability space, their joint CDF is defined as

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}\left(X_1 \leq x_1, \ldots, X_n \leq x_n\right)$$

for all $x_1, \ldots, x_n \in \mathbb{R}$.[10]

[10] Notice the notation:

$$\mathbb{P}\left(X = x, Y = y\right) = \mathbb{P}\left(\{X = x\} \cap \{Y = y\}\right).$$

In what follows, we treat the discrete and continuous random variables separately, and show how the joint distribution can be described in different ways.

Let X, Y be two discrete random variables taking values in $\mathcal{X}, \mathcal{Y}$ respectively. It turns out that probabilities of the form $\mathbb{P}\left(X = x, Y = y\right)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ are sufficient to give a complete characterization of X and Y.

Definition 3.11 (Joint PMF)**.** Let $X_1, \dots, X_n$ be discrete random variables, defined on the same probability space, and taking values in $\mathcal{X}_1, \dots, \mathcal{X}_n$. The function

$$p_{X_1,\dots,X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$$

defined over $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ is called the *joint probability mass function of* $X_1, \dots, X_n$. We write $p(x_1, \dots, x_n)$ if the random variables are clear from the context.

With the same ideas described in Section 3.2.3, we can define the joint PDF of a set of continuous random variables (if it exists):

Definition 3.12 (Joint PDF)**.** Let the random variables $X_1, \dots, X_n$ be defined on the same probability space. We say these random variables have a *joint probability density function* $p_{X_1,\dots,X_n}$ if for all subsets $A \in \mathcal{B}(\mathbb{R}^n)$ it holds

$$\mathbb{P}_{X_1,\dots,X_n}(A) = \mathbb{P}((X_1, \dots, X_n) \in A) = \int_A p_{X_1,\dots,X_n}(x_1, \dots, x_n)\, dx_1 \cdots dx_n.$$

This definition again emphasizes the role of the density: to compute the probability of events about a set of random variables, one can integrate the density.

The same relation with CDF holds in jointly distributed random variables:

Theorem 3.2. *If the random variables $X_1, \dots, X_n$ with joint CDF F have a joint density at $\boldsymbol{x} \in \mathbb{R}^n$, then*

$$p_{X_1,\dots,X_n}(\boldsymbol{x}) = \frac{\partial^n F}{\partial x_1 \cdots \partial x_n}(\boldsymbol{x}).$$

Not all collections of continuous random variables have a joint density:

Example 13. Let X be uniformly distributed on $[0, 1]$ and let $Y = X$. Then X, Y do *not* have a joint density.

Remark. Let X be a continuous and Y be a discrete random variable, both defined on the same probability space. We can still describe the joint distribution by the joint CDF. However, things might get complicated, as something as straightforward as PMF or PDF does not exist in this case. In this course, whenever this situation of mixed joint random variables occurs, it is better to think in terms of conditional distributions, see Section 3.3.5.

Sometimes we talk about a *random vector*. A random vector $\boldsymbol{X} = (X_1, ..., X_n)$ is just a vector made up of random variables X_i defined on the *same* probability space. X_i can have a joint distribution, density and so on. This notion brings forward the fact that every realization is a point in $\mathbb{R}^n$ and bears some geometric meaning.

3.3.2 Marginals

The joint distribution of $X_1, \ldots, X_n$ encodes the distribution of every one of X_i's too.

Definition 3.13 (Marginal Distribution)**.** Suppose we are given the joint distribution $\mathbb{P}_{X,Y}$ of X and Y. Then, for any $A \in \mathcal{B}(\mathbb{R})$, one can compute the distribution $\mathbb{P}_X$ as

$$\mathbb{P}_X(A) = \mathbb{P}_{X,Y}(A \times \mathbb{R}) = \mathbb{P}_{X,Y}\left(\left\{(x,y) \in \mathbb{R}^2 : x \in A\right\}\right).$$

In this context, $\mathbb{P}_X$ is sometimes referred to as the marginal distribution of X. The act of computing $\mathbb{P}_X$ from $\mathbb{P}_{X,Y}$ is called *marginalization on* X. If the joint CDF of X, Y is given, then the CDF of X can be computed as

$$F_X(x) = F_{X,Y}(x, +\infty).$$

If instead of the distribution or CDF, we are given a joint PMF or PDF, we can marginalize as follows:

Let X, Y be discrete random variables. Fix $x \in \mathcal{X}$, and compute

$$\sum_{y \in \mathcal{Y}} p(x,y) = \sum_y \mathbb{P}\left(X = x, Y = y\right) = \mathbb{P}\left(\{X = x\} \cap \left(\bigcup_y \{Y = y\}\right)\right),$$

and notice that $\bigcup_y \{Y = y\} = \Omega$, and hence

$$\sum_{y \in \mathcal{Y}} p(x,y) = \mathbb{P}\left(\{X = x\} \cap \Omega\right) = \mathbb{P}\left(X = x\right) = p_X(x).$$

Thus, summing over all values of Y of the joint PMF gives the PMF of X.

