Compendium of Mathematics & Physics

Nicolas Dewolf

October 10, 2021

Contents

Chapter 1

Introduction

1.1 Goals

This compendium originated out of the necessity for a compact summary of important theorems and formulas during physics and mathematics classes at university. When the interest in more (and more exotic) subjects grew, this collection lost its compactness and became the chaos it now is. Although there should exist some kind of overall structure, it was not always possible to keep every section self-contained or respect the order of the chapters.

It should definitely not be used as a formal introduction to any subject. It is neither a complete work nor a fact-checked one, so the usefulness and correctness is not guaranteed. However it should be used as lookup table for theorems and formulas and as a guide to the literature. To this end each chapter begins with a list of useful references. At the same time only a small number of statements are proven in the text (or appendices). This was done to keep the text as short as possible. However, in some cases the major ideas underlying the proofs are given.

1.2 Conventions

Definitions, properties and formulas marked by a dagger symbol † are explained and/or derived in one of the appendices. This has been done such that the summary itself contains only core notions and theorems. Sections and statements that require more advanced concepts, in particular concepts from later chapters or (higher) category theory, will be labelled by a *clubs* symbol ♣.

Definitions of words in the middle of a text will be indicated by the use of **bold font**. Notions that have not been defined in this summary but that are relevant or that will be defined further on in the compendium are indicated by *italic text*. Names of authors will also be written in *italic*.

Objects from a general category will be denoted by a lower case letter (depending on the context we might also use upper case for clarity), functors will be denoted by upper case letters and the categories themselves will be denoted by symbols in **bold font**. In the later chapters on physics we will often adopt specific conventions for the different types of vectors. Vectors in Euclidean space will be denoted by a bold font letter with an arrow above, e.g. $\vec{\boldsymbol{a}}$. Vectors in Minkowski space (4-vectors) and differential forms will be written without the arrow, e.g. $\mathbf{a}$. Matrices and tensors will always be represented by capital letters and dependent on the context a specific font will be adopted.

Chapter 2

Measure Theory and Lebesgue Integration

The main references for this chapter are [20, 37].

2.1 Measure theory

2.1.1 General definitions

Definition 2.1.1 (Measure). Let X be a set and let Σ be a σ-algebra over X. A function $\mu : \Sigma \to \overline{\mathbb{R}}$ is called a measure if it satisfies the following conditions:

1. **Nonnegativity**: $\forall E \in \Sigma : \mu(E) \geq 0$,
2. **Empty set is null**: $\mu(\emptyset) = 0$, and
3. **σ-additivity**: $\forall i \neq j : E_i \cap E_j = \emptyset \implies \mu(\bigcup_{n=1}^{\infty} E_n) = \sum_{i=n}^{\infty} \mu(E_n)$.

When μ only satisfies countable subadditivity, i.e. the equality in the last condition becomes an inequality $\leq$, it is called an **outer measure**.

Remark 2.1.2. To show that two measures coincide on a σ-algebra, it suffices to show that they coincide on the generating sets and apply the monotone class theorem **??**.

Definition 2.1.3 (Measure space). The pair (X, Σ) is called a measurable space and the triple (X, Σ, μ) is called a measure space. The elements $E \in \Sigma$ are called **measurable sets**.

Definition 2.1.4 (Null set). A set $A \subset \mathbb{R}$ is said to be null if $\mu(A) = 0$.

Definition 2.1.5 (Almost everywhere[1]). Let (X, Σ, μ) be a measure space. A property P is said to hold on X almost everywhere (abbreviated as **a.e.**) if it satisfies the following equation:

$$\mu\big(\{x \in X \,|\, \neg P(x)\}\big) = 0, \tag{2.1}$$

i.e. it holds everywhere except for a null set.

Definition 2.1.6 (Complete measure space). A measure space (X, Σ, μ) is said to be complete if for every $E \in \Sigma$ with $\mu(E) = 0$ the implication $A \subset E \implies A \in \Sigma$ holds. Additivity then necessarily implies that $\mu(A) = 0$.

Definition 2.1.7 (Completion). Let $\mathcal{F} \subseteq \mathcal{G}$ be σ-algebras over a set X. $(X, \mathcal{G}, \overline{\mu})$ is called the completion of $(X, \mathcal{F}, \mu)$ if:

[1] In probability theory this is often called **almost surely**.

1. $\forall A \in \mathcal{F} : \overline{\mu}(A) = \mu(A)$,
2. $(X, \mathcal{G}, \overline{\mu})$ is complete, and
3. $\mathcal{G}$ is the smallest σ-algebra for which the foregoing conditions hold.

Definition 2.1.8 (Borel measure). Consider a topological space together with its Borel σ-algebra **??**. Any measure defined on this measurable space is called a Borel measure.

Definition 2.1.9 (Regular measure). Let μ be a measure on a measurable space (X, Σ). It is called a regular measure if it satisfies the following equations for every measurable set B:

$$\mu(B) = \inf\left\{\mu(O) \,\middle|\, O \text{ open and measurable}, O \supset B\right\} \tag{2.2}$$

$$\mu(B) = \sup\left\{\mu(F) \,\middle|\, F \text{ compact and measurable}, F \subset B\right\}. \tag{2.3}$$

A Borel regular measure can also be characterized as a Borel measure such that for every subset $A \subseteq X$ there exists a Borel set B with $A \subseteq B$ and $\inf\left\{\mu(E) \,\middle|\, A \subseteq E \in \Sigma\right\} = \mu(B)$.

Definition 2.1.10 (Radon measure). A Borel measure on a Hausdorff space that is outer regular, inner regular on open sets and **locally finite**, i.e. every point has a neighbourhood of finite measure. When restricted to locally-compact Hausdorff spaces, this is equivalent to requiring that every compact subset has finite measure.

Definition 2.1.11 (σ-finite measure). Let (X, Σ, μ) be a measure space. The measure μ is said to be σ-finite if there exists a sequence $(A_n)_{n\in\mathbb{N}}$ of measurable sets such that $\bigcup_{n=1}^{\infty} A_n = X$ with $\forall n \in \mathbb{N} : \mu(A_n) < \infty$.

2.1.2 Lebesgue measure

Definition 2.1.12 (Lebesgue outer measure). Let $X \subseteq \mathbb{R}$ be a set. The (Lebesgue) outer measure of X is defined as follows:

$$\lambda^*(X) := \inf\left\{\sum_{n=1}^{\infty} l(I_n) \,\middle|\, (I_n)_{n\in\mathbb{N}} \text{ a sequence of open intervals that covers } X\right\}. \tag{2.4}$$

Property 2.1.13 (Intervals). The outer measure of an interval I equals its length: $\lambda^*(I) = l(I)$.

Property 2.1.14 (Translation). The outer measure is translation-invariant: $\lambda^*(A + t) = \lambda^*(A)$ for all $A \subset \mathbb{R}$ and $t \in \mathbb{R}$.

Property 2.1.15. The Lebesgue outer measure is an outer measure in the sense of Definition 2.1.1.

Theorem 2.1.16 (Carathéodory's criterion). *Let X be a subset of $\mathbb{R}$. If X satisfies the following equation, it is said to be **Lebesgue measurable**:*

$$\forall A \subseteq \mathbb{R} : \lambda^*(A) = \lambda^*(A \cap X) + \lambda^*(A \cap X^c). \tag{2.5}$$

*The collection of all Lebesgue-measurable sets is denoted by $\mathcal{M}$ and the outer measure $\lambda^*(X)$, now denoted by λ, is called the **Lebesgue measure** of X.*

Construction 2.1.17 (Carathéodory's extension theorem). In fact, the above construction can be generalized to arbitrary sets. Every outer measure μ^* gives rise to a σ-algebra consisting of those sets that satisfy Carathéodory's criterion 2.1.16 with respect to μ^*. Furthermore, consider a **premeasure** μ_0, i.e. a σ-additive function defined on an algebra of sets **??** such that $\mu_0(\emptyset) = 0$. Definition 2.1.12 can be used to define an outer measure μ^* in terms of the premeasure μ_0 by replacing intervals with elements from the given algebra of sets. The σ-algebra generated by this outer measure contains the given algebra of sets and μ^* restricts to μ_0. This shows that any premeasure can be extended to a genuine measure. Moreover, it can be shown that this measure is complete.

Corollary 2.1.18. The Lebesgue σ-algebra $\mathcal{M}$ is the completion of the Borel σ-algebra $\mathcal{B}$. (This is how the Lebesgue σ-algebra was introduced historically.)

Property 2.1.19. Any countable set is null with respect to the Lebesgue outer measure.

Property 2.1.20. The Lebesgue measure is a regular Borel measure. For every $A \subset \mathbb{R}$ there exists a sequence $(O_n)_{n\in\mathbb{N}}$ of open sets such that

$$A \subset \bigcap_{n=1}^{\infty} O_n \qquad \text{and} \qquad \lambda\left(\bigcap_{n=1}^{\infty} O_n\right) = \lambda^*(A), \tag{2.6}$$

and for every $E \in \mathcal{M}$ there exists a sequence $(F_n)_{n\in\mathbb{N}}$ of closed sets such that

$$\bigcup_{n=1}^{\infty} F_n \subset E \qquad \text{and} \qquad \lambda\left(\bigcup_{n=1}^{\infty} F_n\right) = \lambda(E). \tag{2.7}$$

Property 2.1.21. Consider a set $A \subset \mathbb{R}$. $A \in \mathcal{M}$ if and only if for every $\varepsilon > 0$ there exist an open set $O \supset A$ and a closed set $F \subset A$ such that $\lambda^*(O\backslash A) < \varepsilon$ and $\lambda^*(A\backslash F) < \varepsilon$.

Property 2.1.22. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of sets in $\mathcal{M}$. The following two properties apply:

$$\forall i \in \mathbb{N} : A_i \subseteq A_{i+1} \implies \lambda\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \lambda(A_n) \tag{2.8}$$

$$\forall i \in \mathbb{N} : A_i \supseteq A_{i+1} \wedge \lambda(A_1) < \infty \implies \lambda\left(\bigcap_{i=n}^{\infty} A_n\right) = \lim_{n\to\infty} \lambda(A_n). \tag{2.9}$$

Remark 2.1.23. This property is valid for every σ-additive set function.

Construction 2.1.24 (Restriction). Let $A \in \mathcal{M}$ have nonzero measure. The restriction of the Lebesgue measure to the set B is defined as follows:

$$\mathcal{M}_A := \{A \cap B \mid B \in \mathcal{M}\} \qquad \text{and} \qquad \forall E \in \mathcal{M}_A : \lambda_A(E) := \lambda(E). \tag{2.10}$$

It can be shown that the measure space $(A, \mathcal{M}_A, \lambda_A)$ is complete.

2.1.3 Measurable functions

Definition 2.1.25 (Measurable function). Consider two measurable spaces (X, Σ_X) and (Y, Σ_Y). A function $f : X \to Y$ is said to be measurable if for every measurable set $A \in \Sigma_Y$ the preimage $f^{-1}(A)$ is also measurable. Equivalently, the σ-algebra generated by the preimages of measurable sets in Σ_Y should be a sub-σ-algebra of Σ_X.

Two important examples are given below:

Example 2.1.26 (Borel-measurable function). A continuous function $f : X \to Y$ such that for every open set $O \in \mathcal{B}_Y : f^{-1}(O) \in \mathcal{B}_X$.

Example 2.1.27 (Lebesgue-measurable function). A function $f : \mathbb{R} \to \mathbb{R}$ such that for every interval $I \subset \mathbb{R} : f^{-1}(I) \in \mathcal{M}$.

Remark 2.1.28. The inclusion $\mathcal{B} \subset \mathcal{M}$ implies that every Borel-measurable function is also Lebesgue-measurable.

Property 2.1.29. The class of Borel/Lebesgue-measurable functions defined on $E \in \mathcal{M}$ forms an algebra.

Example 2.1.30. The following types of functions are Lebesgue-measurable:

- monotonic functions,
- continuous functions,
- indicator functions, and
- compositions of measurable functions.

Corollary 2.1.31. Let f, g be (Lebesgue-)measurable functions and let $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a continuous function. The composition $F(f(x), g(x))$ is also measurable.

Property 2.1.32. Let f be a Lebesgue-measurable function. The level set $\{x \mid f(x) = a\}$ is measurable for all $a \in \mathbb{R}$.

Property 2.1.33. Define the following functions (which are measurable if f is measurable as a result of the previous properties):

$$f^+(x) := \max(f, 0) = \begin{cases} f(x) & f(x) > 0 \\ 0 & f(x) \leq 0, \end{cases} \tag{2.11}$$

$$f^-(x) := \max(-f, 0) = \begin{cases} 0 & f(x) > 0 \\ -f(x) & f(x) \leq 0. \end{cases} \tag{2.12}$$

The function $f : \mathbb{R} \to \mathbb{R}$ is measurable if and only if both f^+ and f^- are measurable. Furthermore, f is measurable if $|f|$ is measurable (the converse is false in general).

Definition 2.1.34 (Pushforward). Consider two measurable spaces (X_1, Σ_1) and (X_2, Σ_2) together with a measurable function $f : X_1 \to X_2$. For every measure μ on X_1 one can define the pushforward measure $f_*\mu$ on X_2 as follows:

$$f_*\mu(A) := \mu\big(f^{-1}(A)\big). \tag{2.13}$$

Definition 2.1.35 (Measure-preserving function). Let (X, Σ, μ) be a measure space and consider a measurable function $T : X \to X$. T is said to be measure-preserving if it satisfies

$$\mu\big(T^{-1}(A)\big) = \mu(A) \tag{2.14}$$

for all $A \in \Sigma$. This can also be concisely written as $T_*\mu = \mu$. These functions form the morphisms in the category **Meas** of measure spaces.[2]

Definition 2.1.36 (Ergodic function). Let (X, Σ, μ) be a measure space and consider a measure-preserving function $T : X \to X$. It is said to be ergodic if the following condition is satisfied:

$$T(A) = A \implies \mu(A) = 0 \vee \mu(X\backslash A) = 0. \tag{2.15}$$

This is equivalent to stating that for every set $A \in \Sigma$ with positive measure the following condition holds:

$$\mu\left(\bigcup_{n=1}^{\infty} T^{-n}(A)\right) = 1. \tag{2.16}$$

[2] The notation **Meas** is also sometimes used to denote the larger category of measurable spaces and measurable functions.

Property 2.1.37. Consider a topological space X with Borel σ-algebra $\mathcal{B}$ and let T be an ergodic function. Almost every T-orbit is dense in the support of μ.

Definition 2.1.38 (Mixing). An endomorphism of a measure spaces (X, Σ, μ) is said to be mixing if for all measurable spaces A, B the following equality holds:

$$\lim_{n\to\infty} \mu\big(T^{-n}(A) \cap B\big) = \mu(A)\mu(B) \tag{2.17}$$

Property 2.1.39. All mixing transformations are ergodic.

Property 2.1.40 (Additivity). Every measurable, additive function $f : \mathbb{R} \to \mathbb{R}$ is linear.

Corollary 2.1.41. From the basic properties of exponential and logarithmic functions, the following results can be obtained:

- Let $f : \mathbb{R} \to \mathbb{R}$ be a measurable function. If $f(x+y) = f(x)f(y)$, then $f(x) = e^{\lambda x}$ for some $\lambda \in \mathbb{R}$.
- Let $f : [0,\infty] \to \mathbb{R}$ be a measurable function. If $f(xy) = f(x) + f(y)$, then $f(x) = \lambda \log(x)$ for some $\lambda \in \mathbb{R}$.
- Let $f : [0,\infty] \to [0,\infty]$ be a measurable function. If $f(xy) = f(x)f(y)$, then $f(x) = x^\lambda$ for some $\lambda \in \mathbb{R}$.

2.1.4 Limit operations

Property 2.1.42. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions. The following functions are also measurable:

- $\min\limits_{i\leq k}(f_i)$ and $\max\limits_{i\leq k}(f_i)$
- $\inf\limits_{n\in\mathbb{N}}(f_n)$ and $\sup\limits_{n\in\mathbb{N}}(f_n)$
- $\liminf\limits_{n\to\infty}(f_n)$ and $\limsup\limits_{n\to\infty}(f_n)$

Property 2.1.43. If f is a measurable function and g is a function such that $f = g$ almost everywhere, then g is measurable as well.

Corollary 2.1.44. As a result of the previous two properties, if a sequence of measurable functions converges pointwise a.e., the limit is also a measurable function.

Definition 2.1.45 (Essential supremum).

$$\operatorname{ess\,sup}(f) := \inf\{z \in \mathbb{R} \mid f \leq z \text{ a.e.}\} \tag{2.18}$$

Definition 2.1.46 (Essential infimum).

$$\operatorname{ess\,inf}(f) := \sup\{z \in \mathbb{R} \mid f \geq z \text{ a.e.}\} \tag{2.19}$$

Property 2.1.47. Every measurable function f satisfies the following inequalities:

- $f \leq \operatorname{ess\,sup}(f)$ a.e. and $f \geq \operatorname{ess\,inf}(f)$ a.e.
- $\operatorname{ess\,sup}(f) \leq \sup(f)$ and $\operatorname{ess\,inf}(f) \geq \inf(f)$.

The latter pair of inequalities becomes a pair of equalities if f is continuous.

Property 2.1.48. If f, g are measurable functions, then $\operatorname{ess\,sup}(f+g) \leq \operatorname{ess\,sup}(f) + \operatorname{ess\,sup}(g)$. An analogous inequality holds for the essential infimum.

Definition 2.1.49 (Weak convergence). A sequence of measures $(\mu_n)_{n\in\mathbb{N}}$ is said to converge weakly to a measure μ on a metrizable space X if any of the following conditions is satisfied:

1. $\int_X f\,d\mu_n \longrightarrow \int_X f\,d\mu$ for all bounded, continuous functions f.
2. $\mu_n(A) \longrightarrow \mu(A)$ for all **continuity sets** A of μ, i.e. for all Borel sets A such that $\mu(\partial A) = 0$.
3. $\liminf \mu_n(U) \geq \mu(U)$ for all open sets U.
4. $\limsup \mu_n(V) \leq \mu(V)$ for all closed sets V.

If $X = \mathbb{R}$ with its canonical topology, the sequence $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ if and only if $\mu_n(\{x \in \mathbb{R} : x \leq y\}) \longrightarrow \mu(\{x \in \mathbb{R} : x \leq y\})$ for all points $y \in \mathbb{R}$ where these functions are continuous.

2.2 Lebesgue integral

2.2.1 Simple functions

Definition 2.2.1 (Indicator function).

$$\mathbb{1}_A(x) := \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases} \tag{2.20}$$

Definition 2.2.2 (Simple function). A function $f : X \to \mathbb{R}$ on a measurable space (X, Σ) that can be expressed as

$$f(x) = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(x) \tag{2.21}$$

for some $\{a_i \geq 0\}_{i\leq n}, \{A_i\}_{i\leq n} \subset \Sigma$ and $n \in \mathbb{N}$.

Definition 2.2.3 (Step function). If $(X, \Sigma) = (\mathbb{R}, \mathcal{M})$ and the sets A_i are intervals, the above function is often called a step function.

Definition 2.2.4 (Lebesgue integral of simple functions). Consider a simple function φ on a measure space (X, Σ, μ). The Lebesgue integral of φ over a measurable set $A \in \Sigma$ with respect to μ is given by

$$\int_A \varphi\,d\mu := \sum_{i=1}^{n} a_i \mu(A \cap A_i). \tag{2.22}$$

As usual, if the domain of integration is not mentioned explicitly, an integral over the whole space X is implied.

Example 2.2.5. Let $\mathbb{1}_{\mathbb{Q}}$ be the indicator function of the rational numbers. Contrary to the case of Riemann integrals, the above definition makes it possible to integrate the rational indicator function over the real line:

$$\int_{\mathbb{R}} \mathbb{1}_{\mathbb{Q}}\,d\lambda = 1 \times \lambda(\mathbb{Q}) + 0 \times \lambda(\mathbb{R}\backslash\mathbb{Q}) = 0, \tag{2.23}$$

where the measure of the rational numbers is 0 because it is a countable set (Corollary 2.1.19).

2.2.2 Measurable functions

Definition 2.2.6 (Integral for nonnegative functions). The definition for simple functions can be generalized to nonnegative measurable functions f as follows:

$$\int_A f\,d\mu := \sup\left\{\int_A \varphi\,d\mu \,\middle|\, \varphi \text{ a simple function such that } \varphi \leq f\right\}. \tag{2.24}$$

This integral is always nonnegative.

Formula 2.2.7. The following equality allows to change the domain of integrals:

$$\int_A f\,d\mu = \int_X f\mathbb{1}_A\,d\mu. \tag{2.25}$$

Property 2.2.8. The Lebesgue integral over a null set is 0.