With continuous random variables one has to "integrate out" the other variable. For the case of two continuous random variables X, Y with joint density $p_{X,Y}$, the (marginal) density of X is given by

$$p_X(x) = \int_{\mathbb{R}} p_{X,Y}(x,y)dy.$$

Marginalization can be easily generalized to a set of random variables. Let $X_1, \ldots, X_n$ be discrete random variables taking values in $\mathcal{X}_1, \ldots, \mathcal{X}_n$ and joint PMF p. If we want to know the joint distribution of a subset $I \subset \{1, \ldots, n\}$, we have to sum over all values of the random variables not in I. For example, if $I = \{1, \ldots, k\}$, then for any $(x_1, \ldots, x_k) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$ we have

$$p_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \sum_{x_{k+1} \in \mathcal{X}_{k+1}} \cdots \sum_{x_n \in \mathcal{X}_n} p_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n).$$

The same holds for continuous random variables $X_1, \ldots, X_n$ with joint density $p_{X_1,\ldots,X_n}$. The marginal PDF of $X_1, \ldots, X_k$ is

$$p_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \int \cdots \int p_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n)\, dx_{k+1} \cdots dx_n.$$

3.3.3 *Independence and Conditional Probability of Events*

First, an intuitive primer on independence is in order. Consider infinitely many tosses of a fair coin. If you look at the ratio of heads in the first n tosses, this number converges to $\frac{1}{2}$ as n increases. Now what if you just look at the tosses in even rounds? What if you just look at the times which are prime, i.e., $2, 3, 5, 7, 11, \ldots$? Our intuition tells us that the ratio of heads is still $\frac{1}{2}$. But now what if we look at the times when the coin has come Heads? The ratio now is going to be 1.

More generally, let A and B be two events. We do the experiment multiple times and try to estimate $\mathbb{P}(A)$ by the ratio of the number of experiments that A happened to the total number of experiments. Now suppose we only look at those experiments that B has happened, and estimate the probability of $\mathbb{P}(A)$ on those. If the estimate changes, it means that A and B are dependent.

Definition 3.14 (Independence and Conditional Probability)**.** Let A, B be two events.[11]

(1) A and B are *independent* iff $\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

(2) If $\mathbb{P}(B) > 0$, the *conditional probability of A given B* is defined as

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)}.$$

(3) Let $A_1, \ldots, A_n$ be events. They are called independent, if for any $k \leq n$ and every k indices $1 \leq i_1 < \cdots < i_k \leq n$,

$$\mathbb{P}\left(A_{i_1}, \ldots, A_{i_k}\right) = \mathbb{P}\left(A_{i_1}\right) \cdots \mathbb{P}\left(A_{i_k}\right)$$

[11] We use comma (,) interchangeably with $\cap$ when writing down probabilities of events. So $\mathbb{P}(A, B)$ reads "probability of A and B" and is the same as $\mathbb{P}(A \cap B)$.

Example 14. In this example, we show that pairwise independence of events does not imply their independence. Throw a red and a blue fair coin. Let A, B, C be the following events: A happens when the red coin turns Heads, B happens when the blue coin turns Heads, and C happens when exactly one of the coins is Heads. Verify that A, B, C are pairwise independent, but $\mathbb{P}(A, B, C) = 0 \neq \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C) = \frac{1}{8}$.

Exercise 3.5. Construct an experiment and n events such that each $(n-1)$ of them are independent, but they are not independent.

Here we bring a list of theorems that concern conditional probabilities and are useful in computations and proofs.

Let A, B be events and $A_1, ..., A_n$ be a partition of the sample space Ω into events, i.e., $\Omega = \bigcup_{i=1}^{n} A_i$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$. Suppose that $\mathbb{P}(A_i) > 0$ for all i and $\mathbb{P}(B), \mathbb{P}(A) > 0$. Take any arbitrary events $C_1, ..., C_n$ with $\mathbb{P}(C_i) > 0$ for all i.

Theorem 3.3 (Law of Total Probability)**.** *One has*

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i).$$

Theorem 3.4 (Bayes Rule)**.** *One has*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Theorem 3.5 (Chain Rule)**.** *It holds that*[12]

$$\mathbb{P}(C_1, \ldots, C_n) = \mathbb{P}(C_1)\,\mathbb{P}(C_2 \mid C_1) \cdots \mathbb{P}(C_n \mid C_1, \ldots, C_{n-1})$$

[12] Be careful of the notation again: the right hand side of this formula is indeed $\prod_{i=1}^{n} \mathbb{P}\left(C_i \middle| \bigcap_{j=1}^{i-1} C_j\right)$.

Exercise 3.6. We have two coins that look similar, one is unbiased (probability of Heads is $\frac{1}{2}$), and the other one is biased (probability of Heads is $\frac{1}{3}$). We pick one of them at random and throw it twice. The result was HT. What is the probability that we have picked the unbiased coin?

3.3.4 Independence of Random Variables

Next, we mention the concept of *independence* for random variables.

Definition 3.15 (Independence of Random Variables)**.** Two random variables X, Y are called independent, if their joint CDF factorizes. That is, for all $x, y \in \mathbb{R}$,

$$F_{X,Y}(x, y) = F_X(x)\,F_Y(y).$$

The following theorem summarizes equivalent conditions under which random variables are guaranteed to be independent:

Theorem 3.6. *Let X and Y be random variables. The following are equivalent:*

- *X and Y are independent.*
- *If X and Y are both discrete, their joint PMF factorizes: $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$.*
- *If X and Y are both continuous and have a joint PDF, their joint PDF factorizes: $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ for all $x, y \in \mathbb{R}$.*

Extension to more than two random variables is similar. For example, the continuous random variables $X_1, \ldots, X_n$ are independent iff

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n)$$

for all $x_1, \ldots, x_n \in \mathbb{R}$.

3.3.5 Conditional Distribution

Given the joint distribution of two (or more) random variables, now we answer the question: "how much the knowledge about a random variable influence our belief/probabilities about others." To be more precise, we expect that if there is dependency among random variables, information about one "changes" the distribution of the others. This concept is formalized as *conditional distribution*. We avoid discussing this concept in full generality and bring three different cases that are of our interest.

3.3.6 Conditional Distribution: Discrete Case

Definition 3.16 (Conditional PMF)**.** Let X, Y be two discrete random variables on the same probability space taking values in $\mathcal{X}$ and $\mathcal{Y}$ respectively. For $y \in \mathcal{Y}$ such that $\mathbb{P}(Y = y) \neq 0$ we define the *conditional distribution of X given $Y = y$* as

$$p_{X|Y}(x \mid y) := \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

Clearly, if X and Y are independent, then $p(x \mid y) = p(x)$.

Notice that for any fixed $y \in \mathcal{Y}$ with $\mathbb{P}(Y = y) > 0$, the function $p_{X|Y}(\cdot \mid y)$ is a probability mass function: by definition, $p_{X|Y}(x \mid y) \geq 0$ for all $x \in \mathcal{X}$ and

$$\sum_{x \in \mathcal{X}} p_{X|Y}(x \mid y) = \sum_{x \in \mathcal{X}} \frac{p_{X,Y}(x, y)}{p_Y(y)} = \frac{1}{p_Y(y)} \sum_x p_{X,Y}(x, y) = 1.$$

Thus, we can consider a random variable with this PMF, and call it *X conditioned on $Y = y$*.

The notion of conditional distribution can be generalized to any number of random variables. Let $X_1, X_2, \ldots, X_n$ be discrete random variables taking values in $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$. For any $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, \ldots, x_k \in \mathcal{X}_k$ such that $\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) > 0$, the conditional distribution of $X_{k+1}, \ldots, X_n$ given $X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k$ is defined as

$$p_{X_{k+1},\ldots,X_n|X_1,\ldots,X_k}(x_{k+1}, \ldots, x_n \mid x_1, \ldots, x_k) := \frac{p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)}{p_{X_1,\ldots,X_k}(x_1, \ldots, x_k)}.$$

3.3.7 Conditional Distribution: Continuous Case

Here, we assume both X and Y are continuous. Recall that since Y is a continuous random variable, it holds that $\mathbb{P}(Y = y) = 0$ for all $y \in \mathbb{R}$. Therefore, the discrete definition that is based on probability mass would not be sensible. Instead, we define the *conditional density*.

Definition 3.17 (Conditional Density)**.** Let X, Y be two continuous random variables with joint density $p_{X,Y}$. The conditional density of X given $Y = y$ is defined as

$$p_{X|Y}(x \mid y) = \begin{cases} \frac{p_{X,Y}(x,y)}{p_Y(y)} & \text{if } p_Y(y) > 0 \\ 0 & \text{else.} \end{cases}$$

As in the discrete case, $X \mid Y = y$ is a new random variable with the PDF above,[13] and is called X conditioned on $Y = y$. If it is clear from the context, we sometimes drop the subscript of $p_{X|Y}(x \mid y)$ and just write $p(x \mid y)$.