Theorem 2.2.9 (Mean value theorem). *If $a \leq f(x) \leq b$, then $a\lambda(A) \leq \int_A f\,d\lambda \leq b\lambda(A)$.*

Property 2.2.10. Let f be a nonnegative measurable function. There exists an increasing sequence $(\varphi_n)_{n\in\mathbb{N}}$ of simple functions such that $\varphi_n \nearrow f$. Moreover, if f is bounded on $A \in \Sigma$, the sequence can be chosen to be uniformly convergent on A.

2.2.3 Integrable functions

Definition 2.2.11 (Integrable function). Let A be a measurable subset of a measure space (X, Σ, μ). A measurable function f is said to be integrable over A if both $\int_A f^+\,d\mu$ and $\int_A f^-\,d\mu$ are finite. The Lebesgue integral of f over A is then defined as

$$\int_A f\,d\mu := \int_A f^+\,d\mu - \int_A f^-\,d\mu. \tag{2.26}$$

If only one of the functions f^+, f^- is finite, f is said to be **quasi-integrable**.

Property 2.2.12 (Absolute integrability). f is integrable if and only if $|f|$ is integrable. Furthermore, $\int_A |f|\,d\mu = \int_A f^+\,d\mu + \int_A f^-\,d\mu$.

Property 2.2.13. Let f, g be integrable functions on a measure space (X, Σ, μ). The following important properties hold:

- **Linearity**: $\int_A (f + \lambda g) d\mu = \int_A f\,d\mu / \lambda \int_A g\,d\mu$ for all $\lambda \in \mathbb{R}$
- **Monotonicity**: $f \leq g$ a.e. implies $\int_A f\,d\mu \leq \int_A g\,d\mu$ and $\forall A \in \Sigma : \int_A f\,d\mu \leq \int_A g\,d\mu \implies f \leq g$ a.e.
- **Finiteness**: f is finite a.e.
- $|\int_A f\,d\mu| \leq \int_A |f|\,d\mu$.
- $\int_A f\,d\mu = 0, \forall A \in \Sigma \implies f = 0$ a.e.

Definition 2.2.14 (Lebesgue integrable functions). The set of integrable functions over a set $A \in \mathcal{M}$ forms the vector space $\mathcal{L}^1(A)$.

Property 2.2.15. Let $f \in \mathcal{L}^1$ and $\varepsilon > 0$. There exists a continuous (or step or even simple) function g, vanishing outside a finite (or even compact) set, such that $\int |f - g|\,d\mu < \varepsilon$.

Definition 2.2.16 (Locally integrable function). A measurable function is said to be locally integrable if it is integrable on every compact subset of its domain. The space of locally integrable functions is denoted by $\mathcal{L}^1_{\text{loc}}$.

Example 2.2.17. All continuous functions are locally integrable.

Property 2.2.18 (Absolute continuity). Let $f \geq 0$ be a measurable function. The mapping $A \mapsto \int_A f\,d\mu$ defines a measure that is σ-finite if f is locally integrable and finite if f is integrable. Furthermore, this measure is said to be absolutely continuous (with respect to μ). See Section 2.5 for a generalization to arbitrary measures.

2.2.4 Convergence theorems

Theorem 2.2.19 (Fatou's lemma). *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative measurable functions.*

$$\int_A \left(\liminf_{n\to\infty} f_n\right) d\mu \leq \liminf_{n\to\infty} \int_A f_n\,d\mu \tag{2.27}$$

Theorem 2.2.20 (Monotone convergence). *Let A be measurable and let $(f_n)_{n\in\mathbb{N}}$ be an increasing sequence of nonnegative measurable functions such that $f_n \nearrow f$ pointwise a.e.*

$$\int_A f\,d\mu = \lim_{n\to\infty} \int_A f_n\,d\mu. \tag{2.28}$$

Method 2.2.21. To prove results concerning integrable functions in spaces such as $\mathcal{L}^1$ it is often useful to proceed as follows:

1. Verify that the property holds for indicator functions. (This often follows by definition.)
2. Use linearity to extend the property to simple functions.
3. Apply the monotone convergence theorem to show that the property holds for all nonnegative measurable functions.
4. Extend the property to all integrable functions by decomposing $f = f^+ - f^-$ and applying linearity again.

Theorem 2.2.22 (Dominated convergence). *Let A be measurable set and consider a sequence of measurable functions $(f_n)_{n\in\mathbb{N}}$ such that $\forall n \in \mathbb{N} : |f_n| \leq g$ a.e. for some function $g \in \mathcal{L}^1(A)$. If $f_n \to f$ pointwise a.e., then f is integrable over A and*

$$\int_A f\,d\mu = \lim_{n\to\infty} \int_A f_n\,d\mu. \tag{2.29}$$

Property 2.2.23. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative measurable functions

$$\int_A \sum_{n=1}^{\infty} f_n\,d\mu = \sum_{n=1}^{\infty} \int_A f_n\,d\mu. \tag{2.30}$$

One cannot conclude that the right-hand side is finite a.e., so the series on the left-hand side need not be integrable.

Theorem 2.2.24 (Beppo Levi[3]). *Suppose that*

$$\sum_{i=1}^{\infty} \int_A |f_n|\,d\mu$$

is finite. The series $\sum_{i=1}^{\infty} f_n(x)$ converges a.e. Furthermore, the series is integrable and

$$\int_A \sum_{i=1}^{\infty} f_n\,d\mu = \sum_{i=1}^{\infty} \int_A f_n\,d\mu. \tag{2.31}$$

[3] Various other theorems and variants of this theorem can be found in the literature under the same name.

Theorem 2.2.25 (Riemann-Lebesgue lemma). *Let $f \in \mathcal{L}^1(\mathbb{R})$. The sequences*

$$s_k = \int_{-\infty}^{\infty} f(x)\sin(kx)dx$$

and

$$c_k = \int_{-\infty}^{\infty} f(x)\cos(kx)dx$$

both converge to 0.

Theorem 2.2.26 (Birkhoff ergodicity). *Let (X, Σ, μ) be a measure space and let T be a μ-ergodic map. For every measurable function f and for μ-almost every element $x \in X$ the integral of f can be computed as an average over the orbit of x:*

$$\lim_{n\to\infty} \frac{1}{n+1}\sum_{t=0}^{n} f(T^n(x)) = \int f\, d\mu. \tag{2.32}$$

2.2.5 Relation to the Riemann integral

Property 2.2.27. Let $f : [a,b] \to \mathbb{R}$ be a bounded function.

- f is Riemann-integrable if and only if f is continuous a.e. with respect to the Lebesgue measure on $[a,b]$, i.e. the set of discontinuities of f has measure zero.
- Riemann-integrable functions on $[a,b]$ are integrable with respect to the Lebesgue measure on $[a,b]$ and the integrals coincide.

Property 2.2.28. If $f \geq 0$ and the improper Riemann integral **??** exists, the Lebesgue integral $\int_{\mathbb{R}} f\, d\mu$ exists and the two integrals coincide. Note that positivity of f is required here. Because the Lebesgue integral is absolute 2.2.12, positive and negative parts cannot cancel (Lebesgue integrals can never be conditionally convergent).

The following definition should be compared to 2.2.1 and **??**.

Definition 2.2.29 (Dirac measure). Define the Dirac measure as follows:

$$\delta_a(A) := \begin{cases} 1 & a \in A \\ 0 & a \notin A. \end{cases} \tag{2.33}$$

Integration with respect to the Dirac measure has the following important property:

$$\int f\, d\delta_a = f(a). \tag{2.34}$$

2.3 Space of integrable functions

2.3.1 Distance

To define a distance between functions, a notion of the length of a function is introduced first. Normally this would not be a problem, one could use the integral of a function to define a norm. However, the fact that two functions differing on a null set have the same integral carries problems with it: a nonzero function could have a zero length. To avoid this issue one quotients out these degenerate functions::

Definition 2.3.1 (L^1-space). Define the set of equivalence classes $L^1 = \mathcal{L}^1_{/\equiv}$ by introducing the following equivalence relation: $f \equiv g$ if and only if $f = g$ a.e.

Property 2.3.2. L^1 is a Banach space **??**. The norm on L^1 is given by

$$\|f\|_1 := \int |f|\, d\mu. \tag{2.35}$$

2.3.2 Hilbert space L^2

Property 2.3.3. L^2 is a Hilbert space **??**. The norm on L^2 is given by

$$\|f\|_2 := \left(\int |f|^2\,d\mu\right)^{\frac{1}{2}}. \tag{2.36}$$

This norm is induced by the following inner product:

$$\langle f|g\rangle := \int \overline{f}g\,d\mu. \tag{2.37}$$

Formula 2.3.4 (Cauchy-Schwarz inequality). Let $f,g\in L^2(X,\mathbb{C})$. Formula 2.3.7 implies that $fg\in L^1(X,\mathbb{C})$ and

$$\left|\int \overline{f}g\,d\mu\right| \leq \|fg\|_1 \leq \|f\|_2\|g\|_2. \tag{2.38}$$

2.3.3 L^p-spaces

Generalizing the previous two function classes leads to the notion of L^p-spaces with the following norm:

Formula 2.3.5. For all $1\leq p\leq\infty$, $L^p(X)$ is a Banach space when equipped with the following norm:

$$\|f\|_p := \left(\int_X |f|^p\,d\mu\right)^{\frac{1}{p}}. \tag{2.39}$$

Remark 2.3.6. Note that L^2 is the only L^p-space that is also a Hilbert space. The other L^p-spaces do not have a norm induced by an inner product.

Formula 2.3.7 (Hölder's inequality). Let $\frac{1}{p}+\frac{1}{q}=1$ with $p\geq 1$ (numbers satisfying this equality are called **Hölder conjugates**). For every $f\in L^p$ and $g\in L^q$ one has that

$$\|fg\|_1 \leq \|f\|_p\|g\|_q. \tag{2.40}$$

This also implies that $fg\in L^1$.

Formula 2.3.8 (Minkowski's inequality). For every $p\geq 1$ and $f,g\in L^p$ one has that

$$\|f+g\|_p \leq \|f\|_p+\|g\|_p. \tag{2.41}$$

This also implies that $f+g\in L^p$.

Property 2.3.9 (Inclusions). $L^1(X)\cap L^\infty(X)\subset L^2(X)$. Moreover, if X has finite measure, then $L^q(X)\subset L^p(X)$ whenever $1\leq p\leq q<\infty$.

Using the Hölder inequality one can prove the following property:

Property 2.3.10. Let p,q be Hölder conjugates. The spaces L^p and L^q are topological duals, i.e. every function $f\in L^p$ can be identified (one-to-one) with a continuous functional on L^q.

Definition 2.3.11 (Essentially bounded function). Let f be a measurable function satisfying $\operatorname{ess\,sup}|f|<\infty$. The function f is said to be essentially bounded and the set of all such functions is denoted by L^∞ (again after quotienting out all functions that are equal a.e.).

Formula 2.3.12. A norm on L^∞ is given by

$$\|f\|_\infty := \operatorname{ess\,sup}|f|. \tag{2.42}$$

This norm is called the **supremum norm** and it induces the supremum metric **??**.

Property 2.3.13. Equipped with the above norm the space L^∞ becomes a Banach space.

2.4 Product measures

2.4.1 Real hyperspace

The notions of intervals and lengths from the one-dimensional case can be generalized to higher dimensions in the following way:

Definition 2.4.1 (Hypercube). Let $I_1, \dots, I_n$ be a sequence of intervals. The hypercube spanned by them is defined as the following set:

$$\mathbf{I} := I_1 \times \cdots \times I_n. \tag{2.43}$$

Definition 2.4.2 (Generalized length). Let $\mathbf{I}$ be a hypercube induced by the set of intervals $I_1, \dots, I_n$. The generalized length (or **volume**) of $\mathbf{I}$ is defined as

$$l(\mathbf{I}) := \prod_{i=1}^{n} l(I_i). \tag{2.44}$$

2.4.2 Construction of the product measure

The general condition for multi-dimensional Lebesgue measures is given by the following equation that should hold for all $A_1 \in \Sigma_1$ and $A_2 \in \Sigma_2$:

$$\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2). \tag{2.45}$$

Definition 2.4.3 (Section). Let $A = A_1 \times A_2$. The following two sets are called sections:

$$\begin{aligned} A_{\omega_1} &:= \{\omega_2 \in X_2 \mid (\omega_1, \omega_2) \in A\} \subset \Sigma_2, \\ A_{\omega_2} &:= \{\omega_1 \in X_1 \mid (\omega_1, \omega_2) \in A\} \subset \Sigma_1. \end{aligned}$$

The following property follows immediately from the definition of product σ-algebras **??**:

Property 2.4.4. Let $\Sigma = \Sigma_1 \times \Sigma_2$. If $A \in \Sigma$, then $A_{\omega_1} \in \Sigma_2$ for each ω_1 and $A_{\omega_2} \in \Sigma_1$ for each ω_2. Equivalently, the sets $\mathcal{G}_1 = \{A \in \Sigma \mid \forall \omega_1 \in X_1 : A_{\omega_1} \in \Sigma_2\}$ and $\mathcal{G}_2 = \{A \in \Sigma \mid \forall \omega_2 \in X_2 : A_{\omega_2} \in \Sigma_1\}$ coincide with the product σ-algebra Σ.

Property 2.4.5. The function $A_{\omega_2} \mapsto \mu(A_{\omega_2})$ is a step function:

$$\mu(A_{\omega_2}) = \begin{cases} \mu_1(A_1) & \omega_2 \in A_2 \\ 0 & \omega_2 \notin A_2. \end{cases}$$

Formula 2.4.6 (Product measure). From the previous property it follows that the product measure $\mu(A)$ can be written in the following way:

$$\mu(A) = \int_{\Omega_2} \mu_1(A_{\omega_2}) d\mu_2(\omega_2). \tag{2.46}$$

Property 2.4.7. Let μ_1, μ_2 be finite measures. If $A \in \Sigma$, the functions

$$\omega_1 \mapsto \mu_2(A_{\omega_1}) \qquad \text{and} \qquad \omega_2 \mapsto \mu_1(A_{\omega_2})$$

are measurable with respect to Σ_1 and Σ_2 respectively and

$$\int_{\Omega_2} \mu_1(A_{\omega_2}) d\mu_2(\omega_2) = \int_{\Omega_1} \mu_2(A_{\omega_1}) d\mu_1(\omega_1). \tag{2.47}$$

Furthermore, the set function μ is countably additive and if any other product measure coincides with μ on all rectangles, it coincides with μ on the whole product σ-algebra.

2.4.3 Fubini's theorem

Property 2.4.8. Let $f : X_1 \times X_2 \to \mathbb{R}$ be a nonnegative function. If f is measurable with respect to $\Sigma_1 \times \Sigma_2$, then for each $\omega_1 \in X$ the function $\omega_2 \mapsto f(\omega_1, \omega_2)$ is measurable with respect to Σ_2 (and vice versa). Their integrals with respect to μ_1 and μ_2 respectively are also measurable.

Definition 2.4.9 (Section). The functions $\omega_1 \mapsto f(\omega_1, \omega_2)$ and $\omega_2 \mapsto f(\omega_1, \omega_2)$ are called sections of f.

Theorem 2.4.10 (Tonelli). *Let $f : X_1 \times X_2 \to \mathbb{R}$ be a nonnegative function. The following equalities hold:*

$$\begin{aligned}\int_{X_1\times X_2} f(\omega_1,\omega_2)d\mu(\omega_1,\omega_2) &= \int_{X_1}\left(\int_{X_2} f(\omega_1,\omega_2)d\mu_2(\omega_2)\right)d\mu_1(\omega_1)\\ &= \int_{X_2}\left(\int_{X_1} f(\omega_1,\omega_2)d\mu_1(\omega_1)\right)d\mu_2(\omega_2). \qquad (2.48)\end{aligned}$$

Corollary 2.4.11 (Fubini). Let $f \in L^1(X_1 \times X_2)$. The sections of f are integrable in the appropriate spaces. Furthermore, the functions $\omega_1 \mapsto \int_{X_2} f d\mu_2$ and $\omega_2 \mapsto \int_{X_1} f d\mu_1$ are in $L^1(\Omega_1)$ and $L^1(\Omega_2)$ respectively and Tonelli's theorem holds.

Remark 2.4.12. The previous construction and theorems also apply to higher-dimensional product spaces. These theorems provide a way to construct higher-dimensional Lebesgue measures by defining them as the completion of the product of multiple one-dimensional Lebesgue measures.

2.5 Radon-Nikodym theorem

Definition 2.5.1 (Absolute continuity). Let (X, Σ) be a measurable space and let μ, ν be two measures defined on this space. Then ν is said to be absolutely continuous with respect to μ if

$$\forall A \in \Sigma : \mu(A) = 0 \implies \nu(A) = 0. \qquad (2.49)$$

This relation is often denoted by $\nu \ll \mu$.

The following property relates the notion of absolute continuity above with that of Definition **??**:

Property 2.5.2 (Absolute continuity). Let μ, ν be finite measures on a measurable space (X, Σ). Then $\nu \ll \mu$ if and only if

$$\forall \varepsilon > 0 : \exists \delta > 0 : \forall A \in \Sigma : \mu(A) < \delta \implies \nu(A) < \varepsilon. \qquad (2.50)$$

Definition 2.5.3 (Singular measures). Consider two measures μ, ν. If there exists a set A such that $\mu(A) = 0 = \nu(A^c)$, they are said to be singular (or **orthogonal**). This is denoted by $\mu \perp \nu$.

Theorem 2.5.4 (Lebesgue's decomposition theorem). *Let μ, ν be two σ-finite measures. There exist two other σ-finite measures ν_a, ν_s such that $\nu = \nu_a + \nu_s$, where $\nu_a \ll \mu$ and $\nu_s \perp \mu$.*

Definition 2.5.5 (Dominated measure). Let μ, ν be two measures defined on a measurable space (X, Σ). Then μ is said to **dominate** ν if $0 \leq \nu(F) \leq \mu(F)$ for every $F \in \Sigma$.

Theorem 2.5.6 (Radon-Nikodym theorem for dominated measures). *Let μ be a finite measure on a measurable space (X, Σ) and let ν be a measure dominated by μ. There exists a nonnegative, measurable function f such that $\nu(A) = \int_A f\,d\mu$ for all $A \in \Sigma$.*

Definition 2.5.7 (Radon-Nikodym derivative). The function f in the previous theorem is called the Radon-Nikodym derivative of ν with respect to μ. It is generally denoted by $\frac{d\nu}{d\mu}$.

Theorem 2.5.8 (Radon-Nikodym theorem). *Let (X, Σ) be a measurable space and let μ, ν be two σ-finite measures defined on Σ such that $\nu \ll \mu$. There exists a nonnegative, measurable function $f : X \rightarrow \mathbb{R}$ such that $\nu(A) = \int_A f\,d\mu$ for all $A \in \Sigma$.*

Remark 2.5.9. The function f in this theorem is unique up to a μ-null (and thus ν-null) set.

Property 2.5.10. In general the Radon-Nikodym derivative is not integrable (unless the measures are finite). However, it is always locally integrable 2.2.16. Together with Property 2.2.18 this implies that (densities of) absolutely continuous measures are in bijection with locally integrable functions.

Property 2.5.11 (Change of variables). Let μ, ν be finite measures such that $\nu \ll \mu$ and let $\frac{d\nu}{d\mu}$ be the associated Radon-Nikodym derivative. For every ν-integrable function f the following equality holds

$$\int_A f\,d\nu = \int_A f h_\nu\,d\mu \tag{2.51}$$

for all $A \in \Sigma$.

Property 2.5.12. Let λ, ν and μ be σ-finite measures. If $\lambda \ll \mu$ and $\nu \ll \mu$, the following two properties hold:

- **Linearity**: $\dfrac{d(\lambda + \nu)}{d\mu} = \dfrac{d\lambda}{d\mu} + \dfrac{d\lambda}{d\mu}$.
- **Chain rule**: If $\lambda \ll \nu$, then $\dfrac{d\lambda}{d\mu} = \dfrac{d\lambda}{d\nu}\dfrac{d\nu}{d\mu}$ a.e.

2.6 Lebesgue-Stieltjes integral

Aside from the Lebesgue measure one can construct some other important measures (and their associated integrals) on the Borel σ-algebra of the real line $\mathbb{R}$. These constructions will be important in the study of density functions in probability theory. To this end consider a function F that is right-continuous, i.e. $F(x^+) = F(x)$, and monotonically increasing. The length of an interval can be generalized in the following way:

Definition 2.6.1 (F-length). The F-length of an interval $]a, b]$ is defined as follows:

$$l_F\big(]a, b]\big) := F(b) - F(a). \tag{2.52}$$

The restriction to half-open intervals assures that this function is additive when taking unions of intervals. The footnote in Definition **??** also assures that the σ-algebra generated by these intervals is the Borel σ-algebra on $\mathbb{R}$.