Similar to the discrete case, the notion of conditional distribution can be extended to the case of more than two continuous random variables.

[13] It is easy to verify that for a fixed $y \in \mathbb{R}$ such that $p_Y(y) > 0$, the function $p(\cdot \mid y)$ defined above is indeed a probability density.

3.3.8 Conditional Distribution: Mixed Case

There are some cases where we study two (or more) random variables where some of them are continuous and some are discrete.

Suppose X is a continuous and Y is a discrete random variable, defined on the same probability space. For each $y \in \mathcal{Y}$, we have a conditional probability distribution $\mathbb{P}_{X|y}$ on $\mathbb{R}$, defined as follows: for each subset $A \in \mathcal{B}(\mathbb{R})$ we have

$$\mathbb{P}_{X|y}(A) = \frac{\mathbb{P}_{X,Y}(\{(x,y) \in \mathbb{R}^2 : x \in A\})}{\mathbb{P}_Y(\{y\})} = \frac{\mathbb{P}(X \in A, Y = y)}{p_Y(y)}$$

When the conditional distribution above has a density, we call it *the conditional probability density of X given $Y = y$* and denote it as $p_{X|Y}(x \mid y)$. In this course, we can always assume that such a density exists for all $y \in \mathcal{Y}$.

What if we want to condition on $X = x$? In this case, the conditional distribution will be discrete, and Bayes theorem (Theorem 3.4) comes to the rescue:

$$p_{Y|X}(y \mid x) = \frac{p_{X|Y}(x \mid y)p_Y(y)}{p_X(x)}, \quad \forall y \in \mathcal{Y}.$$

Example 15 (Three radioactive materials)**.** Suppose we have three different radioactive materials y_1, y_2, y_3, each having a different rate of particle emission, and we mixed different quantities of these materials in a batch. Let X be the (random) time that a particle is emitted from the batch and let Y be the type of the material that caused the emission. Clearly, X is a continuous random variable, while Y is a discrete one. Moreover, there is a clear dependency between X and Y: knowing Y will determine the material, which as a result, changes the distribution of X.

Assume we observed a particle emitted from the batch at time x. A question one may ask is to determine how our perception of the distribution of Y changes after obtaining this new knowledge about X, i.e., we would like to know how likely it is that the materials y_1, y_2, y_3 were the cause of the emission knowing that the emission time was x. This can be translated in mathematical terms to finding the conditional distribution $p_{Y|X}(y \mid x)$ for the three possible values of Y.

At first, this may look difficult: we do not have an explicit description of $p_{Y|X}$. However, we can exploit the knowledge of another conditional distribution which we know well: the conditional $p_{X|Y}$, i.e., the PDF of the emission time for each of the radioactive materials.

The two conditionals are related via Bayes theorem:

$$p_{Y|X}(y \mid x) = \frac{p_{X|Y}(x \mid y)p_Y(y)}{p_X(x)}. \tag{3.3}$$

In (3.3) the PMF p_Y represents our prior beliefs about the distribution of Y, i.e., what we believed the chances of y_1, y_2, y_3 to emit a particle were before observing any particle emission. Using (3.3) we would like to update our prior belief by leveraging the observed emission time x and obtain a posterior distribution $p_{Y|X}(y \mid x)$. The only term that remains unknown in

(3.3) is the PDF p_X, which can be obtained using the law of total probability (Theorem 3.7):

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X|Y}(x \mid y) p_Y(y).$$

3.3.9 Theorems on Conditional Distribution

Here we bring the random variable version of the results in Section 3.3.3. In what follows, X and Y are random variables, and p can be a probability density, or a probability mass function, depending on the context. If nothing is stated about p, it can be either of them.

Theorem 3.7 (Law of Total Probability)**.** *If Y is continuous with a density, then*

$$p_X(x) = \int p_{X|Y}(x \mid y)\, p_Y(y)\, dy.$$

If Y is discrete, then

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X|Y}(x \mid y)\, p_Y(y).$$

Theorem 3.8 (Bayes Rule)**.** *If X is continuous with a density,*

$$p_{X|Y}(x \mid y) = \frac{p_{Y|X}(y \mid x)\, p_X(x)}{\int p_{Y|X}(y \mid x')\, p_X(x')\, dx'}$$

If X is discrete, the integral is replaced by a sum over $x' \in \mathcal{X}$.

Theorem 3.9 (Chain Rule)**.** *Let $X_1, \ldots, X_n$ be random variables. Then*

$$p(x_1, \ldots, x_n) = p(x_1)\, p(x_2 \mid x_1) \cdots p(x_n \mid x_1, \ldots, x_n).$$

3.3.10 Conditional Independence

Let X, Y, Z be random variables. The random variables X and Y are said to be *conditionally independent given* Z if given any value of Z, the probability distribution of X is the same for all values of Y and the probability distribution of Y is the same for all values of X. That is,

$$p_{X,Y|Z}(x, y \mid z) = p_{X|Z}(x \mid z) p_{Y|Z}(y \mid z)$$

for all x, y, z.

3.4 Properties of Expectation

3.4.1 Expectation of Functions of Several Random Variables

Suppose X, Y are two discrete random variables and let $f : \mathbb{R}^2 \to \mathbb{R}$ be a function, whose job is to take the values of X and Y and

produce a number. By the same argument as in Theorem 3.1, we have

$$\mathbb{E}\left[f(X,Y)\right] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)f(x,y).$$

The same holds for continuous random variables. If $X_1,\ldots,X_n$ are continuous random variables and $f:\mathbb{R}^n\to\mathbb{R}$ is a function, then

$$\mathbb{E}\left[f(X_1,\ldots,X_n)\right] = \int_{\mathbb{R}^n} f(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x},$$

where $p(\boldsymbol{x})$ is the joint density of $X_1,\ldots,X_n$ evaluated at $\boldsymbol{x} = (x_1,\ldots,x_n)$.

A special case is when $f(\boldsymbol{x}) = x_1+\cdots+x_n$. Then, regardless of X_i being independent or not, we have

$$\mathbb{E}\left[X_1+\cdots+X_n\right] = \mathbb{E}\left[X_1\right]+\cdots+\mathbb{E}\left[X_n\right].$$

This property, together with the fact that for $\alpha\in\mathbb{R}$, $\mathbb{E}\left[\alpha X\right] = \alpha\mathbb{E}\left[X\right]$, constitute the *linearity of expectation*.

In the special case where X and Y are independent, the expectation also respects products, as the following theorem states.

Theorem 3.10. *Let X,Y be independent random variables. Then*

$$\mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right].$$

3.4.2 Conditional Expectation

Conditional expectation in its simplest form is the expected value of the conditional distribution. To be more precise, let X,Y be random variables. Suppose we want to condition on $Y=y$ and see what is the expected value of X given this information. Using the conditional distribution, this is an easy task: If X is continuous and a conditional density exists, then

$$\mathbb{E}\left[X \mid Y=y\right] := \int x\, p_{X|Y}(x\mid y)\, dx.$$

The same holds if X is discrete; in that case, we replace the integral by a sum over $x\in\mathcal{X}$.

Notice that for each value of Y, we can compute the conditional expectation. This naturally introduces a new random variable:

$$\omega\in\Omega \mapsto Y(\omega) \mapsto \mathbb{E}\left[X \mid Y=Y(\omega)\right].$$

We define $\mathbb{E}\left[X\mid Y\right]$ to denote this new random variable and call it *the conditional expectation of X given Y*. Note that if Y is discrete, then this random variable is discrete, and if it is continuous, it can be discrete or continuous (depending on what values $\mathbb{E}\left[X\mid Y=y\right]$ can produce for different values of $y\in\mathbb{R}$). In any case, the conditional expectation is a *function* of the random variable Y whose value at

$Y = y$ is $\mathbb{E}[X \mid Y = y]$. This viewpoint of treating the conditional expectation as a random variable is very important and is ubiquitous in the course.

A very important property of conditional expectation is the following theorem, that is incredibly useful in calculations:

Theorem 3.11 (Tower Property)**.** *Let X, Y be two random variables. Then*

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]].$$

The proof follows from the definition of conditional expectation and Theorem 3.1.

If Y is discrete, the theorem reads:

$$\mathbb{E}[X] = \sum_{y \in \mathcal{Y}} \mathbb{E}[X \mid Y = y] \; p_Y(y),$$

and if Y is continuous,

$$\mathbb{E}[X] = \int \mathbb{E}[X \mid Y = y] \; p_Y(y) \, dy.$$

The following exercise is from [Ros14].

Exercise 3.7. Suppose that the number of people entering a store on a given day is a random variable with mean 50. Suppose further that the amounts of money spent by these customers are independent random variables having a common mean of \$8. Finally, suppose that the amount of money spent by a customer is also independent of the total number of customers who enter the store. What is the expected amount of money spent in the store on a given day?