An immediate extension of Definition 2.1.12 gives the outer measure associated to F:

Definition 2.6.2 (F-outer measure). Let $X \subseteq \mathbb{R}$ be a set. The F-outer measure of X is defined as follows:

$$\mu_F^*(X) := \inf\left\{\sum_{n=1}^{\infty} l_F(I_n)\ \middle|\ (I_n)_{n\in\mathbb{N}} \text{ a sequence of half-open intervals that cover } X\right\}. \tag{2.53}$$

Using this outer measure one can define the μ_F-measurable sets as those sets satisfying Carathéodory's criterion (with respect to μ_F^*). The main difference with the Lebesgue measure is that μ_F is not necessarily translation-invariant and that singletons are not necessarily null:

Property 2.6.3 (Singletons). The F-measure of a singleton $\{x\}$ is equal to the jump of F at x:

$$\mu_F\big(\{x\}\big) = F(x) - F(x^-). \tag{2.54}$$

Such elements are examples of **atoms**, sets of positive measure for which every proper measurable subset is null.

Corollary 2.6.4. It follows that the Lebesgue-Stieltjes measures having null singletons are exactly those for which F is continuous.

Property 2.6.5 (Regularity). The Lebesgue-Stieltjes measure is a regular Borel measure. Furthermore, every (finite) regular Borel measure μ on $\mathbb{R}$ is equal to a Lebesgue-Stieltjes measure induced by the function

$$F(x) = \mu\big(]-\infty, x]\big). \tag{2.55}$$

Example 2.6.6 (Lebesgue measure). The Lebesgue measure is the Lebesgue-Stieltjes measure associated to $F(x) = x$.

Example 2.6.7 (Dirac measure). The Dirac measure at $x \in \mathbb{R}$ can be obtained as the Lebesgue-Stieltjes measure for $F(x) = \mathbb{1}_{[x,\infty[}$.

2.6.1 Signed measures

Definition 2.6.8 (Signed measure). Consider a measurable space (X, Σ). A function $\mu : \Sigma \to \overline{\mathbb{R}}$ is called a signed measure if it satisfies the following conditions:

1. **Measure zero**: $\mu(\emptyset) = 0$, and
2. **Countable additivity** : $\forall i \neq j : E_i \cap E_j = \emptyset \implies \mu(\bigcup_{n=1}^\infty E_n) = \sum_{i=n}^\infty \mu(E_n)$.

Note that these requirements are the same as for an ordinary measure 2.1.1, except that now the function is allowed to become negative. The function is, however, not allowed to attain $-\infty$ to exclude undefined expressions such as $\infty - \infty$.

Remark 2.6.9. An important consequence of this generalization is that signed measures are not necessarily monotonic, i.e. $A \subseteq B \not\Longrightarrow \mu(A) \leq \mu(B)$. In fact this is a strict relation. A signed measure is monotonic if and only if it is a genuine measure.

Definition 2.6.10 (Total variation). Consider a signed measure μ on a measurable space (X, Σ). The total variation $|\mu|$ is the measure defined as follows:

$$|\mu|(A) := \sup\left\{\sum_{P\in\mathcal{P}} |\mu(P)| \,\middle|\, \mathcal{P} \subset \Sigma, \mathcal{P} \text{ covers } A\right\}. \tag{2.56}$$

Using this measure one can decompose the signed measure μ as a difference of two genuine measures:

$$\begin{aligned}\mu &= \mu^+ - \mu^- \\ &= \frac{1}{2}(|\mu| + \mu) + \frac{1}{2}(|\mu| - \mu).\end{aligned} \tag{2.57}$$

Furthermore, this decomposition is minimal in the sense that if $\mu = \lambda_1 - \lambda_2$ for any two measures, then $\mu^+ \leq \lambda_1$ and $\mu^- \leq \lambda_2$.

The following theorem generalizes both the Radon-Nikodym and Lebesgue decomposition theorems to the case of signed measures:

Theorem 2.6.11. *Consider a σ-finite signed measure μ and a σ-finite measure ν on a measurable space (X, Σ). There exists a ν-a.e. unique integrable function $f \in L^1(\nu)$ and a σ-finite measure $\mu_s \perp \nu$ such that for all $A \in \Sigma$:*

$$\mu(A) = \int_A f d\nu + \mu_s(A). \tag{2.58}$$

As before, the function is called f the Radon-Nikodym derivative of μ.

Theorem 2.6.12 (Hahn-Jordan). *Consider a signed measure μ on a measurable space (X, Σ). There exists a set $A \in \Sigma$ such that the minimal decomposition $\mu = \mu^+ - \mu^-$ in terms of two measures $\mu^\pm$ is given by*

$$\mu^+(B) = \mu(A \cap B) \qquad\qquad \mu^-(B) = \mu(A^c \cap B). \tag{2.59}$$

Definition 2.6.13 (Integral with respect to a signed measure). Let μ be a signed measure on a measurable space (X, Σ) and consider a measurable function f on $A \in \Sigma$. The integral of f with respect to μ is defined as follows:

$$\int_A f\, d\mu := \int_A f\, d\mu^+ - \int_A f\, d\mu^-. \tag{2.60}$$

Definition 2.6.14 (Lebesgue-Stieltjes signed measure). Let F be a function of bounded variation. According to Property **??** it can be written as $F = F_1 - F_2$, where F_1, F_2 are monotonically increasing, absolutely continuous functions. The Lebesgue-Stieltjes (signed) measure associated to F is defined as $\mu_F := \mu_{F_1} - \mu_{F_2}$.

Theorem 2.6.15 (Fundamental theorem of calculus). *Let F be an absolutely continuous function on the closed interval $[a, b]$. Then F is differentiable λ-a.e. (λ being the Lebesgue measure) and its associated Lebesgue-Stieltjes measure μ_F has Radon-Nikodym derivative $\frac{d\mu_F}{d\lambda} = F'$ λ-a.e. Furthermore, for all $x \in [a, b]$ one has*

$$F(x) - F(a) = \mu_F([a, x]) = \int_a^x F'(t) dt. \tag{2.61}$$

Corollary 2.6.16. If F is absolutely continuous and $F' = 0$ λ-a.e., then F is constant.

Chapter 3

Probability

The majority of this chapter uses the language of measure theory. For an introduction see Chapter 2. The section on *imprecise probabilities* is mainly based on [1].

3.1 Probability

The Kolmogorov axioms of probability state when a set admits the definition of a probability theory:

Definition 3.1.1 (Kolmogorov axioms). A probability space (Ω, Σ, P) is a measure space 2.1.3 with finite measure $P(X) = 1$. The set Ω is called the **sample space**.

Definition 3.1.2 (Random variable). Let (Ω, Σ, P) be a probability space. A function $X : \Omega \to \mathbb{R}$ is called a random variable if $\forall a \in \mathbb{R} : X^{-1}\big([a, \infty[\big) = \{\omega \in \Omega \mid X(\omega) \geq a\} \in \Sigma$.

Definition 3.1.3 (σ-algebra of a random variable). Let X be a random variable defined on a probability space (Ω, Σ, P). The following family of sets is a σ-algebra:

$$X^{-1}(\mathcal{B}) := \{S \in \Sigma \mid \exists B \in \mathcal{B} : S = X^{-1}(B)\}. \tag{3.1}$$

Notation 3.1.4. The σ-algebra generated by the random variable X is often denoted by $\mathcal{F}_X$, analogous to **??**.

Definition 3.1.5 (Event). Let (Ω, Σ, P) be a probability space. An element S of the σ-algebra Σ is called an event.

From this definition it is clear that a single possible outcome of a measurement can be a part of multiple events. So, although only one outcome can occur at the same time, multiple events can occur simultaneously.

Remark. The Kolmogorov axioms use the σ-algebra **??** of events instead of the power set **??** of all events. Intuitively this seems to mean that some possible outcomes are not treated as events. However, one can make sure that the σ-algebra still contains all ‘‘useful’’ events by using a ‘‘nice’’ definition of probability spaces.

Formula 3.1.6 (Union). Let A, B be two events. The probability that at least one of them occurs is given by the following formula:

$$P(A \cup B) = P(A) + P(B) + P(A \cap B). \tag{3.2}$$

Definition 3.1.7 (Disjoint events). Two events A and B are said to be disjoint if they cannot happen at the same time:

$$P(A \cap B) = 0. \tag{3.3}$$

Corollary 3.1.8. If A and B are disjoint, the probability that both A and B occur is just the sum of their individual probabilities.

Formula 3.1.9 (Complement). Let A be an event. The probability of A being false is denoted as $P(\overline{A})$ and is given by

$$P(\overline{A}) = 1 - P(A). \tag{3.4}$$

Corollary 3.1.10. From the previous equation and de Morgan's laws (**??**) and (**??**), one can derive the following formula:

$$P(\overline{A} \cap \overline{B}) = 1 - P(A \cup B). \tag{3.5}$$

3.2 Conditional probability

Definition 3.2.1 (Conditional probability). Let A, B be two events. The probability of A given that B is true is denoted as $P(A|B)$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.6}$$

By interchanging A and B in previous equation and by observing that this has no effect on the quantity $P(A \cap B)$ the following important result can be derived:

Theorem 3.2.2 (Bayes). *Let A, B be two events.*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{3.7}$$

Formula 3.2.3. Let $(B_n)_{n\in\mathbb{N}}$ be a sequence of pairwise disjoint events. If $\bigsqcup_{n=1}^{\infty} B_n = \Omega$, the total probability of a given event A can be calculated as follows:

$$P(A) = \sum_{n=1}^{\infty} P(A|B_n)P(B_n). \tag{3.8}$$

Definition 3.2.4 (Independent events). Let A, B be two events. A and B are said to be independent if they satisfy the following relation:

$$P(A \cap B) = P(A)P(B). \tag{3.9}$$

Corollary 3.2.5. If A and B are two independent events, Bayes's theorem simplifies to

$$P(A|B) = P(A). \tag{3.10}$$

The above definition can be generalized to multiple events:

Definition 3.2.6. The events $A_1, \ldots, A_n$ are said to be independent if for each choice of k events the probability of their intersection is equal to the product of their individual probabilities.

This definition can be stated in terms of σ-algebras:

Definition 3.2.7 (Independence). The σ-algebras $\mathcal{F}_1, \ldots, \mathcal{F}_n$ defined on a probability space $(\Omega, \mathcal{F}, P)$ are said to be independent if for all choices of distinct indices $i_1, \ldots, i_k$ and for all choices of sets $F_{i_n} \in \mathcal{F}_{i_n}$ the following equation holds:

$$P(F_{i_1} \cap \cdots \cap F_{i_k}) = P(F_{i_1}) \cdots P(F_{i_k}). \tag{3.11}$$

Corollary 3.2.8. Let X, Y be two random variables. X and Y are independent if the σ-algebras generated by them are independent.

3.3 Probability distribution

Definition 3.3.1 (Probability distribution). Let X be a random variable defined on a probability space (Ω, Σ, P). The following function is a measure on the Borel σ-algebra of $\mathbb{R}$:

$$P_X(B) = P(X^{-1}(B)). \tag{3.12}$$

This measure is called the probability distribution of X.

Definition 3.3.2 (Density). Let $f \geq 0$ be an integrable function and recall Property 2.2.18. The function f is called the density of the measure $P(A) := \int_A f\, d\lambda$ (with respect to the Lebesgue measure λ). Measures of this form are often called **cumulative distribution functions** and denoted by F. More generally, by the Radon-Nikodym theorem from Section 2.5, every absolutely continuous distribution function F is of the form

$$F(A) = \int_A f\, d\lambda \tag{3.13}$$

for some integrable function f.

Theorem 3.3.3 (Skorokhod's representation theorem). *Let $F : \mathbb{R} \to [0,1]$ be a function that satisfies the following three properties:*

- *F is nondecreasing.*
- $\lim\limits_{x\to-\infty} F(x) = 0$ *and* $\lim\limits_{x\to\infty} F(x) = 1$.
- *F is right-continuous, i.e.* $\lim\limits_{y\nearrow y_0} F(y) = F(y_0)$.

There exists a random variable $X : [0,1] \to \mathbb{R}$ defined on the probability space $([0,1], \mathcal{B}, m_{[0,1]})$ such that $F = F_X$.

Theorem 3.3.4 (Theorem of the unconscious statistician). *Consider a random variable X on a probability space (Ω, Σ, P). The following equality holds for every integrable function $g \in L^1(\mathbb{R})$:*

$$\int_\Omega g \circ X\, dP = \int_\mathbb{R} g(x) dP_X(x). \tag{3.14}$$

Remark 3.3.5. The name of this theorem stems from the fact that many scientists take this equality to be a definition of the expectation value $\mathrm{E}[g(X)]$. However, this equality should be proven since the measure on the right-hand side is the one belonging to the random variable X and not $g(X)$.

Formula 3.3.6. Consider an absolutely continuous probability function F defined on $\mathbb{R}^n$ and let f be the associated density. Let $g : \mathbb{R}^n \to \mathbb{R}$ be integrable with respect to F.

$$\int_{\mathbb{R}^n} g\, dF = \int_{\mathbb{R}^n} f(x) g(x) dx \tag{3.15}$$

Corollary 3.3.7. The previous formula together with Theorem 3.3.4 gives rise to

$$\int_\Omega g \circ X\, dP = \int_{\mathbb{R}^n} f_X(x) g(x) dx. \tag{3.16}$$

Formula 3.3.8. Let X be a random variable with density function f_X and let $g : \mathbb{R} \to \mathbb{R}$ be smooth and strictly monotone. The random variable $g \circ X$ has an associated density f_g given by

$$f_g(y) = f(g^{-1}(y)) \left| \frac{dg^{-1}}{dy}(y) \right|. \tag{3.17}$$

Weak convergence of measures 2.1.49 induces a notion of convergence of random variables:

Definition 3.3.9 (Convergence in distribution). A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is said to converge in distribution to a random variable Y if the associated distribution functions F_{X_n} converge pointwise to F_Y, i.e. $\lim_{n\to\infty} F_{X_n}(x) = F_Y(x)$ for all $x \in \mathbb{R}$ where F is continuous.

Notation 3.3.10. If a sequence $(X_n)_{n\in\mathbb{N}}$ converges in distribution to a random variable Y, this is often denoted by $X_n \xrightarrow{d} Y$. Sometimes the d (for "distribution") is replaced by the $\mathcal{L}$ (for "law").

Theorem 3.3.11 (Slutsky). *Let $(X_n)_{n\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}}$ be two sequences of random variables converging in probability to a random variable X and a constant c, respectively. The following statements hold:*

- $X_n + Y_n \xrightarrow{d} X + c,$
- $X_nY_n \xrightarrow{d} cX,$ *and*
- $X_n/Y_n \xrightarrow{d} X/c.$

Definition 3.3.12 (Giry monad ♣). Consider the category **Meas** of measurable spaces. On this space one can define a monad **??** that sends a set X to its collection of probability distributions equipped with the σ-algebra generated by all evaluation maps ev_U, where U runs over the measurable subsets of X.

The unit of the Giry monad $\mathbb{P}$ is defined by assigning Dirac measures:

$$\eta_X(x) := \delta_x. \tag{3.18}$$

The multiplication map is defined as follows:

$$\mu_X(Q)(U) := \int_{P\in\mathbb{P}X} \mathrm{ev}_U(P)\, dQ. \tag{3.19}$$

3.4 Moments

3.4.1 Expectation value

Definition 3.4.1 (Expectation value). Let X be random variable defined on a probability space (Ω, Σ, P).

$$\mathrm{E}[X] := \int_\Omega X\, dP \tag{3.20}$$

Notation 3.4.2. Other notations that are common in the literature are $\langle X\rangle$ and μ_X.

Definition 3.4.3 (Moment of order r). The moment of order r is defined as the expectation value of the r^{th} power of X. By Equation (3.16) this becomes

$$\mathrm{E}[X^r] = \int_\mathbb{R} x^r f_X(x)dx. \tag{3.21}$$

Definition 3.4.4 (Central moment of order r).

$$\mathrm{E}[(X-\mu)^r] = \int_\mathbb{R} (x-\mu)^r f_X(x)dx \tag{3.22}$$

Remark 3.4.5. Moments of order n are determined by central moments of order $k \leq n$ and, conversely, central moments of order n are determined by moments of order $k \leq n$.

Definition 3.4.6 (Variance). The central moment of order 2 is called the variance:

$$\mathrm{Var}[X] := \mathrm{E}\big[(X-\mu)^2\big]. \tag{3.23}$$

Definition 3.4.7 (Standard deviation).

$$\sigma_X := \sqrt{V[X]} \tag{3.24}$$

Property 3.4.8. If $\mathrm{E}[|X|^n]$ is finite for $n > 0$, then $\mathrm{E}\big[X^k\big]$ exist and is finite for all $k \leq n$.

Definition 3.4.9 (Moment generating function).

$$M_X(t) := \mathrm{E}\big[e^{tX}\big] = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx \tag{3.25}$$

Property 3.4.10. If the moment generating function exists, the moments $\mathrm{E}[X^n]$ can be expressed in terms of M_X (using the series expansion of the exponential function):

$$\mathrm{E}[X^n] = \left.\frac{d^n M_X(t)}{dt^n}\right|_{t=0}. \tag{3.26}$$

Definition 3.4.11 (Characteristic function).

$$\varphi_X(t) := \mathrm{E}\big[e^{itX}\big] \tag{3.27}$$

Property 3.4.12. The characteristic function has the following properties:

- $\varphi_X(0) = 1$,
- $|\varphi_X(t)| \leq 1$, and
- $\varphi_{aX+b}(t) = e^{itb}\varphi_X(at)$ for all $a, b \in \mathbb{R}$.

Formula 3.4.13. If $\varphi_X(t)$ is k times continuously differentiable, then X has a finite k^{th} moment and

$$\mathrm{E}\Big[X^k\Big] = \frac{1}{i^k}\frac{d^k}{dt^k}\varphi_X(0). \tag{3.28}$$

Conversely, if X has a finite k^{th} moment, then $\varphi_X(t)$ is k times continuously differentiable and the above formula holds.

Formula 3.4.14 (Inversion formula). Let X be a random variable. If the CDF of X is continuous at $a, b \in \mathbb{R}$, then

$$F_X(b) - F_X(a) = \lim_{c\to\infty} \frac{1}{2\pi}\int_{-c}^{c} \frac{e^{-ita} - e^{-itb}}{it}\varphi_X(t)dt. \tag{3.29}$$

Formula 3.4.15. If $\varphi_X(t)$ is integrable, the CDF is given by:

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_X(t)dt. \tag{3.30}$$

Remark 3.4.16. This formula implies that the density function and the characteristic function form a Fourier transform pair.

3.4.2 Correlation

Property 3.4.17. Two random variables X, Y are independent if and only if $\mathrm{E}[f(X)g(Y)] = \mathrm{E}[f(X)]\mathrm{E}[g(Y)]$ holds for all Borel-measurable bounded functions f, g.

The value $\mathrm{E}[XY]$ is equal to the inner product $\langle X|Y\rangle$ as defined in (2.37). It follows that independence of random variables implies orthogonality. To generalize this concept, the following notions are introduced:

Definition 3.4.18 (Centred random variable). Let X be a random variable with finite expectation value $\mathrm{E}[X]$. The centred random variable X_c is defined as $X_c = X - \mathrm{E}[X]$.

Definition 3.4.19 (Covariance). The covariance of two random variables X, Y is defined as follows:

$$\mathrm{cov}(X, Y) := \langle X_c|Y_c\rangle = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]. \tag{3.31}$$

Some basic math gives

$$\mathrm{cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]. \tag{3.32}$$

Definition 3.4.20 (Correlation). The correlation of two random variables X, Y is defined as the cosine of the angle between X_c and Y_c:

$$\rho_{XY} := \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{3.33}$$

Corollary 3.4.21. From Theorem 3.4.17 it follows that independent random variables are uncorrelated.