Solution. Let $X_1, X_2, \ldots$ denote the money spent by each customer, and let N be the number of customers entering the shop. What we want is to compute the expected value of $\sum_{i=1}^{N} X_i$. For this computation, we first compute the conditional expectation given on $N = n$, then use the tower property.

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N} X_i \mid N = n\right] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i \mid N = n\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] && X_i \text{ and } N \text{ are independent} \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i] && \text{linearity of expectation} \\
&= 8n.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N} X_i \mid N\right]\right] && \text{tower property} \\
&= \mathbb{E}[8N] \\
&= 400.
\end{aligned}$$

What we have shown is that $\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[N]\,\mathbb{E}[X_1]$. □

3.4.3 Conditional Expectation Notations

In this small section, we bring a new notation that is not standard in the literature, but is intuitive and useful in our study.

Oftentimes in the lecture, we deal with expectations of functions of several variables (random or deterministic). In some cases, we use a subscript to indicate which variable is being averaged over (as opposed to the default case where we average over *all* variables; see Section 3.4.1). For instance,

$$\mathbb{E}_X [f(X, Y)]$$

denotes the average of the function f with respect to the (marginal) distribution of X. Notice that $\mathbb{E}_X [f(X, Y)]$ will be a function of Y. Moreover, Y can be a random variable, or a parameter. If Y is a random variable and we want to average over conditional distribution given $Y = y$, we sometimes write $\mathbb{E}_{X|y} [f(X, Y)]$.

In the following list, we bring these notational examples and their meaning:

1. $\mathbb{E}_X [f(X, Y)] := \int f(x, Y)\, p_X(x)\, dx$. If Y is a deterministic variable, then this is a function of Y, otherwise, this becomes a random variable which is a function of Y.
2. $\mathbb{E}_{X|y} [f(X, Y)] := \int f(x, y)\, p_{X|Y}(x \mid y)\, dx$.

3.4.4 Covariance

When studying two or more random variables, we sometimes want a quantitative measure of how dependent they are. The covariance is an example of such a measure, that quantizes *linear* dependency between random variables.

Definition 3.18 (Covariance)**.** Let X and Y be random variables. The covariance of X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

Note that $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

For random vectors $\boldsymbol{X} = (X_1, ..., X_p)^\top$ we define the expectation

$$\mathbb{E}[\boldsymbol{X}] = (\mathbb{E}[X_1], ..., \mathbb{E}[X_p])^\top \in \mathbb{R}^p$$

and the covariance matrix

$$\mathrm{Cov}(\boldsymbol{X}) = \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^\top\right] \in \mathbb{R}^{p \times p}.$$

Note that the covariance matrix is of the form

$$\mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} \mathrm{Var}(X_1) & \dots & \mathrm{Cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \dots & \mathrm{Var}(X_p) \end{pmatrix} \in \mathbb{R}^{p \times p}.$$

Exercise 3.8. Prove that for any random vector $\boldsymbol{X}$ with finite second moments, $\mathrm{Cov}(\boldsymbol{X})$ is symmetric and positive semi-definite.

3.5 Normal Distributions

One of the most important distributions a random variable or random vector can have is the *normal distribution*. In particular, it is the subject of the central limit theorem (Theorem 3.15), which we will discuss later.

Definition 3.19 (Normal Distribution)**.** Define $p_{\mu,\Sigma} : \mathbb{R}^p \to [0,\infty)$ as

$$p_{\mu,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

for a vector $\mu \in \mathbb{R}^p$ and a symmetric, positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$. If a random vector $\boldsymbol{X}$ has the density $p_{\mu,\Sigma}$, we say that it follows a normal distribution with mean μ and covariance matrix Σ, and we write $\boldsymbol{X} \sim \mathcal{N}(\mu,\Sigma)$. Conveniently, it holds that if $\boldsymbol{X} \sim \mathcal{N}(\mu,\Sigma)$, $\mathbb{E}[\boldsymbol{X}] = \mu$ and $\mathrm{Cov}(\boldsymbol{X}) = \Sigma$.

Note that if $p = 1, \mu = 0$ and $\Sigma = \sigma^2 = 1$, this translates to

$$\phi(x) := p_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

We call this the standard normal distribution $\mathcal{N}(0,1)$ and denote its c.d.f. by Φ. Figure 3.2 visualizes the standard normal distribution and the Cauchy distribution.