Corollary 3.4.22. If the random variables X and Y are uncorrelated, they satisfy $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

Formula 3.4.23 (Bienaymé formula). Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent (or uncorrelated) random variables. Their variances satisfy the following equation:

$$\mathrm{Var}\left[\sum_{i=1}^{\infty} X_i\right] = \sum_{i=1}^{\infty} \mathrm{Var}[X_i]. \tag{3.34}$$

3.4.3 Conditional expectation

Let (Ω, Σ, P) be a probability space. Consider a random variable $X \in L^2(\Omega, \Sigma, P)$ and a sub-σ-algebra $\mathcal{G} \subset \Sigma$. Property 2.3.3 implies that the spaces $L^2(\Sigma)$ and $L^2(\mathcal{G})$ are complete and, hence, the projection theorem **??** can be applied. For every $X \in L^2(\Sigma)$ there exists a random variable $Y \in L^2(\mathcal{G})$ such that $X - Y$ is orthogonal to $L^2(\mathcal{G})$. This has the following result:

$$\forall Z \in L^2(\mathcal{G}) : \langle X - Y|Z\rangle \equiv \int_\Omega (X - Y)Z dP = 0. \tag{3.35}$$

Since $\mathbb{1}_G \in L^2(\mathcal{G})$ for every $G \in \mathcal{G}$, Equation (2.25) can be rewritten as

$$\int_G X\, dP = \int_G Y\, dP \tag{3.36}$$

for all $G \in \mathcal{G}$. This leads to the introduction of the following definition:

Definition 3.4.24 (Conditional expectation). Let (Ω, Σ, P) be a probability space and let $\mathcal{G}$ be a sub-σ-algebra of Σ. For every Σ-measurable random variable $X \in L^2(\Sigma)$ there exists a unique (up to a null set) random variable $Y \in L^2(\mathcal{G})$ that satisfies Equation (3.36) for every $G \in \mathcal{G}$. This variable Y is called the conditional expectation of X given $\mathcal{G}$ and it is denoted by $\mathrm{E}[X|\mathcal{G}]$:

$$\int_G \mathrm{E}[X|\mathcal{G}]\, dP = \int_G X\, dP. \tag{3.37}$$

Remark 3.4.25. Although this construction was based on orthogonal projections, one could as well have used the (signed) Radon-Nikodym theorem 2.6.11 since $G \mapsto \int_G X\, dP$ is absolutely continuous with respect to $P|_{\mathcal{G}}$.

Property 3.4.26. Let (Ω, Σ, P) be a probability space and consider a sub-σ-algebra $\mathcal{G} \subset \Sigma$. If the random variable X is $\mathcal{G}$-measurable, then

$$\mathrm{E}[X|\mathcal{G}] = X \text{ a.s.} \tag{3.38}$$

On the other hand, if X is independent of $\mathcal{G}$, then

$$\mathrm{E}[X|\mathcal{G}] = \mathrm{E}[X] \text{ a.s.} \tag{3.39}$$

3.5 Joint distributions

Definition 3.5.1 (Joint distribution). Let X, Y be two random variables defined on the same probability space (Ω, Σ, P) and consider the vector random variable $(X, Y) : \Omega \to \mathbb{R}^2$. The distribution of (X, Y) isa probability measure defined on the Borel algebra of $\mathbb{R}^2$ defined by

$$P_{(X,Y)}(B) = P((X,Y)^{-1}(B)). \tag{3.40}$$

Definition 3.5.2 (Joint density). If the probability measure from the previous definition can be written as

$$P_{(X,Y)}(B) = \int_B f_{(X,Y)}(x,y) dx dy \tag{3.41}$$

for some integrable $f_{(X,Y)}$, it is said that X and Y have a joint density.

Definition 3.5.3 (Marginal distribution). The distributions of the one-dimensional random variables is determined by the joint distribution:

$$P_X(A) = P_{(X,Y)}(A \times \mathbb{R}) \tag{3.42}$$

$$P_Y(A) = P_{(X,Y)}(\mathbb{R} \times A). \tag{3.43}$$

Corollary 3.5.4. If the joint density exists, the marginal distributions are absolutely continuous and the associated density functions are given by

$$f_X(x) = \int_{\mathbb{R}} f_{(X,Y)}(x,y) dy \tag{3.44}$$

$$f_Y(y) = \int_{\mathbb{R}} f_{(X,Y)}(x,y) dx. \tag{3.45}$$

The converse, however, is not always true. The one-dimensional distributions can be absolutely continuous without the existence of a joint density.

Property 3.5.5 (Independence). Let X, Y be two random variables with joint distribution $P_{(X,Y)}$. X and Y are independent if and only if the joint distribution coincides with the product measure:

$$P_{(X,Y)} = P_X \otimes P_Y. \tag{3.46}$$

If X and Y are absolutely continuous, the previous properties also applies to the densities instead of the distributions.

Formula 3.5.6 (Sum of random variables). Consider two independent random variables X, Y and let $Z = X + Y$ denote their sum. The density f_Z is given by the following convolution:

$$f_Z(z) := f * g(z) = \int_{-\infty}^{\infty} g(x)h(z-x)dx = \int_{-\infty}^{\infty} g(z-y)h(y)dy, \tag{3.47}$$

where g, h denote the densities of X, Y respectively.

Formula 3.5.7 (Product of random variables). Consider two independent random variables X, Y and let $Z = XY$ denote their product. The density f_Z is given by

$$f_Z(z) = \int_{-\infty}^{\infty} g(x)h(z/x)\frac{dx}{|x|} = \int_{-\infty}^{\infty} g(z/y)h(y)\frac{dy}{|y|}, \tag{3.48}$$

where g, h denote the densities of X, Y respectively.

Corollary 3.5.8. Taking the Mellin transform **??** of both the positive and negative part of the above integrand (to be able to handle the absolute value) gives the following relation:

$$\mathcal{M}\{f\} = \mathcal{M}\{g\}\mathcal{M}\{h\}. \tag{3.49}$$

Formula 3.5.9 (Conditional density). Let X, Y be two random variables with joint density $f_{(X,Y)}$. The conditional density of Y given $X \in A$ is

$$h(y|X \in A) = \frac{\int_A f_{(X,Y)}(x,y)dx}{\int_A f_X(x)dx}. \tag{3.50}$$

For $X = \{a\}$ this equation is ill-defined since the denominator would become 0. However, it is possible to avoid this problem by formally setting

$$h(y|A = a) := \frac{f_{(X,Y)}(a,y)}{f_X(a)}, \tag{3.51}$$

where $f_X(a) \neq 0$. This last condition is nonrestrictive because the probability of having a measurement $(X,Y) \in \{(x,y) \mid f_X(x) = 0\}$ is 0 (for nonsingular measures). One can thus define the conditional probability of Y given $X = a$ as follows:

$$P(Y \in B|X = a) := \int_B h(y|X = a)dy. \tag{3.52}$$

Formula 3.5.10 (Conditional expectation).

$$\mathrm{E}[Y|X](\omega) = \int_{\mathbb{R}} yh(y|X(\omega))dy \tag{3.53}$$

Let $\mathcal{F}_X$ denote the σ-algebra generated by the random variable X as before. Using Fubini's theorem one can prove that for all sets $A \in \mathcal{F}_X$ the following equality holds:

$$\int_A \mathrm{E}[Y|X]\,dP = \int_A Y\,dP. \tag{3.54}$$

This implies that the conditional expectation $\mathrm{E}[Y|X]$ on $\mathcal{F}_X$ coincides with Definition 3.4.24.

Applying Property 3.4.26 to the case $\mathcal{G} = \mathcal{F}_X$ gives the law of total expectation:

Property 3.5.11 (Law of total expectation).

$$\mathrm{E}[\mathrm{E}[Y|X]] = \mathrm{E}[Y] \tag{3.55}$$

Theorem 3.5.12 (Bayes's theorem). *The conditional density can be computed without prior knowledge of the joint density:*

$$g(x|y) = \frac{h(y|x)f_X(x)}{f_Y(y)}. \tag{3.56}$$

3.6 Stochastic calculus

Definition 3.6.1 (Stochastic process). A sequence of random variables $(X_t)_{t\in T}$ for some index set T. In practice T will often be a totally ordered set, e.g. $(\mathbb{R}, \leq)$ in the case of a time series. This will be assumed from here on.

Definition 3.6.2 (Filtered probability space). Consider a probability space (Ω, Σ, P) together with a filtration **??** of Σ, i.e. a collection of σ-algebras $\mathbb{F} \equiv (\mathbb{F}_t)_{t\in T}$, such that $i \leq j \implies \mathbb{F}_i \subseteq \mathbb{F}_j$. The quadruple $(\Omega, \Sigma, \mathbb{F}, P)$ is called a filtered probability space.

Often the filtration is required to be exhaustive and separated (where $\emptyset$ is replaced by $\mathbb{F}_0 = \{\emptyset, \Omega\}$ since any σ-algebra has to contain the total space).

Definition 3.6.3 (Adapted process). A stochastic process $(X_t)_{t\in T}$ on a filtered probability space $(\Omega, \Sigma, \mathbb{F}, P)$ is said to be adapted to the filtration $\mathbb{F}$ if X_t is $\mathbb{F}_t$-measurable for all $t \in T$.

Definition 3.6.4 (Predictable process). A stochastic process $(X_t)_{t\in T}$ on a filtered probability space $(\Omega, \Sigma, \mathbb{F}, P)$ is said to be predictable if X_{t+1} is $\mathbb{F}_t$-measurable for all $t \in T$.

Definition 3.6.5 (Stopping time). Consider a random variable τ on filtered probability space $(\Omega, \Sigma, \mathbb{F}, P)$ where the codomain of τ coincides with the index set of $\mathbb{F}$. This variable is called a stopping time for $\mathbb{F}$ if

$$\{\tau \leq t\} \in \mathbb{F}_t \tag{3.57}$$

for all t. The stopping time is a "time indicator" that only depends on the knowledge of the process up to time $t \in T$.

3.6.1 Martingales

From here on the index set T will be $\mathbb{R}_+ \equiv [0, \infty[$ so that the index t can be interpreted as a true time parameter. The discrete case $T = \mathbb{N}$ can be obtained as the restriction of most definitions or properties and, if necessary, this will be made explicit.

Definition 3.6.6 (Martingale). Consider a filtered probability space $(\Omega, \Sigma, \mathbb{F}, P)$. A stochastic process $(X_t)_{t\in T}$ is called a martingale relative to $\mathbb{F}$ if it satisfies the following conditions:

1. $(X_t)_{t\in T}$ is adapted to $\mathbb{F}$.
2. Each random variable X_t is integrable, i.e. $X_t \in L^1(P)$ for all $t \geq 0$.
3. For all $t > s \geq 0 : \mathrm{E}[X_t \mid \mathbb{F}_s] = X_s$.

If the equality in the last condition is replaced by the inequality $\leq$ (resp. $\geq$), the stochastic process is called a **supermartingale** (resp. **submartingale**).

Property 3.6.7 (Doob-Ville inequality). Consider a càdlàg submartinagle $(X_t)_{t\in T}$.

$$P\left(\sup_{t\leq\tau} X_t \geq C\right) \leq \frac{\mathrm{E}[\max(0, X_\tau)]}{C} \tag{3.58}$$

for all $C \geq 1$.

Theorem 3.6.8 (Doob decomposition). *Any integrable adapted process $(X_t)_{t\in T}$ can be decomposed as $X_t = X_0 + M_t + A_t$, where $(M_t)_{t\in T}$ is a martingale and $(A_t)_{t\in T}$ is a predictable process. These two processes are constructed iteratively as follows:*

$$A_0 = 0 \qquad\qquad M_0 = 0 \tag{3.59}$$

$$\Delta A_t = \mathrm{E}[\Delta X_t \mid \mathbb{F}_{t-1}] \qquad\qquad \Delta M_t = \Delta X_t - \Delta A_t. \tag{3.60}$$

Furthermore, $(X_t)_{t\in T}$ is a submartingale if and only if $(A_t)_{t\in T}$ is (almost surely) increasing.

Corollary 3.6.9. Consider the special case $X = Y^2$ for some martingale Y. One can show the following property:

$$\Delta A_t = \mathrm{E}\big[(\Delta Y_t)^2 \mid \mathbb{F}_{t-1}\big] \qquad \forall t \in \mathbb{R}_+. \tag{3.61}$$

The process $(A_t)_{t\in T}$ is often called the **quadratic variation process** of $(X_t)_{t\in T}$ and is denoted by $([X]_t)_{t\in T}$.

Definition 3.6.10 (Discrete stochastic integral[1]). Let $(M_n)_{n\in\mathbb{N}}$ be a martingale on a filtered probability space $(\Omega, \Sigma, \mathbb{F}, P)$ and let $(X_n)_{n\in\mathbb{N}}$ be a predictable stochastic process with respect to $\mathbb{F}$. The (discrete) stochastic integral of X with respect to M is defined as follows:

$$(X \cdot M)_t(\omega) := \sum_{i=1}^{t} X(\omega)_i \Delta M_i(\omega), \tag{3.62}$$

where $\omega \in \Omega$. For $t = 0$ the convention $(X \cdot M)_0 = 0$ is used.

Property 3.6.11. If the process $(X_n)_{n\in\mathbb{N}}$ is bounded, the stochastic integral itself defines a martingale.

Property 3.6.12 (Itô isometry). Consider a martingale $(M_n)_{n\in\mathbb{N}}$ and a predictable process $(X_n)_{n\in\mathbb{N}}$. Using the Doob decomposition theorem one can show the following equality for all $n \geq 0$:

$$\mathrm{E}\Big[(X \cdot M)_n^2\Big] = \mathrm{E}\big[(X^2 \cdot [M])_n\big]. \tag{3.63}$$

It is this property that allows for the definition of integrals with respect to continuous martingales, since although the martingales are not in general of bounded variation (and hence do not induce a well-defined Lebesgue-Stieltjes integral), their quadratic variations are (e.g. the Wiener process).

3.6.2 Markov processes

Definition 3.6.13 (Markov process). A Markov process (or chain) is a stochastic process $(X_t)_{t\in T}$ adapted to a filtration $(\mathbb{F}_t)_{t\in T}$ such that

$$P(X_t \mid \mathbb{F}_s) = P(X_t \mid X_s) \tag{3.64}$$

for all $t, s \in T$. For discrete processes, the first-order Markov chains are the most common. These satisfy

$$P(X_t \mid X_{t-1}, \dots, X_{t-r}) = P(X_t \mid X_{t-1}) \tag{3.65}$$

for all $t, r \in \mathbb{N}$.

[1] Sometimes called the **martingale transform**.

3.7 Information theory

Definition 3.7.1 (Self-information). The self-information of an event x described by a distribution P is defined as follows:

$$I(x) := -\ln P(x). \tag{3.66}$$

This definition is modeled on the following (reasonable) requirements:

- Events that are almost surely going to happen, i.e. events x such that $P(x) = 1$, contain only little information: $I(x) = 0$.[2]
- Events that are very rare contain a lot of information.
- Independent events contribute additively to the information.

Definition 3.7.2 (Shannon entropy). The amount of uncertainty in a discrete distribution P is characterized by its (Shannon) entropy

$$H(P) := \mathrm{E}[I(X)] = -\sum_i P_i \ln(P_i). \tag{3.67}$$

Definition 3.7.3 (Kullback-Leibler divergence). Let P, Q be two probability distributions. The Kullback-Leibler divergence (or **relative entropy**) of P with respect to Q is defined as follows:

$$D_{\mathrm{KL}}(P\|Q) := \int_\Omega \log\left(\frac{P}{Q}\right) dP. \tag{3.68}$$

This quantity can be interpreted as the information gained when using the distribution P instead of Q. Instead of a base-10 logarithm, any other logarithm can be used since this simply changes the result by a (positive) scaling constant.

Property 3.7.4 (Gibbs's inequality). By noting that the logarithm is a concave function and applying Jensen's equality **??**, one can prove that the Kullback-Leibler divergence is nonnegative:

$$D_{\mathrm{KL}}(P\|Q) \geq 0. \tag{3.69}$$

Furthermore, the Kullback-Leibler divergence is zero if and only if P and Q are equal almost everywhere.

3.8 Extreme value theory

Definition 3.8.1 (Conditional excess). Consider a random variable X with distribution P. The conditional probability that X is larger than a given threshold is given by the conditional excess distribution:

$$F_u(y) = \Pr(X - u \leq y \mid X > u) = \frac{P(u+y) - P(u)}{1 - P(u)}. \tag{3.70}$$

Definition 3.8.2 (Extreme value distribution). The extreme value distribution is given by the following formula:

$$F(x;\xi) = \exp\Big(-(1 + x\xi)^{-1/\xi}\Big). \tag{3.71}$$

In the case that $\xi = 0$, one can use the definition of the Euler number to rewrite the definition as

$$F(x;0) = \exp(-e^{-x}). \tag{3.72}$$

The number ξ is called the **extreme value index**.

[2] And by extension $P(x) \approx 1 \implies I(x) \approx 0$.

Definition 3.8.3 (Maximum domain of attraction). The (maximum) domain of attraction of a distribution function H consist of all distribution functions F for which there exist sequences $(a_n > 0)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ such that $F^n(a_nx+b_n) \longrightarrow H(x)$.

Theorem 3.8.4 (Fischer, Tippett & Gnedenko). *Consider a sequence of i.i.d. random variables with distribution F. If F lies in the domain of attraction of G, then G has the form of an extreme value distribution.*

Theorem 3.8.5 (Pickands, Balkema & de Haan). *Consider a sequence of i.i.d. random variables with conditional excess distribution F_u. If the distribution F lies in the domain of attraction of the extreme value distribution, the conditional excess distribution F_u converges to the generalised Pareto distribution when $u \longrightarrow \infty$.*

3.9 Copulas

Property 3.9.1 (Uniformization transform). Consider a continuous random variable X and let U be the result of the probability integral transformation, i.e. $U := F_X(X)$. This transformed random variable has a uniform cumulative distribution, i.e. $F_U(u) = u$.

Definition 3.9.2 (Copula). The joint cumulative distribution function of a random variable with uniform marginal distributions.

The following alternative definition is more analytic in nature:

Alternative Definition 3.9.3 (Copula). A function $C : [0,1]^d \to [0,1]$ satisfying the following properties:

1. **Normalization** $C(x_1, \ldots, x_d) = 0$ if any of the x_i is zero.
2. **Uniformity:** $C(1, 1, \ldots, x_i, 1, \ldots) = x_i$ for all $1 \leq i \leq d$.
3. **d-nondecreasing:** For every box $B = \prod_{1\leq i\leq d}[a_i, b_i] \subseteq [0,1]^d$ the C-volume is nonnegative:

$$\int_B dC := \sum_{\mathbf{z}\in\prod_i\{a_i,b_i\}} (-1)^{N_b(\mathbf{z})} C(\mathbf{z}) \geq 0, \tag{3.73}$$

   where $N_B(\mathbf{z}) = \text{Card}(\{i \mid a_i = z_i\})$.

Theorem 3.9.4 (Sklar). *For every joint distribution function H with marginals F_i there exists a unique copula C such that*

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{3.74}$$

Property 3.9.5 (Fréchet-Hoeffding bounds). Every copula $C : [0,1]^d \to [0,1]$ is bounded in the following way:

$$\max\left(\sum_{i=1}^d u_i - d + 1, 0\right) \leq C(u_1, \ldots, u_d) \leq \min_i u_i \tag{3.75}$$

for all $(u_1, \ldots, u_d) \in [0,1]^d$. Furthermore, the upper bound is sharp, i.e. $\min_i u_i$ is itself a copula.[3]

[3] The lower bound is only a copula for $d = 2$. In general this bound is only pointwise sharp.

Definition 3.9.6 (Extreme value copula). A copula C for which there exists a copula $\widetilde{C}$ such that

$$\left[\widetilde{C}(u_1^{1/n},\dots,u_d^{1/n})\right]^n \longrightarrow C(u_1,\dots,u_d) \tag{3.76}$$

for all $(u_1,\dots,u_d)\in[0,1]^d$.

Property 3.9.7. A copula C is an extreme value copula if and only if it is stable in the following sense:

$$C(u_1,\dots,u_d) = \left[C(u_1^{1/n},\dots,u_d^{1/n})\right]^n \tag{3.77}$$

for all $n\geq 1$.

3.10 Randomness ♣

This section is strongly related to Section **??** on computability theory.

Definition 3.10.1 (Kolmogorov randomness). Consider a *universal Turing machine* U. The **Kolmogorov complexity** $C(\kappa)$ of a finite bit string κ (with respect to U) is defined as

$$C(\kappa) := \min\{|\sigma| \mid \sigma \text{ is finite}, U(\sigma)=\kappa\}. \tag{3.78}$$

A finite bit string is said to be Kolmogorov random (with respect to U) if there exists an integer $n\in\mathbb{N}$ such that $C(\kappa)\geq|\sigma|-n$.

Property 3.10.2. For every universal Turing machine there exists at least one Kolmogorov random string. This easily follows from the pigeonhole principle since for every $n\in\mathbb{N}$ there are 2^n strings of length n but only 2^n-1 programs of length less than n.

Remark 3.10.3. Note that, although universal Turing machines can emulate each other, the randomness of a string is not absolute. Its randomness depends on the chosen machine.

It would be pleasing if this notion of randomness could easily be extended to infinite bit strings, for example by giving such a string the label random if there exists a uniform choice of constant k such that all initial segments of the string are k-random. However, by a result of *Martin-Löf*, there does not exist any string satisfying this condition.

3.11 Imprecise probability

Definition 3.11.1 (Gamble). Consider a set X. The set of gambles over X is the (Banach) space of bounded real-valued functions $\mathcal{L}(X) := \{f : X\rightarrow\mathbb{R}\}$. A subset $\mathcal{D}\subset\mathcal{L}(X)$ of "desirable" gambles is said to be **coherent** if it satisfies the following condition:

1. **Positivity:** $\lambda>0 \implies \lambda\mathcal{D}=\mathcal{D}$,
2. **Additivity:** $\mathcal{D}+\mathcal{D}\subseteq\mathcal{D}$,
3. **Accepting partial gains:** $\mathcal{L}^+(X)\subseteq\mathcal{D}$, where $\mathcal{L}^+(X) := \{f\in\mathcal{L}(X)\mid f>0\}$, and
4. **Avoiding partial losses:** $\mathcal{L}^-(X)\cap\mathcal{D}=\emptyset$, $\mathcal{L}^-(X) := \{f\in\mathcal{L}(X)\mid f<0\}$.

The first two axioms imply that the desirable gambles form a convex cone. It is also clear that the positive orthant $\mathcal{L}^+(X)$ is the smallest coherent set of desirable gambles.

Property 3.11.2 (Order structure). The collection of all coherent sets of desirable gambles over a space X can be given a poset structure by inclusion, with $\mathcal{L}^+(X)$ as its least element. If $\mathcal{D}\subset\mathcal{D}^+$, then $\mathcal{D}$ is said to be less **committal** than $\mathcal{D}'$.

Definition 3.11.3 (Credal set). A subset of the set of probability measures $\mathbb{P}(X)$.

Credal sets are often used to represent the lack of knowledge about a probability distribution. For this reason it is natural to assume that credal sets are convex. If one is uncertain about both $P_1, P_2 \in K \subseteq \mathbb{P}(X)$, one is also uncertain about the mixtures $\lambda P_1 + (1-\lambda)P_2, \lambda \in [0,1]$.

Chapter 4

Statistics

In this chapter, most definitions and formulas will be based on either a standard calculus approach or a data-driven approach. For a measure-theoretic approach, see Chapter 3. For some sections the language of information geometry will be used as introduced in the previous chapter.

4.1 Data samples

4.1.1 Moment estimators

Formula 4.1.1 (r^{th} sample moment).

$$\overline{x^r} := \frac{1}{N}\sum_{i=1}^{N} x_i^r \tag{4.1}$$

Example 4.1.2 (Arithmetic mean). The arithmetic mean is used to average out differences between measurements. It is defined as the first sample moment:

$$\overline{x} := \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{4.2}$$

Formula 4.1.3 (r^{th} central sample moment).

$$m_r := \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^r \tag{4.3}$$

Definition 4.1.4 (Weighted mean). Let $f : \mathbb{R} \to \mathbb{R}^+$ be a weight function. The weighted mean is given by:

$$\overline{x} := \frac{\sum_i f(x_i)x_i}{\sum_i f(x_i)}. \tag{4.4}$$

Example 4.1.5 (Binned mean). If the data has been grouped in bins, the weight function is given by the number of elements in each bin.

$$\overline{x} = \frac{1}{N}\sum_{i=1} n_i x_i. \tag{4.5}$$

Remark 4.1.6. In the above definitions, the measurements x_i can be replaced by function values $f(x_i)$ to calculate the mean of the function $\overline{f(x)}$. This follows from Theorem 3.3.4. However, it is also important to keep in mind that $\overline{f(x)} \neq f(\overline{x})$. The equality only holds for linear functions.

Definition 4.1.7 (Geometric mean). Let $\{x_i\}$ be a data set taking values in either $\mathbb{R}_+$ or $\mathbb{R}_-$. The geometric mean is used to average out *normalised* measurements, i.e. ratios with respect to a reference value.

$$g := \left(\prod_{i=1}^{N} x_i\right)^{1/N} \tag{4.6}$$

The following relation exists between the arithmetic and geometic mean:

$$\ln g = \overline{\ln x}. \tag{4.7}$$

Definition 4.1.8 (Harmonic mean).

$$h := \left(\frac{1}{N}\sum_{i=1}^{N} x_i^{-1}\right)^{-1} \tag{4.8}$$

The following relation exists between the arithmetic and harmonic mean:

$$\frac{1}{h} = \overline{x^{-1}}. \tag{4.9}$$

Property 4.1.9. Let $\{x_i\}$ be a data set taking values in $\mathbb{R}_+$.

$$h \leq g \leq \overline{x} \tag{4.10}$$

The equalities only hold when all x_i are equal.

Definition 4.1.10 (Mode). The most occurring value in a data set.

Definition 4.1.11 (Median). The element x_i in a data set such that half of the values is greater than x_i and half of the values is smaller than x_i.

4.1.2 Dispersion

Definition 4.1.12 (Range). The simplest indicator for statistical dispersion:

$$R := x_{\text{max}} - x_{\text{min}}. \tag{4.11}$$

However, it is very sensitive for outliers.

Definition 4.1.13 (Mean absolute difference).

$$\text{MD} := \frac{1}{N}\sum_{i=1}^{N} |x_i - \overline{x}| \tag{4.12}$$

Definition 4.1.14 (Sample variance).

$$\text{Var}(x) := \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{4.13}$$

Formula 4.1.15. The variance can also be rewritten in the following way:

$$\mathrm{Var}(x) = \overline{x^2} - \overline{x}^2. \tag{4.14}$$

Remark 4.1.16 (Bessel corection). A better estimator for the variance of a sample is given by the following formula:

$$\hat{s} := \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2. \tag{4.15}$$

See Remark 4.4.9 for more information.

Definition 4.1.17 (Skewness). The skewness γ describes the asymmetry of a distribution. It is defined as the proportionality constant relating the third central moment m_3 and the standard deviation σ:

$$m_3 = \gamma\sigma^3. \tag{4.16}$$

A positive skewness indicates a tail to the right or alternatively a median smaller than $\overline{x}$. A negative skewness indicates a median larger than $\overline{x}$.

Definition 4.1.18 (Pearson's mode skewness).

$$\gamma_P := \frac{\overline{x} - \mathrm{mode}}{\sigma} \tag{4.17}$$

Definition 4.1.19 (Kurtosis). The kurtosis c is an indicator for the "tailedness". It is defined as the proportionality constant relating the fourth central moment m_4 and the standard deviation σ:

$$m_4 = c\sigma^4. \tag{4.18}$$

Definition 4.1.20 (Excess kurtosis). The excess kurtosis is defined as $c - 3$. This fixes the excess kurtosis of all univariate normal distributions at 0. A positive excess is an indicator for long "fat" tails, a negative excess indicates short "thin" tails.

Definition 4.1.21 (Percentile). The p-percentile c_p is defined as the value that is larger than $p\%$ of the measurements. The median is the 50-percentile.

Definition 4.1.22 (Interquartile range). The difference between the upper and lower quartile (75- and 25-percentiles respectively).

Definition 4.1.23 (Full Width at Half Maximum). The difference between the two values of the independent variable where the dependent variable is half of its maximum. This quantity is often denoted by the abbreviation **FWHM**.

Property 4.1.24. For Gaussian distributions the following relation exists between the FWHM and the standard deviation σ:

$$\mathrm{FWHM} = 2.35\sigma. \tag{4.19}$$

4.1.3 Multivariate data sets

When working with bivariate (or even multivariate) distributions it is useful to describe the relationship between the different random variables.

Definition 4.1.25 (Covariance). The covariance of two data sequences is defined as follows:

$$\mathrm{cov}(x,y) := \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \overline{xy} - \overline{x}\,\overline{y}. \tag{4.20}$$

The covariance is also often denoted by σ_{xy}.

Formula 4.1.26. The covariance and standard deviation are related by the following equality:

$$\sigma_x^2 = \sigma_{xx}. \tag{4.21}$$

Definition 4.1.27 (Correlation coefficent).

$$\rho_{xy} := \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} \tag{4.22}$$

The correlation coefficient is bounded to the interval $[-1, 1]$. It should be noted that its magnitude is only an indicator for the linear dependence.

Remark 4.1.28. For multivariate distributions the above definitions can be generalized using matrices:

$$V_{ij} = \mathrm{cov}(x_{(i)}, x_{(j)}) \tag{4.23}$$

$$\rho_{ij} = \rho_{(i)(j)}, \tag{4.24}$$

where $\mathrm{cov}(x_{(i)}, x_{(j)})$ and $\rho_{(i)(j)}$ are defined as in Formulas 4.1.25 and 4.1.27.

4.2 Probability distributions

In the following sections and subsections, all distributions will be taken to be continuous. The formulas can be generalized to discrete distributions by replacing the integral with a summation.

Definition 4.2.1 (Percentile). The p-percentile c_p of a distribution F is defined as:

$$c_p = F^{-1}(p). \tag{4.25}$$

Definition 4.2.2 (Parametric family). A family of probability densities indexed by one or more parameters θ.

Example 4.2.3 (Mixture family). Consider a collection of distributions $\mathcal{P} = \{P_i\}_{i\leq n}$. The mixture family generated by $\mathcal{P}$ consist of all convex combintations of elements in $\mathcal{P}$:

$$\left\{\sum_{i=1}^{n} w_i P_i \,\middle|\, w_i \geq 0, \sum_{i=1}^{n} w_i = 1\right\}. \tag{4.26}$$

Every element of this family is called a **mixture distribution**.

4.2.1 Empirical distribution

Definition 4.2.4 (Empirical distribution function). The (discrete) empirical probability distribution function is defined as the uniform mixture distribution with Dirac measures at the observations:

$$F_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}. \tag{4.27}$$

Theorem 4.2.5 (Borel's law of large numbers). *If the sample size approaches infinity, the observed frequencies approach the theoretical propabilities.*

Corollary 4.2.6 (Frequentist probability[1]).

$$\Pr(x) := \lim_{n\to\infty} \frac{f_n(x)}{n} \tag{4.28}$$

The law of large numbers can also be phrased in terms of the empirical distribution function:

Theorem 4.2.7 (Glivenko-Cantelli). *Consider a cumulative distribution function F on a probability space Ω. Denote the empirical distribution function of n random variables on Ω by F_n. If the random variables are i.i.d. according to F, then*

$$\sup_{x\in\Omega} |F(x) - F_n(x)| \xrightarrow{a.s.} 0. \tag{4.29}$$

Remark 4.2.8. The law of the large numbers implies pointwise convergence of the empirical distribution function, while the Glivenko-Cantelli theorem strengthens this to uniform convergence.

The quantity in the Glivenko-Cantelli theorem is important enough to get its own name:

Definition 4.2.9 (Kolmogorov-Smirnov statistic). Let F be a given cumulative distribution function. The n^{th} Kolmogorov-Smirnov statistic is defined as follows:

$$D_n := \sup_x |F_n(x) - F(x)|. \tag{4.30}$$

Definition 4.2.10 (Kolmogorov distribution).

$$F_{\text{Kol}}(x) := 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty} e^{-(2i-1)^2\pi^2/(8x^2)} \tag{4.31}$$

Property 4.2.11 (Kolmogorov-Smirnov test). Let the null hypothesis H_0 state that a given data sample is described by a distribution function F. The null hypothesis is rejected at significance level α if

$$\sqrt{n}D_n > K_\alpha, \tag{4.32}$$

where K_α is defined by the Kolmogorov distribution: $F_{\text{Kol}}(K_\alpha) = 1 - \alpha$.

4.2.2 Common distributions

Formula 4.2.12 (Uniform distribution).

$$f(x;a,b) := \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases} \tag{4.33}$$

$$\text{E}[x] = \frac{a+b}{2} \tag{4.34}$$

$$\text{Var}[x] = \frac{(b-a)^2}{12} \tag{4.35}$$

[1] Also called the **empirical probability**.

Formula 4.2.13 (Gaussian distribution).

$$\mathcal{G}(x;\mu,\sigma) := \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4.36}$$

This distribution is also called a (univariate) **normal distribution**.

Formula 4.2.14 (Standard normal distribution).

$$\mathcal{N}(z) := \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \tag{4.37}$$

The cumulative distribution of $\mathcal{N}$ is given by the *error function*.

Remark 4.2.15. Every Gaussian distribution can be transformed into a standard normal distribution by passing to the random variable $Z = \frac{X-\mu}{\sigma}$. This transformation is often called **standardization**.

Theorem 4.2.16 (Central limit theorem). *A sum of n i.i.d. random variables X_i distributed according to a distribution with mean μ and variance σ^2 satisfies the following property:*

$$\sqrt{n}\left(\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} \mathcal{N}(0,\sigma^2). \tag{4.38}$$

Remark 4.2.17. If the random variables are not independent, property 2 will not be fulfilled. However, a generalization to distributions that are not identical exists. These are the *Lyapunov* and *Lindeberg* CLTs. (This generalization does require additional conditions on the higher moments.)

Remark 4.2.18. The sum of Gaussians is always Gaussian.

Formula 4.2.19 (Exponential distribution).

$$f(x;\tau) := \frac{1}{\tau}e^{-\frac{x}{\tau}} \tag{4.39}$$

$$\mathrm{E}[x] = \tau \tag{4.40}$$

$$\mathrm{Var}[x] = \tau^2. \tag{4.41}$$

Property 4.2.20. The exponential distribution is **memoryless**:

$$\Pr(X > x_1 + x_2 | X > x_2) = \Pr(X > x_1). \tag{4.42}$$

Formula 4.2.21 (Bernoulli distribution). A random variable that can only take 2 possible values is described by a Bernoulli distribution. When the possible values are 0 and 1, with respective chances ρ and $1-\rho$, the distribution is given by

$$p(k;\rho) := \rho^k(1-\rho)^{1-k} \tag{4.43}$$

$$\mathrm{E}[k] = \rho \tag{4.44}$$

$$\mathrm{Var}[k] = \rho(1-\rho). \tag{4.45}$$

Formula 4.2.22 (Binomial distribution). A process with n i.i.d. Bernoulli trials with probability ρ, is described by a binomial distribution:

$$p(k;\rho,n) := \binom{n}{k}\rho^k(1-\rho)^{n-k} \tag{4.46}$$

$$\mathrm{E}[k] = n\rho \tag{4.47}$$

$$\mathrm{Var}[k] = n\rho(1-\rho). \tag{4.48}$$

Formula 4.2.23 (Poisson distribution). A process with known possible outcomes but an unknown number of events is described by a Poisson distribution with average expected number of events λ.

$$p(r;\lambda) := \frac{e^{-\lambda}\lambda^r}{r!} \tag{4.49}$$

$$\mathrm{E}[r] = \mathrm{Var}[r] = \lambda. \tag{4.50}$$

Property 4.2.24. If two Poisson processes, with expectations λ_a and λ_b respectively, occur simultaneously, the probability of r events is also described by a Poisson distribution with average $\lambda_a + \lambda_b$. The number of events coming from the process described by λ_a is given by a binomial distribution $p(r_a;\Lambda_a,r)$ with $\Lambda_a = \frac{\lambda_a}{\lambda_a+\lambda_b}$.

Remark 4.2.25. For $\lambda \longrightarrow \infty$, the Poisson distribution $p(r;\lambda)$ can be approximated by a Gaussian distribution $\mathcal{G}(x;\lambda,\sqrt{\lambda})$.

Formula 4.2.26 (χ^2-distribution). The sum of k squared, independent (standard), normally distributed random variables Y_i defines the random variable:

$$\chi_k^2 := \sum_{i=1}^k Y_i^2, \tag{4.51}$$

where k is said to be the number of **degrees of freedom**. The associated density is

$$f(\chi^2;n) := \frac{\chi^{n-2}e^{-\frac{\chi^2}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}. \tag{4.52}$$

Remark 4.2.27. Due to the CLT 4.2.16 the χ^2-distribution approximates a Gaussian distribution for large k: $f(\chi^2;k) \overset{k>30}{\longrightarrow} \mathcal{G}(\sqrt{2\chi^2};\sqrt{2k-1},1)$.

Formula 4.2.28 (Student-t distribution). The Student-t distribution describes the difference between the true mean and a sample average with estimated standard deviation $\hat{\sigma}$:

$$f(t;n) := \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}, \tag{4.53}$$

where

$$t := \frac{(x-\mu)/\sigma}{\hat{\sigma}/\sigma} = \frac{z}{\sqrt{\chi^2/n}}. \tag{4.54}$$

Formula 4.2.29 (Cauchy distribution[2]). The general density $f(x;x_0,\gamma)$ is given by

$$f(x;x_0,\gamma) := \frac{1}{\pi}\frac{\gamma}{(x-x_0)^2+\gamma^2}. \tag{4.55}$$

The associated characteristic function is given by

$$\mathrm{E}\big[e^{itx}\big] = e^{ix_0t-\gamma|t|}. \tag{4.56}$$

Property 4.2.30. Both the mean and variance of the Cauchy distribution are undefined.

4.3 Errors

Definition 4.3.1 (Systematic error). Errors that always have the same effect independent of the measurements itself, i.e. they shift all values in the same way and cannot be directly inferred from the measurements. Note that they are not necessarily independent of each other.

Formula 4.3.2 (Inverse-variance averaging). When performing a sequence of measurements x_i with different variances σ_i^2, it is impossible to use the arithmetic mean 4.1.2 in a meaningful way because the measurements are not of the same type. Therefore it is also impossible to apply the CLT 4.2.16.

These problems can be resolved by the using the weighted mean 4.1.4:

$$\overline{x} := \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}. \tag{4.57}$$

The variation of the weighted mean is given by

$$\mathrm{Var}(\overline{x}) := \frac{1}{\sum_i \sigma_i^{-2}}. \tag{4.58}$$

Formula 4.3.3. Let X be random variable with variance $\mathrm{Var}[X]$. The variance of a linear function $f(X) = aX + b$ is given by

$$\mathrm{Var}[f] = a^2\mathrm{Var}[X]. \tag{4.59}$$

Formula 4.3.4. Let X be random variable with a small variance $\mathrm{Var}[X]$. The variance of a general function $f(X)$ is given by

$$\mathrm{Var}[f] \approx \left(\frac{df}{dx}\right)^2\mathrm{Var}[x]. \tag{4.60}$$

Corollary 4.3.5. The correlation coefficient 4.1.27 of a random variable X and a **linear** function of X is independent of σ_x and is always equal to ±1.

Formula 4.3.6 (Law of error propagation). Let $\mathbf{X}$ be a vector random variable with small variance. The variance of a general function $f(\mathbf{X})$ is given by

$$\mathrm{Var}[f] = \sum_p\left(\frac{\partial f}{\partial X_{(p)}}\right)^2\mathrm{Var}\big[X_{(p)}\big] + \sum_{p\neq q}\left(\frac{\partial f}{\partial X_{(p)}}\right)\left(\frac{\partial f}{\partial X_{(q)}}\right)\mathrm{cov}[X_{(p)},X_{(q)}]. \tag{4.61}$$

[2] Also known (especially in particle physics) as the **Breit-Wigner** distribution.

Definition 4.3.7 (Fractional error). Let X, Y be two independent random variables. The standard deviation of $f(X,Y) = XY$ is given by the fractional error:

$$\left(\frac{\sigma_f}{f}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2. \tag{4.62}$$

The fractional error of a variable is equal to the fractional error of the reciprocal of that variable.

Property 4.3.8 (Logarithm). Let X be a random variable. The error of the logarithm of X is equal to the fractional error of X.

Formula 4.3.9 (Covariance of functions).

$$\mathrm{cov}[f,g] = \sum_{p,q}\left(\frac{\partial f}{\partial X_{(p)}}\right)\left(\frac{\partial g}{\partial X_{(q)}}\right)\mathrm{cov}[X_{(p)}, X_{(q)}] \tag{4.63}$$

Corollary 4.3.10. Let $\mathbf{f} = (f_1, \ldots, f_k)$ be a vector-valued function. The covariance matrix $\mathrm{Var}[\mathbf{f}]$ is given by

$$\mathrm{Var}[\mathbf{f}] = J\mathrm{Var}[\mathbf{X}]J^T, \tag{4.64}$$

where J is the Jacobian matrix of $\mathbf{f}$.