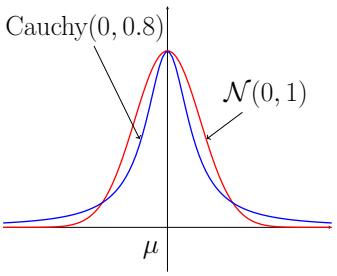

Figure 3.2: Densities of the normal distribution (red) and Cauchy distribution (blue) in comparison. Note the heavy tails of the Cauchy distribution.

Theorem 3.12 (Normal Distribution under Affine Transformations)**.** *Let $\boldsymbol{X} \sim \mathcal{N}(\mu,\Sigma)$. Then for any $A \in \mathbb{R}^{p\times p}$ and any $b \in \mathbb{R}^p$ it holds that*

$$AX + b \sim \mathcal{N}\left(A\mu + b, A\Sigma A^\top\right).$$

Let $1 \le k \le p$. By taking the matrix A as $A_{i,j} = 1$ if $i = j = k$ and $A_{i,j} = 0$ otherwise, we see the following corollary.

Bonus Material

Corollary 3.13 (Marginals of Normal Distributions)**.** *Let $\boldsymbol{X} \sim \mathcal{N}(\mu,\Sigma)$. Then for any $1 \le k \le p$, it holds that*

$$X_k = A\boldsymbol{X} \sim \mathcal{N}\left(A\mu, A\Sigma A^\top\right) = \mathcal{N}(\mu_k, \Sigma_{kk}).$$

Specifically, the marginals of a normal distribution are also normal distributions.

Remark. The converse is *not* true. If you have two random variables X_1 and X_2 following normal distributions, (X_1, X_2) does not have to follow a two-dimensional normal distribution! For a collection of counterexamples, see [Kow73].

3.6 Convergence of Empirical Averages to Expectation

In this section we deal with sequences of random variables and the behaviour of their average.

Definition 3.20 (i.i.d.)**.** The random variables $X_1, X_2, \ldots$ defined on the same probability space are called *independent with identical*

distribution (i.i.d.) if they are independent and each X_i has the same distribution.

Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite first moment. Recall the definition of the sample mean:

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The laws of large numbers are central to many statistical procedures. Roughly speaking, they say that as the sample size n grows, the sample mean $\bar{X}_n$ converges to the expectation $\mathbb{E}[X_1]$. This is why the sample mean is often used as an *estimator* of the expectation.

Bonus Material

First, we need to define what convergence means for random variables.

Definition 3.21 (Convergence of Random Variables)**.** Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of random variables and X another random variable. We say that

1. X_n converges to X almost surely, if

$$\mathbb{P}\left(\left\{\omega \in \Omega \mid \lim_{n \to \infty} X_n(\omega) = X(\omega)\right\}\right) = 1,$$

   and we write $X_n \xrightarrow{a.s.} X$ as $n \to \infty$.

2. X_n converges to X in probability, if for any $\varepsilon > 0$ we have

$$\lim_{n \to \infty} \mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0$$

   and we write $X_n \xrightarrow{\mathbb{P}} X$ as $n \to \infty$.

3. X_n converges to X in distribution, if for all continuity points x of F_X we have

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

   and we write $X_n \xrightarrow{\mathcal{D}} X$ as $n \to \infty$.

   Without proving it here, it is worth noting that as $n \to \infty$,

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{\mathcal{D}} X.$$

   We can now proceed to derive the laws of the large numbers.

Theorem 3.14 (Laws of Large Numbers)**.** *Let $X_1, ..., X_n$ be i.i.d. random variables with finite first moment $\mu := \mathbb{E}[X_1]$ and finite second moment $\sigma^2 := \mathrm{Var}(X_1)$. Then the weak law of large numbers (WLLN) states that as $n \to \infty$*

$$\bar{X}_n \xrightarrow{\mathbb{P}} \mu$$

and the strong law of large numbers (SLLN) states that as $n \to \infty$

$$\bar{X}_n \xrightarrow{a.s.} \mu.$$

From a statistical perspective, Theorem 3.14 means that the estimator $\bar{X}_n$ is *consistent*. It is noteworthy that $\bar{X}_n$ is also *unbiased*, because $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X_1]$.

3.6.1 Central Limit Theorem

We will now come to a slightly more advanced result, called the *Central Limit Theorem* (CLT). Consider the setting of Theorem 3.14. The law of large numbers states that the sum $S_n := \sum_{i=1}^n X_i$ with the scaling factor of $1/n$ *converges* to a fixed number (the expected value). This means that $1/n$ is small enough to make sure that S_n/n does not blow up, but simultaneously is not too small, which would make it converge to zero. The main idea of the CLT is that "interchanging $1/n$ with a slightly larger factor of $1/\sqrt{n}$ will make $S_n/\sqrt{n}$ also converge, but this time in distribution to a *normally distributed random variable*". What is so astonishing about this fact, is, that we make only very weak assumptions on the distribution of $X_1, ..., X_n$!