4.4 Parameter estimation

4.4.1 General properties

Definition 4.4.1 (Consistency). An estimator $\hat{a}$ is said to be consistent if it is asymptotically equal to the true parameter:

$$\lim_{N\to\infty} \hat{a} = a. \tag{4.65}$$

Definition 4.4.2 (Unbiased estimator). An estimator $\hat{a}$ is said to be unbiased if its expectation value is equal to the true parameter:

$$\langle\hat{a}\rangle = a. \tag{4.66}$$

Note that neither consistency, nor unbiasedness implies the other.

Definition 4.4.3 (Bias).

$$B(\hat{a}) := |\langle\hat{a}\rangle - a|. \tag{4.67}$$

Definition 4.4.4 (Mean squared error).

$$\mathrm{MSE}(\hat{a}) := B(\hat{a})^2 + \mathrm{Var}(\hat{a}). \tag{4.68}$$

Remark 4.4.5. If an estimator is unbiased, the MSE is equal to the variance of the estimator.

4.4.2 Common estimators

Property 4.4.6 (Unbiased mean). The CLT 4.2.16 implies that the sample mean 4.1.2 is a consistent and unbiased estimator of the population mean.

Formula 4.4.7 (Standard error of the mean). Using the Bienaymé formula 3.4.23 one can show that the standard error of the mean, i.e. the standard deviation of the sample mean, is given by the following formula:

$$\mathrm{Var}(\overline{x}) = \frac{\sigma^2}{N}. \tag{4.69}$$

Formula 4.4.8 (Variance estimator for known mean). If the true mean μ is known, a consistent and unbiased estimator for the variance is given by

$$\widehat{\mathrm{Var}[X]} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{4.70}$$

Formula 4.4.9 (Variance estimator for unknown mean). If the true mean is unknown and the sample mean has been used to estimate it, a consistent and unbiased estimator is given by

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2. \tag{4.71}$$

The modified factor $\frac{1}{N-1}$ is called the **Bessel correction**. It corrects the bias of the estimator given by the sample variance 4.1.14. The consistency is guaranteed by the CLT.

4.4.3 Estimation error

Formula 4.4.10 (Variance of the estimator of the variance).

$$\mathrm{Var}\left(\widehat{\mathrm{Var}[X]}\right) = \frac{(N-1)^2}{N^3}\langle(x - \langle x\rangle)^4\rangle - \frac{(N-1)(N-3)}{N^3}\langle(x - \langle x\rangle)^2\rangle^2 \tag{4.72}$$

Formula 4.4.11 (Variance of the estimator of the standard deviation).

$$\mathrm{Var}(\widehat{\sigma}) = \frac{1}{4\sigma^2}\mathrm{Var}\left(\widehat{\mathrm{Var}[X]}\right) \tag{4.73}$$

Remark 4.4.12. The previous result is a little odd, as one has to know the true standard deviation to compute the variance of the estimator. This problem can be solved in two ways. Either a value (hopefully close to the real one) inferred from the sample is used as an estimator, or a guess is used in the design phase of an experiment to see what the possible outcomes are.

4.4.4 Likelihood function

Definition 4.4.13 (Likelihood). The likelihood $\mathcal{L}(a;\vec{x})$ is the probability to find a set of measurements $\vec{\boldsymbol{x}} = \{x_1, \ldots, x_N\}$ given a density $f(X;a)$:

$$\mathcal{L}(a;\vec{\boldsymbol{x}}) = \prod_{i=1}^{N} f(x_i;a). \tag{4.74}$$

Definition 4.4.14 (Log-likelihood).

$$\log\mathcal{L}(a;\vec{\boldsymbol{x}}) = \sum_i \ln f(x_i;a) \tag{4.75}$$

Property 4.4.15. The expectation value of an estimator $\hat{a}$ is given by

$$\langle\hat{a}\rangle = \int \hat{a}\mathcal{L}(\hat{a};x)dx. \tag{4.76}$$

Theorem 4.4.16 (Cramer-Rao bound). *The variance of an* ***unbiased*** *estimator has a lower bound called the Cramer-Rao bound or* ***minimum variance bound (MVB)****:*

$$\mathrm{Var}(\hat{a}) \geq \frac{1}{\left\langle\left(\frac{d\ln\mathcal{L}}{da}\right)^2\right\rangle}. \tag{4.77}$$

For a biased estimator with bias b, the MVB takes on the following form:

$$\mathrm{Var}(\hat{a}) \geq \frac{\left(1+\frac{db}{da}\right)^2}{\left\langle\left(\frac{d\ln\mathcal{L}}{da}\right)^2\right\rangle}. \tag{4.78}$$

Remark 4.4.17.

$$\left\langle\left(\frac{d\ln\mathcal{L}}{da}\right)^2\right\rangle = -\left\langle\frac{d^2\ln\mathcal{L}}{da^2}\right\rangle \tag{4.79}$$

Definition 4.4.18 (Fisher information).

$$I_X(a) := \left\langle\left(\frac{d\ln\mathcal{L}}{da}\right)^2\right\rangle = N\int\left(\frac{d\ln f}{da}\right)^2 f\,dX \tag{4.80}$$

Using this definition one can rewrite the Cramer-Rao inequality as follows:

$$\mathrm{Var}(\hat{a}) \geq I_X(a). \tag{4.81}$$

Definition 4.4.19 (Finite-sample efficiency). An unbiased estimator is said to be (finite-sample) efficient if it saturates the Cramer-Rao bound. In general the **efficiency** of (unbiased) estimators is defined through the Cramer-Rao bound as follows:

$$e(\hat{a}) := \frac{I_X(a)^{-1}}{\mathrm{Var}(\hat{a})}. \tag{4.82}$$

4.4.5 Maximum likelihood estimation

From Definition 4.4.13 it follows that the estimator $\hat{a}_{\mathrm{MLE}}$ that makes the given measurements most probable is the value of a for which the likelihood function is maximal. It is therefore not the most probable estimator.

Using Bayes's theorem one finds $f(a|x) = f(x|a)\frac{f(a)}{f(x)}$. The prior density $f(x)$ is fixed since the values x_i are given by the measurement and, hence, does not vary. The density $f(a)$ is generally assumed to be uniform if there is no prior knowledge about a. It follows that $f(a|x)$ and $f(x|a)$ are proportional and, hence, the logarithms of these functions differ only by an additive constant. This leads to following method for finding an estimator $\hat{a}$:

Method 4.4.20 (Maximum likelihood estimator). The maximum likelihood estimator $\hat{a}$ is obtained by solving the following equation:

$$\left.\frac{d\ln\mathcal{L}}{da}\right|_{a=\hat{a}} = 0. \tag{4.83}$$

Remark 4.4.21. MLE estimators are mostly consistent but often biased.

Property 4.4.22. MLE estimators are invariant under parameter transformations.

Corollary 4.4.23. The invariance implies that the two estimators $\hat{a}$ and $\widehat{f(a)}$ cannot both be unbiased at the same time.

Property 4.4.24. Every consistent estimator asymptotically becomes unbiased and efficient.

Property 4.4.25 (Minimizing KL-divergence). It can be shown that maximizing the log-likelihood is equivalent to minimizing the Kullback-Leibler divergence 3.7.3 between the would-be distribution $p(x;\theta)$ and the true distribution $q(x)$:

$$\begin{aligned}\arg\max_\theta \ln\mathcal{L} &= \arg\max_\theta \sum_{i\in I} \ln p(x_i;\theta)\\ &= \arg\max_\theta \sum_{i\in I} \ln p(x_i;\theta) - \ln q(x_i)\\ &= \arg\min_\theta \frac{1}{n}\sum_{i\in I} \ln\frac{q(x_i)}{p(x_i;\theta)}\\ &\longrightarrow \arg\min\theta \int p(x;\theta)\ln\frac{q(x)}{p(x;\theta)}dx = \arg\min_\theta D_{\mathrm{KL}}(p_\theta\|q),\end{aligned}$$

where the law of large numbers was used in the last line.

4.4.6 Least squares estimation

To fit a (parametric) function $y = f(x;a)$ to a set of 2 variables (x,y), where the x values are exact and the y values have an uncertainty σ_i, one can use the following method:

Method 4.4.26 (Least squares).

1. For every event (x_i, y_i) define the residual $d_i := y_i - f(x_i;a)$.
2. Determine the χ^2-statistic (analytically):

$$\chi^2 := \sum_i \frac{d_i^2}{f_i}, \tag{4.84}$$

   where $f_i = f(x_i;a)$.

3. Find the most probable value of $\hat{a}$ by solving the equation

$$\frac{d\chi^2}{da} = 0. \tag{4.85}$$

Property 4.4.27. The optimal χ^2-value is asymptotically distributed according to a χ^2-distribution $f(\chi^2;n)$. The number of degrees of freedom n is equal to the number of events N minus the number of fitted parameters k. (See more in Section 4.9.1.)

Formula 4.4.28 (Linear fit). When all uncertainties σ_i are equal, the slope $\hat{m}$ and intercept $\hat{c}$ are given by the following formulas:

$$\hat{m} := \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{Var}(x)} \tag{4.86}$$

$$\hat{c} := \overline{y} - \hat{m}\overline{x} = \frac{\overline{x^2} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}. \tag{4.87}$$

Remark 4.4.29. The equation $\overline{y} = \hat{c} + \hat{m}\overline{x}$ says that the linear fit passes through the center of mass $(\overline{x}, \overline{y})$.

Formula 4.4.30 (Errors of linear fit).

$$\mathrm{Var}(\hat{m}) = \frac{1}{N(\overline{x^2} - \overline{x}^2)}\sigma^2 \tag{4.88}$$

$$\mathrm{Var}(\hat{c}) = \frac{\overline{x^2}}{N(\overline{x^2} - \overline{x}^2)}\sigma^2 \tag{4.89}$$

$$\mathrm{cov}(\hat{m}, \hat{c}) = \frac{-\overline{x}}{N(\overline{x^2} - \overline{x}^2)}\sigma^2 \tag{4.90}$$

The least squares method is very useful to fit data that has been grouped in bins (histograms):

Method 4.4.31 (Binned least squares).

1. N i.i.d. events with distributions $f(X; a)$ divided in N_B intervals, where the interval j is centered on the value x_j, has a width W_j and contains n_j events.
2. The ideally expected number of events in the j^{th} interval: $f_j = NW_jf(x_j; a)$.
3. The real number of events has a Poisson distribution: $\overline{n}_j = \sigma_j^2 = f_j$.
4. Define the binned χ^2 as

$$\chi^2 := \sum_i^{N_B} \frac{(n_i - f_i)^2}{f_i^2}. \tag{4.91}$$

4.4.7 Geometric approach

Consider a sample $\mathbf{x} := \{x_1, \ldots, x_n\}$ drawn from a distribution $P(\mathbf{x}; \theta)$ in an exponential family. The likelihood 4.4.13 is given by

$$p(\mathbf{x}; \theta) = \prod_{i=1}^n P(x_i; \theta).$$

The m-coordinates of the observed point are

$$\eta = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x}. \tag{4.92}$$

The optimal value for θ can be found by maximizing the log-likelihood $\log p(\mathbf{x}; \theta)$ or, equivalently according to Property 4.4.25, by minimizing the Kullback-Leibler divergence between the observed point and the "true" distribution $P(x; \xi)$. The latter is found by m-projecting the observed point η on the submanifold S of "admissible" distributions.

Theorem 4.4.32 (Sanov). *Consider a probability distribution q on a finite set S and draw n i.i.d. samples. Let P_n be the empirical distribution function of the samples* (4.2.4)*. Further, let Γ be a collection of probability distributions such that $P_n \in \Gamma$. The joint distribution q^n satisfies the following inequality:*

$$q^n(P_n \in \Gamma) \leq (n+1)^{|S|}2^{-nD_{\mathrm{KL}}(p^*\|q)}, \tag{4.93}$$

where p^* is the information projection of q on Γ. If $\Gamma = \overline{\Gamma^\circ}$, this can be restated as

$$\lim_{n\to\infty} \frac{1}{n}q^n(P_n \in \Gamma) = -D_{\mathrm{KL}}(p^*\|q). \tag{4.94}$$

4.5 Bayesian modelling

Definition 4.5.1 (Conjugate distributions). Consider a prior distribution $F(\theta)$ and a posterior distribution $F(\theta|X)$. If these distributions belong to the same family, e.g. they are both Gaussians, they are said to be conjugate. In this case the prior $F(\theta)$ is said to be a **conjugate prior** for the likelihood $F(X|\theta)$.

Example 4.5.2. The simplest example is the case of binomial distributions, where the conjugate prior is the *Beta distribution*. This can be generalized to multi-class situations. The conjugate prior of a categorical (or even *multinomial*) distribution is the *Dirichlet distribution*.

4.6 Confidence intervals

The true value of a parameter ε can never be known exactly. However, it is possible to construct an interval I in which this value should lie with a certain confidence C.

Example 4.6.1 (Prediction interval). Let X be a normally distributed random variable. A measurement will lie in the interval $[\mu - 1.96\sigma, \mu + 1.96\sigma]$ with 95% probability. The true value μ lies in the interval $[x - 2\sigma, x + 2\sigma]$ with 95% confidence.

Remark. In the previous example some assumptions were made. All possible values (left or right side of peak) are given the same probability due to the Gaussian distribution. If one removes this symmetry condition, a more careful approach is required. Furthermore, the apparent symmetry between the uncertainty and confidence levels is only valid for Gaussian distributions.

4.6.1 Interval types

Definition 4.6.2 (Two-sided confidence interval).

$$\Pr(x_- \leq X \leq x_+) = \int_{x_-}^{x_+} f(x)dx = C \tag{4.95}$$

There are three possible (often used) two-sided intervals:

- **symmetric interval**: $x_+ - \mu = \mu - x_-$,
- **shortest interval**: $|x_+ - x_-|$ is minimal, or
- **central interval**: $\int_{-\infty}^{x_-} f(x)dx = \int_{x_+}^{\infty} f(x)dx = \frac{1-C}{2}$.

The central interval is the most widely used confidence interval.

Remark 4.6.3. For Gaussian distributions these three definitions are equivalent.

Definition 4.6.4 (One-sided confidence interval).

$$\Pr(x \geq x_-) = \int_{x_-}^{\infty} f(x)dx = C \tag{4.96}$$

$$\Pr(x \leq x_+) = \int_{-\infty}^{x_+} f(x)dx = C \tag{4.97}$$

Definition 4.6.5 (Discrete central confidence interval). For a discrete distribution it is often impossible to find integers $x_\pm$ such that the real value lies with exact confidence C in the interval $[x_-, x_+]$.

$$x_- = \arg\min_\theta \left[\frac{1-C}{2} - \sum_{x=0}^{\theta-1} p(x)\right] \tag{4.98}$$

$$x_+ = \arg\min_\theta \left[\frac{1-C}{2} - \sum_{x=\theta+1}^{\infty} p(x)\right] \tag{4.99}$$

4.6.2 General construction

For every value of the true parameter X it is possible to construct a confidence interval. This leads to the construction of two functions $x_-(X)$ and $x_+(X)$. The 2D diagram obtained by plotting $x_-(X)$ and $x_+(X)$ with the x-axis horizontally and X-axis vertically is called the **confidence region**.

Method 4.6.6. Let x_0 be a point estimate of the parameter X. From the confidence region it is possible to infere a confidence interval $[X_-(x), X_+(x)]$, where the upper limit X_+ is not the limit such that there is only a $\frac{1-C}{2}$ chance of having a true parameter $X \geq X_+$, but the limit such that if the true parameter $X \geq X_+$ then there is a chance of $\frac{1-C}{2}$ to have a measurement x_0 or smaller.

4.6.3 Interval for a sample mean

Formula 4.6.7 (Interval with known variance). If the sample size is large enough, the real distribution is unimportant, because the CLT ensures a Gaussian distribution of the sample mean $\overline{X}$. The α-level confidence interval such that $\Pr(-z_{\alpha/2} < Z < z_{\alpha/2})$ with $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{N}}$ is given by

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right]. \tag{4.100}$$

Remark 4.6.8. If the sample size is not sufficiently large, the measured quantity must follow a normal distribution.

Formula 4.6.9 (Interval with unknown variance). To account for the uncertainty of the estimated standard deviation $\hat{\sigma}$, the student-t distribution 4.2.28 is used instead of a Gaussian distribution to describe the sample mean $\overline{X}$. The α-level confidence interval is given by

$$\left[\overline{X} - t_{\alpha/2;(n-1)}\frac{s}{\sqrt{N}}, \overline{X} + t_{\alpha/2;(n-1)}\frac{s}{\sqrt{N}}\right], \tag{4.101}$$

where s is the estimated standard deviation 4.4.9.

Formula 4.6.10 (Wilson score interval). For a sufficiently large sample, a sample proportion $\hat{P}$ is approximately Gaussian distributed with expectation value π and variance $\frac{\pi(\pi-1)}{N}$. The α-level confidence interval is given by

$$\left[\frac{(2N\hat{P} + z^2_{\alpha/2}) - z_{\alpha/2}\sqrt{z^2_{\alpha/2} + 4N\hat{P}(1-\hat{P})}}{2(N + z^2_{\alpha/2})}, \frac{(2N\hat{P} + z^2_{\alpha/2}) + z_{\alpha/2}\sqrt{z^2_{\alpha/2} + 4N\hat{P}(1-\hat{P})}}{2(N + z^2_{\alpha/2})}\right]. \tag{4.102}$$

Remark. The expectation value and variance are these of a binomial distribution 4.2.22 with $r = X/N$.

4.7 Hypothesis testing

Definition 4.7.1 (Simple hypothesis). A hypothesis where the distribution is fully specified.

Definition 4.7.2 (Composite hypothesis). A hypothesis where the distribution is given relative to some parameter values.

4.7.1 Testing

Definition 4.7.3 (Type I error). Rejection of a true null hypothesis.

Definition 4.7.4 (Type II error). Acceptance of a false null hypothesis.

Definition 4.7.5 (Significance). The probability of making a type I error:

$$\alpha := \int P_{\text{I}}(x)dx. \tag{4.103}$$

Property 4.7.6. Let $\alpha_1 > \alpha_2$. An α_2-level test is also significant at the α_1-level.

Remark 4.7.7. For discrete distributions it is not always possible to achieve an exact level of significance.

Remark. Type I errors occur occasionally. They cannot be prevented, one can only try to control them.

Definition 4.7.8 (Power). The probability of not making a type II error:

$$\beta := \int P_{\text{II}}(x)dx \qquad \longrightarrow \qquad \text{power: } 1-\beta. \tag{4.104}$$

Remark 4.7.9. A good test is a test with a small significance and a large power. The probabilities P_{I} and P_{II} should be as different as possible.

Definition 4.7.10 (Likelihood ratio test). The null hypothesis $H_0 : \theta = \theta_0$ is rejected in favour of the alternative hypothesis $H_1 : \theta = \theta_1$ if the likelihood ratio Λ satisfies the following condition:

$$\Lambda(x) = \frac{\mathcal{L}(\theta_0|x)}{\mathcal{L}(\theta_1|x)} \leq \eta, \tag{4.105}$$

where $P(\Lambda(x) \leq \eta|H_0) = \alpha$.

Remark. In some references the reciprocal of Λ is used as the definition of the likelihood ratio.

Theorem 4.7.11 (Neyman-Pearson lemma). *The likelihood ratio test is the most powerful test at significance level α.*

Definition 4.7.12 (Family-wise error). Given a collection of hypothesis tests, the family-wise error is defined as the probability of making at least one type-I error.

Construction 4.7.13 (Bonferroni correction). Consider a set of hypotheses $\{H_i\}_{1\leq i\leq n}$. The higher the number of tests, the higher the chance that by statistical fluctuations at least one of these hypotheses will be rejected. To avoid this problem of multiple comparisons, one can try to control the family-wise error rate, i.e. the probability of falsely rejecting at least one hypothesis. The easiest way to control this error rate is by modifying the individual significance levels:

$$\alpha \longrightarrow \frac{\alpha}{n}. \tag{4.106}$$

4.8 Comparison tests

Definition 4.8.1 (McNemar test). Consider two models or hypotheses describing a given data set. Construct the contingency table describing the number of true positives and true negatives for both models:

	TP (model 1)	TN (model 1)
TP (model 2)	a	b
TN (model 2)	c	d

(4.107)

The null hypothesis of the McNemar test is that there is no significant difference between the predictive power of the model, i.e. $p_a + p_c = p_a + p_b$ and $p_b + p_d = p_c + p_d$ where p_i indicates the proportion of the variable i. By noting that the diagonal values are redundant in this description, one can write the hypotheses more concisely:

$$\begin{aligned} H_0 &: b = c \\ H_1 &: b \neq c. \end{aligned}$$

The test statistic is the McNemar chi-squared statistic:

$$\chi^2 = \frac{(b-c)^2}{b+c}. \tag{4.108}$$

When the values of b and c are large enough (> 25), one can approximate this distribution by an ordinary χ^2-distribution with 1 degree of freedom.

Remark 4.8.2 (Edwards correction). It is common to apply a continuity correction (similar to the *Yates-correction* for the ordinary chi-squared test):

$$\chi^2 := \frac{(|b-c|-1)^2}{b+c}. \tag{4.109}$$

This follows from the fact that for small b, c the exact p-values should be compared with a binomial test which compares b to $b + c$ (note the factor of 2):

$$p = 2\sum_{i=b}^{b+c}\binom{b+c}{i}0.5^i(1-0.5)^{b+c-i}. \tag{4.110}$$

Definition 4.8.3 (Wilcoxon signed-rank test). Consider a paired data sample, i.e. two dependent data samples for which the i^{th} entries are paired together. This test checks if the population means are different. The test statistic is defined as follows:

First, calculate the differences d_i and rank their absolute values (ties are assigned an average rank). Then, calculate the sums of the ranks R_+, R_- for positive and negative differences and take smallest of these:

$$T := \min(R_+, R_-). \tag{4.111}$$

For small data samples ($n < 25$) one can look up critical values in the literature. For larger data samples one can (approximately) use a standard normal distribution with statistic

$$z := \frac{T - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}.$$

Remark 4.8.4. The biggest benefit of this test over a signed t-test is that the Wilcoxon test does not require the data samples to be drawn from a normal distribution. However in the case where the assumptions for a paired t-test are met, the t-test is more powerful.

Definition 4.8.5 (Friedman test). Consider k models tested on N data sets. For every data set one ranks the models according to decreasing performance. For every $i \leq k$ one defines the average rank $R_i = \frac{1}{N}\sum_{j\leq N} r_i^j$, where r_i^j is the rank of the i^{th} model on the j^{th} data set. Under the null hypothesis "all models perform equally well", the average ranks should be the same for all models.

The Friedman statistic

$$\chi_F^2 := \frac{12N}{k(k+1)}\left(\sum_{i\leq k} R_i^2 - \frac{k(k+1)^2}{4}\right) \tag{4.112}$$

follows a χ^2-distribution with $k-1$ degrees of freedom when $N > 10$ and $k > 5$. For smaller values of these parameters one can look up the exact critical values in the literature.

Remark 4.8.6. It was shown that the original Friedman test is rather conservative and that a better statistic is

$$F := \frac{(N-1)\chi_F^2}{N(k+1)-\chi_F^2}. \tag{4.113}$$

This follows an F-distribution with $k-1$ and $(N-1)(k-1)$ degrees of freedom. As a further remark is that the (nonparametric) Friedman test is weaker than the (parametric) *repeated-measures ANOVA* whenever the assumptions for the latter hold (similar to the case of the Wilcoxon signed-rank test).

4.8.1 Post-hoc tests

After successfully using one of the multi-model tests from the previous section to reject the null hypothesis of equal performance, one is often interested in exactly which model outperforms the others. For this one can use one of the following pairwise tests:

Definition 4.8.7 (Nemenyi test). Consider the average ranks R_i from the Friedman test. As a test statistic one uses

$$z := \frac{R_i - R_j}{\sqrt{\frac{k(k+1)}{6N}}}, \tag{4.114}$$

where k is the number of models and N is the number of data sets. The exact critical values can either be found in the literature or one can approximately use a normal distribution.

Definition 4.8.8 (Bonferroni-Dunn test). If all one wants to do is see if a particular model performs better than a given baseline model, the Nemenyi test is too conservative since it corrects for $k(k-1)/2$ model comparisons instead of $k-1$. Therefore it is better to use a general method to control the family-wise error for multiple measurements. The Bonferroni-Dunn test modifies the Nemenyi test by performing a Bonferroni correction with $n-1$ degrees of freedom.

A more powerful test is given by the following strategy:

Definition 4.8.9 (Holm test). Consider the p-values of the Nemenyi test. Instead of comparing all values to a single Bonferroni-corrected significance, one can use a so-called "step-down" method. First one orders the p-values in ascending order and compares the smallest one to $\frac{\alpha}{k-1}$. If this value is significant, i.e. the hypothesis that the associated models perform equally well is rejected, one compares p_2 to $\frac{\alpha}{k-2}$ and so on until one finds a hypothesis that cannot be rejected. All remaining hypotheses are retained as well.

Remark 4.8.10. It is possible that the post hoc test fails to report a significant difference even though the Friedman test rejected the null hypothesis. This is a consequence of the lower power of post hoc tests.

4.9 Goodness of fit

Definition 4.9.1 (Akaike information criterion). Consider a model $f(x;\theta)$ with k parameters fitted to a given data sample and let $\mathcal{L}_0$ be the maximum of the associated likelihood function. The Akaike information criterion is defined a follows:

$$\text{AIC} := 2k - 2\ln(\mathcal{L}_0). \tag{4.115}$$

From this definition it is immediately clear that the AIC rewards goodness-of-fit but penalizes overfitting due to the first term.

This criterion is often useful when trying to select the best model/parameters to describe a certain data set. However it should be noted that it is not an absolute measure of quality.

4.9.1 χ^2-test

Property 4.9.2. If there are $N - n$ fitted parameters one has:

$$\int_{\chi^2}^{\infty} f(\chi^2|n)d\chi^2 \approx 1 \implies \begin{cases} \circ \text{ good fit} \\ \circ \text{ errors were overestimated} \\ \circ \text{ selected measurements} \\ \circ \text{ lucky shot} \end{cases} \tag{4.116}$$

Property 4.9.3 (Reduced chi-squared). The reduced chi-squared statistic is defined as follows:

$$\chi^2_{\text{red}} := \chi^2/n, \tag{4.117}$$

where n is the number of degrees of freedom. Depending on the value of this statistic one can draw the following conclusions (under the right assumptions):

- $\chi^2_{\text{red}} \gg 1$: poor modelling,
- $\chi^2_{\text{red}} > 1$: bad modelling or underestimation of the uncertainties,
- $\chi^2_{\text{red}} \approx 1$: good fit, or
- $\chi^2_{\text{red}} < 1$: (improbable) overestimation of the uncertainties.

4.9.2 Runs test

A good χ^2-test does not mean that the fit is good. As mentioned in Property 4.9.2, it is possible that the errors were overestimated. Another condition for a good fit is that the data points vary around the fit, i.e. there are no long sequences of points that lie above/underneath the fit. This condition is tested with a runs test 4.9.5.

Remark 4.9.4. The χ^2-test and runs test are complementary. The χ^2-test only takes the absolute value of the differences between the fit and data points into account, the runs test only takes the signs of the differences into account.

Formula 4.9.5 (Runs distribution). Let N_+ and N_- denote the number of points above and below the fit. Under the hypothesis that all points were independently drawn from the same distribution the number of runs is distributed as follows (approximately Gaussian):

$$\begin{aligned} P(r_{\text{even}}) &= 2\frac{\binom{N_+-1}{\frac{r}{2}-1}\binom{N_--1}{\frac{r}{2}-1}}{\binom{N}{N_+}} \\ P(r_{\text{odd}}) &= \frac{\binom{N_+-1}{\frac{r-3}{2}}\binom{N_--1}{\frac{r-1}{2}}+\binom{N_--1}{\frac{r-3}{2}}\binom{N_+-1}{\frac{r-1}{2}}}{\binom{N}{N_+}}, \end{aligned} \tag{4.118}$$

where C_k^n is the binomial coefficient $\binom{n}{k}$. The first two moments of this distribution are given by the following formulas:

$$\text{E}[r] = 1 + 2\frac{N_+N_-}{N} \tag{4.119}$$

$$\text{Var}[r] = 2\frac{N_+N_-}{N}\frac{2N_+N_- - N}{N(N-1)}. \tag{4.120}$$

Remark 4.9.6. For $r > 15$, the runs distribution approximates a Gaussian distribution.

Chapter 5

Data Analysis

The main reference for the sections on optimization problems is [116]. For the geometry of clustering methods, see [69]. The main references for the section on *conformal prediction* are [8, 78]. Although a part of this chapter is a continuation of the previous one, the focus here lies more on the computational aspect of the analysis of large data sets. For this reason the chapter starts with some sections on applied linear algebra (for a refresher see Chapter **??**).

5.1 Optimization

5.1.1 Linear equations

Method 5.1.1 (Normal equation). Given the equation

$$Ax = b$$

as in Section **??**, one can try to numerically solve for x by minimizing the ℓ^2-norm $\|Ax - b\|^2$:

$$\hat{x} := \arg\min_x (Ax - b)^T (Ax - b). \tag{5.1}$$

This leads to the so-called normal equation[1]

$$A^T Ax = A^T b. \tag{5.2}$$

This can be formally be solved by $x = (A^T A)^{-1} A^T b$, where $(A^T A)^{-1} A^T$ is the pseudoinverse of A.

Remark 5.1.2. It is easy to see that the above linear problem is obtained when trying to extremize the quadratic form associated to a symmetric matrix.

Method 5.1.3 (Tikhonov regularization). Consider a linear (regression) problem

$$Ax = b.$$

The most straightforward way to solve for x is the least squares method introduced in Chapter 4, where the solution is (formally) given by the normal equation: $x = (A^T A)^{-1} A^T b$. However, sometimes it might happen that A is nearly singular (it is said to be **ill-conditioned**). In this case a regularization term can be added to the minimization problem:

$$\|Ax - b\|^2 + \|\Gamma x\|^2, \tag{5.3}$$

where Γ is called the **Tikhonov matrix**. In the case that $\Gamma = \lambda \mathbb{1}$, one speaks of **$\ell^2$-regularization**. This regularization technique benefits solutions with smaller norms.

[1] The name stems from the fact that the equation $A^T Ax = A^T b$ implies that the residual is orthogonal (normal) to the range of A.

Remark 5.1.4. The ℓ^2-regularization can be generalized by replacing the 2-norm by any p-norm $\|\cdot<|_p$. For $p=1$ and $p=2$ the names **lasso** and **ridge** regression are often used. For general $p\geq 0$ one sometimes speaks of **bridge** regression.

The minimization procedures for $p\leq 1$ have the important property that they not only shrink the coefficients, but even perform feature selection, i.e. some coefficients become identically zero. However, it can be shown that the optimization problem for $p<1$ is nonconvex and, hence, is harder to solve. In general it is found that lasso regression gives the best results.

A benefit of ℓ^2-regularization is that it can be derived from a Bayesian approach. By choosing a Gaussian prior $\mathcal{N}(0,\lambda^{-1})$, Bayesian inference immediately gives the ℓ^2-regularized cost function as the posterior distribution. Accordingly the ℓ^2-regularized linear regressor is equivalent to the maximum a posteriori estimator with Gaussian priors. One can obtain ℓ^p-regularization in a similar way by replacing the Gaussian priors with generalized normal distributions (such as the Laplace distribution for $p=1$).

Definition 5.1.5 (Multicollinearity). Consider a finite set of random variables $\{X_i\}_{1\leq i\leq n}$. These random variables are said to be perfectly (multi)collinear if there exists an affine relation between them, i.e. there exist variables $\{\lambda_i\}_{0\leq i\leq n}$ such that

$$\lambda_0+\lambda_1X_1+\cdots+\lambda_nX_n=0. \tag{5.4}$$

The same concept can be applied to data samples. The data is said to be (multi)collinear if the above equation holds for all entries of the data set. However, in this case one also define "near multicollinearity" if the variables X_i are related as above up to some error term ε. If the variance of ε is small, the matrix X^TX might have an ill-conditioned inverse which might render the algorithms unstable.

Definition 5.1.6 (Variance inflation factor). The VIF is an estimate for how much the variance of a coefficient is inflated by multicollinearity. The VIF of a coefficient β_i is defined as follows:

$$\mathrm{VIF}_i:=\frac{1}{1-R_i^2}, \tag{5.5}$$

where R_i^2 is the R^2-value obtained after regressing the predictor $\hat{X}_i$ on all other predictors. The rule of thumb is that VIF ≥ 10 implies that a significant amount of multicollinearity is present in the model.

5.1.2 Descent methods

The gradient descent algorithm is first introduced in the case of quadratic forms:

Method 5.1.7 (Steepest descent). Consider the quadratic form

$$f(x)=\frac{1}{2}x^TAx-b^Tx+c.$$

Assume that A is symmetric and positive-definite such that $Ax=b$ gives the minimum of f. Like most recursive algorithms, gradient descent starts from an arbitrary guess x_0. It then takes a step in the direction of steepest descent (or largest gradient), i.e. in the direction opposite to $f'(x_0)=Ax_0-b=:-r_0$:

$$x_{i+1}:=x_i+\alpha r_i. \tag{5.6}$$

The quantities r_i are called the **residuals**. This procedure is repeated until convergence, i.e. until the residual vanishes up to a fixed numerical tolerance.

A naive gradient descent method would require to fine-tune the step size α. However, a more efficient method is given by the **line search algorithm**, where the value of α is optimized in every step as to minimize f along the line defined by r_i. A standard calculus argument leads to the following form of the step size:

$$\alpha_i = \frac{r_i^T r_i}{r_i^T A r_i}. \tag{5.7}$$

This choice forces the descent direction to be orthogonal to the previous one since $\frac{d}{d\alpha} f(x_i) = -f'(x_i)^T f'(x_{i-1})$. As a consequence, this minimization scheme often results in a chaotic zigzag trajectory through the configuration space. The higher the **condition number** $\kappa = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$, the worse the zigzag motion will be. A very narrow valley (or some higher-dimensional analogue) will make the trajectory bounce back and forth between the walls, instead of moving towards the minimum.

5.1.3 Conjugate gradient

As noted in the previous section, a common problem with gradient descent is that the direction of steepest descent is often not the same as the direction pointing to the optimal solution and, hence, convergence might only occur after a long time.

A simple solution can be obtained by considering multiple orthogonal directions and taking a suitable step once in every direction. This way one obtains an algorithm that converges in n steps, where n is the dimension of the coefficient matrix A. By requiring that the error at step $i+1$ is orthogonal to the direction d_i, it is assured that no direction is used twice. However, the main problem with this idea is that the exact error e_i is not known and, hence, one cannot calculate the required steps.

By modifying the orthogonality condition one can avoid this problem. This is the idea behind conjugate direction methods:

Definition 5.1.8 (Conjugate vectors). Consider a symmetric positive-definite matrix A. Any such matrix induces an inner product as follows:

$$\langle v|w\rangle_A := v^T A w. \tag{5.8}$$

Two vectors v, w are said to be (A-)conjugate if they are orthogonal with respect to $\langle\cdot|\cdot\rangle_A$. The general approach to obtain a basis of A-conjugate vectors is a modified version of the Gram-Schmidt procedure **??** where the ordinary Euclidean inner product is replaced by (5.8). This modification is called the **Arnoldi method**.

By taking the input vectors of the Arnoldi method to be the residuals r_i, one obtains the **conjugate gradient** (CG) algorithm. It is interesting to note that the residuals themselves satisfy a recursion relation:

$$r_{i+1} = r_i - \alpha_i A d_i, \tag{5.9}$$

where the step size α_i is defined similar to the step size for ordinary steepest descent:

$$\alpha_i = \frac{d_i^T r_i}{d_i^T A d_i}. \tag{5.10}$$

Since the directions are constructed using the residuals, they span the same subspace. By denoting the subspace spanned by the first i directions by $\mathcal{D}_i$, the relation $r_{i+1} \in \mathcal{D}_i + Ad_i$ leads to the following expression because of the above recursion relation:

$$\mathcal{D}_i = \operatorname{span}\{r_0, Ar_0, \ldots, A^{i-1}r_0\}. \tag{5.11}$$

Because of their prominence in the literature on numeric optimization techniques, these subspaces have earned their own name:

Definition 5.1.9 (Krylov subspace). A vector space $\mathcal{K}$ of the form

$$\mathcal{K} := \operatorname{span}\{v, Av, \dots, A^n v\} \tag{5.12}$$

for some matrix A, vector v and natural number $n \in \mathbb{N}$. Given such an A and v, one often denotes the associated Krylov subspace of dimension n by $\mathcal{K}_n(A, v)$.

The fact that the spaces $\mathcal{D}_i$ are Krylov spaces also has an import implication for the numerical complexity of the CG algorithm. The residual r_{i+1} can be shown to be orthogonal to the space $\mathcal{D}_{i+1}$ (this is generally called the **Galerkin condition**). But since $A\mathcal{D}_i \subset \mathcal{D}_{i+1}$, this also implies that r_{i+1} is A-conjugate to $\mathcal{D}_i$. It follows that the only relevant contribution in the Arnoldi method is given by the last direction d_i. This reduces the complexity (both time-wise and memory-wise) per iteration from $O(n^2)$ to $O(n)$.

The steps in the CG algorithm are summarized below:

Method 5.1.10 (Conjugate gradient). Let x_0 be the initial guess with the associated residual $r_0 := b - Ax_0$ acting as the first direction vector d_0. The following scheme gives an iterative n-step (n being the dimension of the coefficient matrix A) algorithm to obtain the solution to $Ax = b$:

$$\alpha_i := \frac{r_i^T r_i}{d_i^T A d_i} \tag{5.13}$$

$$x_{i+1} := x_i + \alpha_i d_i \tag{5.14}$$

$$r_{i+1} := r_i - \alpha_i A d_i \tag{5.15}$$

$$d_{i+1} := r_{i+1} + \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i} d_i. \tag{5.16}$$

Remark 5.1.11. In exact arithmetic the above optimization scheme would result in an exact solution after n iterations (in fact the number of iterations is bounded by the number of distinct eigenvalues of A). However, in real life one is not working in exact arithmetic and one has to take into account the occurrence of floating-point errors. These not only ruin the accuracy of the residual recursion relation (5.15), but more importantly[2] it might result in the search directions not being A-conjugate.

Now, what about general coefficient matrices A, for example those resulting in under- or overdetermined systems? For nonsymmetric or nondefinite square matrices one can still solve the normal equation (5.2) using the same methods, since $A^T A$ is both symmetric and positive-definite. For underdetermined systems an exact solution does not always exact, but the numerical methods will always be able to find a solution that minimizes the ℓ^2-error. For overdetermined systems $A^T A$ will be nonsingular and the numerical methods can find an exact solution. However, the condition number of $A^T A$ is the square of that of A and, hence, the algorithms will convergence much slower.

A different approach exists where the CG algorithm is not applied to the matrix $A^T A$, but the individual matrices are used A, A^T directly. This way not one Krylov space is generated, but two dual "copies" are constructed:

$$\begin{aligned}\mathcal{D}_i &:= \operatorname{span}\big\{r_0, Ar_0, \dots, A^{i-1}r_0\big\}\\ \widetilde{\mathcal{D}}_i &:= \operatorname{span}\big\{\widetilde{r}_0, A^T\widetilde{r}_0, \dots, (A^T)^{i-1}\widetilde{r}_0\big\},\end{aligned}$$

[2] The residual problem can be solved by computing the residual "exactly", i.e. by the formula $r_i = b - Ax_i$, every k iterations.

where $\widetilde{r}_0$ does not have to be related to r_0. In this case there are two Galerkin conditions $r_i \perp \mathcal{D}_i$ and $\widetilde{r}_i \perp \widetilde{\mathcal{D}}_i$ (only the first one is relevant). The residuals form biorthogonal bases of the Krylov subspaces:

$$\langle r_i | r_j \rangle = \|r_i\|^2 \delta_{ij}. \tag{5.17}$$

As a consequence the search directions also form biconjugate bases:

$$\langle d_i | d_j \rangle_A = \|d_i\|_A^2 \delta_{ij}. \tag{5.18}$$

5.1.4 Nonlinear conjugate gradients

Of course, many real-world applications are determined by nonlinear equations and, hence, it would be pleasant if one could salvage some of the above ideas even when linear algebra is not the natural language. The main requirement would be that one can calculate the gradient of the function to be minimized.

On the level of the implementation, the structure of the algorithm remains more or less the same. What does change is the form of the Arnoldi method, in particular, the prefactor in Equation (5.16). For linear CG there are multiple equivalent formulas, but for nonlinear CG these do not lead to the same algorithm. The two most common choices are given below.