Bonus Material

Theorem 3.15 (Central Limit Theorem by Lindeberg-Lévy)**.** *Let* $X_1, ..., X_n$ *be i.i.d. random variables with finite first and second moments, i.e., their expectation* μ *and variance* σ^2 *exists. Then it holds that*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) = \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} Z$$

where Z *is a random variable with normal distribution* $\mathcal{N}(0, \sigma^2)$.

Example 16 (Central Limit Theorem for Binomial Distribution)**.** Consider the case of $X_1, ..., X_n$ being i.i.d. $\text{Ber}(p)$ random variables, i.e., $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = p$, $\mathbb{E}[X_i] = p$ and $\text{Var}(X_i) = p(1-p)$. Then $S_n = \sum_{i=1}^n X_i$ is a random variable that follows the binomial distribution with parameters n and p. Therefore, S_n has expectation np and variance $np(1-p)$. By Theorem 3.15, for large enough n, this distribution should roughly be a normal distribution. Specifically,

$$\frac{1}{\sqrt{n}} S_n - \sqrt{n}p$$

should roughly be $\mathcal{N}(0, p(1-p))$ distributed. By rearranging, we see that S_n should roughly be $\mathcal{N}(np, np(1-p))$ distributed. Figure 3.3 visualises this phenomenon.

3.7 Useful Inequalities and Lemmas

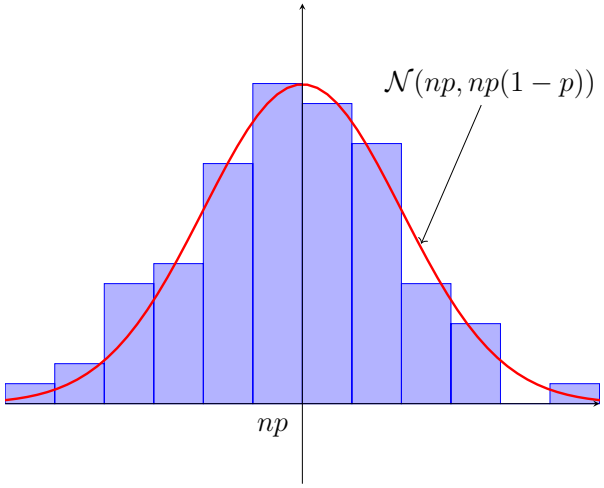

Figure 3.3: The central limit Theorem for the binomial distribution with parameters n and p. The figure shows the histogram (blue) of 100 realisations with $n = 100$ and $p = 0.5$. The respective normal distribution in red.

Lemma 3.16 (Jensen's Inequality). *Let X be a random variable with finite first moment and $g : \mathbb{R} \to \mathbb{R}$ a convex function. Then*

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

Lemma 3.17 (Chebychev-Markov Inequality). *Let X be a random variable and $g : [0, \infty) \to [0, \infty)$ an increasing function with $g(x) > 0$ for all $x > 0$. Then it holds for all $\varepsilon > 0$ that*

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbb{E}[g(|X|)]}{g(\varepsilon)}.$$

Proof. It holds that

$$\begin{aligned} g(\varepsilon)\mathbb{P}(|X| \geq \varepsilon) &= \int g(\varepsilon)\mathbb{I}_{\{|X| \geq \varepsilon\}} d\mathbb{P} \\ &\leq \int g(|X|)\mathbb{I}_{\{|X| \geq \varepsilon\}} d\mathbb{P} \\ &\leq \int g(|X|) d\mathbb{P} \\ &= \mathbb{E}[g(|X|)] \end{aligned}$$

where we used that g is increasing in the second inequality and $g \geq 0$ in the last inequality. □

Bibliography

[Dur19] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge University Press, 2019.

[Kow73] Charles J. Kowalski. "Non-Normal Bivariate Distributions with Normal Marginals". In: *The American Statistician* 27.3 (1973), pp. 103–106.

[Ros14] Sheldon M Ross. *A first course in probability*. Pearson, 2014.

[TB97] Lloyd N. Trefethen and David Bau. *Numerical linear algebra*. SIAM, 1997.

[Wik22] Wikipedia. *Big O notation*. Feb. 2022. URL: https://en.wikipedia.org/wiki/Big_O_notation.

[Wil91] David Williams. *Probability with Martingales*. Cambridge mathematical textbooks. Cambridge University Press, 1991.