Method 5.1.12 (Nonlinear CG). Since there is no linear equation related to the minimization problem, the residuals are always defined as $r_i := -f'(x_i)$. The algorithm consists of the following iterations:

$$\alpha_i := \arg\min_\alpha f(x_i + \alpha d_i) \tag{5.19}$$

$$x_{i+1} := x_i + \alpha_i d_i \tag{5.20}$$

$$r_{i+1} := -f'(x_i) \tag{5.21}$$

$$d_{i+1} := r_{i+1} + \beta_{i+1} d_i, \tag{5.22}$$

where β_{i+1} is computed by one of the following formulas:

- **Fletcher-Reeves formula**:

$$\beta_{i+1} := \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i}. \tag{5.23}$$

- **Polak-Ribière formula**:

$$\beta_{i+1} := \max\left\{ \frac{r_{i+1}^T (r_{i+1} - r_i)}{r_i^T r_i}, 0 \right\}. \tag{5.24}$$

Some general remarks have to be made concerning the nonlinear CG algorithm:

Remark 5.1.13. As was already mentioned for the linear version, floating-point errors might lead to a loss of conjugacy. For the nonlinear extension this becomes worse. The more f deviates from a quadratic function, the quicker conjugacy is lost (for quadratic formulas the Hessian is exactly the matrix A, but for higher-degree functions the Hessian varies from point to point). Another problem, one that did not occur for quadratic functions, is that nonlinear functions might have multiple local minima. The CG method does not care about local vs. global and, hence, it will not necessarily converge to the global minimum. A last remark concerns the fact that there is no theoretical guarantee that the method will converge in n steps. Since the Gram-Schmidt procedure can only construct n conjugate vectors, the simplest solution is to perform a restart of the algorithm every n iterations.[3]

[3] The max operation in Equation (5.24) is already a form of restarting, due to the fact that the Polak-Ribière version of nonlinear CG sometimes results in cyclic behaviour.

For linear CG a simple formula for finding the optimal value of α_i was obtained. However, for nonlinear CG one cannot solve Equation (5.19) as easily. The main idea, i.e. that f' should be orthogonal to the previous search direction remains, is still valid. Here, only the **Newton-Raphson approach** is considered:[4]

$$\alpha_i = \frac{f'(x_i)^T d_i}{d_i^T f''(x_i) d_i}. \tag{5.25}$$

To obtain the optimal α-value, one should iteratively apply the Newton-Raphson method in every CG iteration. If the action of the Hessian f'' on d_i cannot be simplified, i.e. if the full Hessian has to be computed in every iteration, this can lead to considerable computational overhead. The general rule of thumb is to perform only a few Newton-Raphson iterations and obtain a less accurate but more efficient algorithm. To make sure that the search descent direction is indeed a direction of descent (and not one of ascent), one can check that $r^T d \geq 0$ and restart the procedure if it is negative.

5.1.5 Krylov methods

Generally one starts from an iterative fixed-point based technique to solve the linear equation $Ax = b$ as before, i.e. one iterates $x_{i+1} = b + (\mathbb{1} - A)x_i$. Using the residuals $r_i = b - Ax_i$ this can be rewritten as

$$x_i = x_0 + \sum_{k=0}^{i-1} r_k = x_0 + \sum_{k=0}^{i-1} (\mathbb{1} - A)^k r_0. \tag{5.26}$$

It is clear that this results in $x_i - x_0 \in \mathcal{K}_i(A, r_0)$. The main idea is then to find optimal degree-k polynomials P_k such that $x_i - x_0 = \sum_{k=0}^{i-1} P_k(A) r_0$.

Method 5.1.14 (Jacobi method). Consider a linear problem $Ax = b$ where A has spectral radius less than 1. First, decompose A as the sum of a diagonal matrix D and and a matrix E with zero diagonal elements. If one assumes that D is invertible, the following recursive scheme is obtained:

$$x_{i+1} := D^{-1}(b - Ex_i). \tag{5.27}$$

A sufficient condition for convergence is strict diagonal dominance, i.e. $|D_{ii}| > \sum_{j\neq i} |E_{ij}|$.

?? COMPLETE (e.g. Lanczos)??

5.1.6 Constrained optimization

A common generalization of the above optimization problems is the addition of constraints involving equalities:

$$\arg\min_x f(x) \qquad \text{such that} \qquad g_i(x) = 0 \qquad \forall 1 \leq i \leq n. \tag{5.28}$$

The general approach to solving such constrained problems is by extending the optimization loss:

Method 5.1.15 (Lagrange multipliers). Given a constrained optimization problem of the form (5.28), one can construct the enhanced loss function

$$\mathcal{L}(x, \lambda_1, \ldots, \lambda_n) := f(x) + \sum_{i=1}^{n} \lambda_i g_i(x). \tag{5.29}$$

[4] Another common method is the *secant method*.

The solution to the original problem is obtained by extremizing this loss with respect to x and the Lagrange multipliers λ_i (as usual this might fail globally for nonconvex problems):

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial x} = 0 \\ \\ \dfrac{\partial\mathcal{L}}{\partial\lambda_i} = 0 \quad \forall 1 \leq i \leq n. \end{cases} \tag{5.30}$$

The situation becomes even more interesting when one also allows constraints involving inequalities:

$$\arg\min_x f(x) \qquad \text{such that} \qquad \begin{cases} g_i(x) = 0 & \forall 1 \leq i \leq m \\ h_j(x) \leq 0 & \forall 1 \leq j \leq n. \end{cases} \tag{5.31}$$

Problems of this form are called **primal optimization problems**. By defining an enhanced loss using Lagrange multipliers as before

$$\mathcal{L}(x,\alpha,\beta) := f(x) + \sum_{i=1}^m \alpha_i g_i(x) + \sum_{i=1}^n \beta_i h_i(x), \tag{5.32}$$

it is not hard to see that

$$\max_{\alpha,\beta;\beta_j\geq 0} \mathcal{L}(x,\alpha,\beta) = \begin{cases} \infty & \text{if a constraint is violated} \\ f(x) & \text{if all constraints are satisfied.} \end{cases} \tag{5.33}$$

Definition 5.1.16 (Primal optimization problem). Denote the maximum of $\mathcal{L}(x,\alpha,\beta)$ by $\theta_P(x)$.

$$p^* := \min_x \theta_P(x) = \min_x \max_{\alpha,\beta;\beta_i\geq 0} \mathcal{L}(x,\alpha,\beta). \tag{5.34}$$

By interchanging the max and min operators in the primal formulation, another problem is obtained:

Definition 5.1.17 (Dual optimization problem).

$$d^* := \max_{\alpha,\beta;\beta_i\geq 0} \theta_D(\alpha,\beta) = \max_{\alpha,\beta;\beta_i\geq 0} \min_x \mathcal{L}(x,\alpha,\beta). \tag{5.35}$$

From basic calculus it is known that $\max\min \leq \min\max$ and, hence, that $d^* \leq p^*$. The difference $p^* - d^*$ is called the **duality gap** and, if $d^* = p^*$, one says that **strong duality** holds. The real question then becomes: "*When does strong duality hold?*".

Definition 5.1.18 (Slater conditions). Consider a convex optimization problem, i.e. a problem of the form (5.31) where f is convex, the g_i are convex and the h_j are affine. This problem is said to satisfy the Slater condition(s) if there exists an x that is strictly **feasible**, i.e. $h_j(x) < 0$ for all $1 \leq j \leq n$.

Property 5.1.19 (Strong duality). If a convex problem satisfies the Slater conditions, strong duality holds. The solutions x and (α,β) that attain this duality are called primal optima and dual optima respectively.

The following property gives a set of sufficient conditions:

Property 5.1.20 (Karush-Kuhn-Tucker conditions). If there exist x, α and β such that strong duality holds, the following conditions are satisfied:

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial x} = 0 \\ \\ \dfrac{\partial\mathcal{L}}{\partial\alpha_i} = 0 \end{cases} \qquad \forall 1 \leq i \leq m \qquad \text{and} \qquad \begin{cases} \beta_j h_j(x) = 0 \\ h_j(x) \leq 0 \\ \beta_j \geq 0. \end{cases} \qquad \forall 1 \leq j \leq n \tag{5.36}$$

Conversely, if there exists values x, α and β that satisfy the KKT conditions, they give strongly dual solutions for the primal and dual problems.

Remark 5.1.21 (Complementary slackness). The third equation in the KKT conditions has an important implication. It says that if there is an index j such that the constraint h_j is not **active**, i.e. $h_j(x) < 0$, the associated Lagrange multiplier is 0 and, conversely, if there is an index j such that the Lagrange multiplier $\beta_j > 0$, the constraint h_j is active.

Remark 5.1.22. It is not hard to see that the KKT conditions reduce to the conditions for Lagrange multipliers when all h_j are identically 0. For this reason the quantities α and β are called the **KKT multipliers**.

5.2 Classification problems

5.2.1 Clustering

Probably the most well-known and simplest algorithm for clustering in the unsupervised setting is the k-means algorithm:

Method 5.2.1 (k-means algorithm). Assume that an unlabelled dataset $\mathcal{D} \subset \mathbb{R}^n$ is given. For every integer $k \in \mathbb{N}$, usually satisfying $k \ll |\mathcal{D}|$, and any choice of k distinct **centroids** $\{c_i \in \mathbb{R}^n\}_{i\leq k}$, the k-means algorithm is defined through the following iterative scheme:

1. To every point $d \in \mathcal{D}$ assign a cluster C_i based on the following criterion:

$$i = \arg\min_{j\leq k} \|d - c_j\|^2. \tag{5.37}$$

2. Update the centroids c_i to represent the center of mass of the associated cluster C_i:

$$c_i \longleftarrow \frac{1}{|C_i|}\sum_{d\in C_i} d. \tag{5.38}$$

This algorithm optimizes the following global cost function with respect to the centroids c_i:

$$\mathcal{L}_{k\text{-means}}(c_1, \ldots, c_k) = \sum_{i=1}^{k}\sum_{d\in C_i} \|d - c_i\|^2. \tag{5.39}$$

Given the above idea, one could ask for a more general algorithm where clustering is performed with respect to a divergence function **??**. In the case of Bregman divergences **??** it can be shown that all one needs to do is replace the Euclidean distance by the divergence D_f:

Property 5.2.2 (Centroid position). Let D_f be a Bregman divergence. The minimizer

$$\arg\min_{\kappa} \sum_{i=1}^{k} D_f(x_i\|\kappa) \tag{5.40}$$

is given by the arithmetic average

$$\kappa = \frac{1}{k}\sum_{i=1}^{k} x_i. \tag{5.41}$$

If instead of a cluster $C = \{x_i \in \mathbb{R}^n\}_{i\leq k}$, one is given a probability distribution p, one simply has to replace the arithmetic average by the expectation value with respect to p. It can be furthermore be shown that for any Bregman divergence the k-means algorithm always converges in a finite number of steps (however, the clustering is not necessarily optimal).

The cluster boundaries $H(c_1, c_2) = \{x \in \mathbb{R}^n \mid D_f(x\|c_1) = D_f(x\|c_2)\}$ admit a simple geometric construction:

Property 5.2.3 (Cluster boundaries). Let D_f be a Bregman divergence and consider the k-means problem associated to D_f for $k = 2$ (higher-dimensional problems can be treated similarly). The boundary $H(c_1, c_2)$ is exactly the dual geodesic hypersurface orthogonal to the affine geodesic connecting c_1 and c_2. This partitioning of the data manifold is a generalization of *Voronoi diagrams* to (Bregman) divergences.[5]

5.2.2 Nearest neighbour search

?? COMPLETE ??

5.3 Garden

?? ADD (e.g. trees, forests)??

5.4 Support-vector machines

5.4.1 Kernel methods

This section will introduce the mathematics of kernel methods. This mainly involves the language of Hilbert spaces (see Chapter **??** for a refresher).

Definition 5.4.1 (Kernel[6]). A function $k : X \times X \to \mathbb{C}$ that is (conjugate) symmetric and for which the Gram-matrix $K_{ij} := K(x_i, x_j)$ is positive-definite for all $n \in \mathbb{N}$ and $\{x_i \in X\}_{i\leq n}$.

Definition 5.4.2 (Reproducing kernel Hilbert space). A Hilbert space $\mathcal{H} \subset \mathrm{Map}(X, \mathbb{C})$ for some set X for which all evaluation functionals $\delta_x : f \mapsto f(x)$ are bounded (or continuous by Property **??**). Reproducing kernel Hilbert spaces are often abbreviated as **RKHS**s.

Using the Riesz representation theorem **??** one can express every evaluation functional δ_x on a RKHS $\mathcal{H}$ as a function $K_x \in \mathcal{H}$. This allows for the introduction of a kernel on X:

Definition 5.4.3 (Reproducing kernel). Let $\mathcal{H}$ be a RKHS on a set X. The (reproducing) kernel k on X is defined as follows:

$$k(x, y) := \delta_x(K_y) \overset{\text{Riesz}}{=} \langle K_x | K_y\rangle_{\mathcal{H}}. \tag{5.42}$$

Because k is given by a metric, it is not hard to see that the reproducing kernel is a Mercer kernel 5.4.1.

[5] See [70] for more information. This is also introduced in [12], but there the author has confusingly interchanged the affine and dual coordinates.

[6] Also called a **Mercer kernel**. See Mercer's theorem below for more information.

Starting from a kernel one can also characterize an RKHS as follows:

Alternative Definition 5.4.4 (RKHS). A Hilbert space $\mathcal{H} \subset \text{Map}(X, \mathbb{C})$ of functions over a set X such that there exists a kernel k on X with the following properties:

1. **Reproducing property**: For all $x \in X, f \in \mathcal{H}$ the evaluation functional δ_x satisfies $\delta_x(f) = \langle k(\cdot, x)|f\rangle_{\mathcal{H}}$.
2. **Density**: The span of $\{k(\cdot, x) \mid x \in X\}$ is dense in $\mathcal{H}$.

The density property is often replaced by the property that $k(\cdot, x) \in \mathcal{H}$ for all $x \in X$.

Property 5.4.5 (Convergence). In an RKHS, convergence in norm implies pointwise convergence.

Theorem 5.4.6 (Moore-Aronszajn). *There exists a bijection between RKHSs and kernels.*

Proof. One direction of the theorem is, as mentioned before, rather simple to see. The other direction is constructive:

> Given a kernel k, one defines for all $x \in X$ the function $K_x := k(\cdot, x)$. The RKHS is then constructed as the Hilbert completion of $\text{span}\{k_x \mid x \in X\}$, where the inner product is defined as follows
>
> $$\left\langle \sum_{x\in X} a_x K_x \middle| \sum_{y \in X} b_y K_y \right\rangle_{\mathcal{H}} := \sum_{x,y\in X} \overline{a_x} b_y k(x, y). \tag{5.43}$$

Formula 5.4.7. Let $\mathcal{H}$ be an RKHS with kernel k. If $\{e_i\}_{i\leq\dim(\mathcal{H})}$ is an orthonormal basis for $\mathcal{H}$, then

$$k(x, y) = \sum_{i=1}^{\dim(\mathcal{H})} e_i(x)\overline{e_i(y)}. \tag{5.44}$$

Remark 5.4.8. Note that one can use different conventions in the above definitions, e.g. the definition $k(x, y) = \langle K_y|K_x\rangle_{\mathcal{H}}$ is also valid.

Theorem 5.4.9 (Mercer). *Let X be a finite measure space and consider a (conjugate) symmetric function $k \in L^2(X \times X, \mathbb{C})$. If k satisfies the* ***Mercer condition***

$$\iint_{X\times X} k(x, y)\overline{f(x)}f(y)dxdy \geq 0 \tag{5.45}$$

for all $f \in L^2(X, \mathbb{C})$, the Hilbert-Schmidt operator

$$T_k : L^2(X, \mathbb{C}) \to L^2(X, \mathbb{C}) : f \mapsto \int_X k(\cdot, x)f(x)dx \tag{5.46}$$

admits a countable orthonormal basis $\{e_i\}_{i\in\mathbb{N}}$ with nonnegative eigenvalues $\{\lambda_i\}_{i\in\mathbb{N}}$ such that

$$k(x, y) = \sum_{i=1}^{\infty} \lambda_i e_i(x)\overline{e_i(y)}. \tag{5.47}$$

Theorem 5.4.10 (Bochner). *A continuous function satisfies the Mercer condition if and only if it is a kernel.*

Alternative Definition 5.4.11 (Kernel). Consider a set X. A function $k : X \times X \to \mathbb{C}$ is called a kernel on X if there exists a Hilbert space $\mathcal{H}$ together with a function $\phi : X \to \mathcal{H}$ such that

$$k(x, y) = \langle\phi(x)|\phi(y)\rangle_{\mathcal{H}}. \tag{5.48}$$

When using Mercer's theorem, the feature maps are given by $\phi_i : x \mapsto \sqrt{\lambda_i}e_i(x)$.

5.4.2 Decision boundaries

Consider a linear model for a classification problem $y = w^T x + b$. The object x_i is said to belong to the positive (resp. negative) class if $y > 0$ (resp. $y < 0$). This is implemented by the sign activation function

$$\mathrm{sgn}(y) = \begin{cases} 1 & y > 0 \\ -1 & y < 0 \end{cases} \tag{5.49}$$

to the linear model. The **decision boundary** $y = 0$, where the decision becomes ambiguous, forms a hyperplane in the feature space. However, it should be clear that in generic situations there are multiple hyperplanes that can separate the two classes for a finite number of data points. The problem then becomes to obtain the hyperplane with the maximal separation, i.e. the hyperplane for which the distance to the nearest data point is maximal.

The unit vector $\frac{w}{\|w\|}$ defines the normal to the hyperplane and, therefore, one can obtain the distance $d(x)$ from a data point x to the decision boundary by projecting onto this unit vector. The point $x - d(x)\frac{w}{\|w\|}$ is an element of the decision boundary and, hence, satisfies the hyperplane equation. Rewriting this gives an expression for the distance

$$d(x) = \frac{w^T x + b}{\|w\|}. \tag{5.50}$$

To account for the direction of the arrow, this number should be multiplied by the class $\mathrm{sgn}(y) = \pm 1$. This result is called the **geometric margin** $\gamma(x) := \mathrm{sgn}(y)d(x)$. The numerator in the geometric margin is called the **functional margin**. The geometric margin is preferable since it is invariant under simultaneous scale transformations of the parameters w, b.

The optimization objective now becomes

$$\max_w \frac{\gamma}{\|w\|} \qquad \text{such that} \qquad y_i(w^T x_i + b) \geq \gamma\|w\| \qquad \forall 1 \leq i \leq n, \tag{5.51}$$

where $\gamma = \min_{i\in\{1,\dots,n\}} \gamma(x_i)$ for x_i ranging over the training set. The problem is formulated in terms of the functional margin $\gamma\|w\|$ to avoid the nonconvex constraint $\|w\| = 1$. This allows the application of the Slater conditions for strong duality. Since the geometric margin is invariant under scale transformations, one can without loss of generality work with the assumption $\gamma\|w\| = 1$. The optimization problem is then equivalent to the following minimization problem:

$$\min_w \|w\|^2 \qquad \text{such that} \qquad y_i(w^T x_i + b) \geq 1 \qquad \forall 1 \leq i \leq n. \tag{5.52}$$

The KKT conditions for this problem give the following results:

$$w = \sum_{i=1}^n \beta_i y_i x_i \tag{5.53}$$

and

$$\sum_{i=1}^n \beta_i y_i = 0, \tag{5.54}$$

where the quantities β_i are the KKT multipliers for the affine constraints $1 - y_i(w^T x_i + b) \leq 0$. Using these relations the quantity y can be expressed for a new data point as follows:

$$y \equiv w^T x + b = \sum_{i=1}^n \beta_i y_i \langle x_i | x\rangle + b. \tag{5.55}$$

Two observations can be made at this point. First of all, complementary slackness 5.1.21 implies that the only relevant vectors x_i in this calculation are the ones that satisfy $\gamma(x_i) = 0$. These are called the **support vectors** and they give their name to a class of models called **support-vector machines** (SVMs). These are the models that are trained using the above optimization problem. Furthermore, y can be written in terms of an inner product. It is exactly this last observation that allows for the generalization of the above model to nonlinear decision boundaries. The previous section showed that inner products are equivalent to (Mercer) kernels. Hence, by choosing a nonlinear kernel function, one can implicitly work with nonlinear feature maps. This is often called the **kernel trick**. As an example, polynomial kernels represent feature maps from x to monomials in the coefficients of x.

However, as often happens with data analysis algorithms, this procedure is sensitive to outliers. This is especially the case for kernels that are based on feature maps to infinite-dimensional spaces (e.g. the *RBF kernel*). To solve this problem one can introduce a regularization term in the cost function. The simplest such term for support-vector machines is a simple ℓ^1-penalty:

$$\min_w \|w\|^2 + C\sum_{i=1}^n \xi \qquad \text{such that} \qquad \begin{cases} \xi_i \geq 0 & \forall 1 \leq i \leq n \\ y_i(w^Tx_i + b) \geq 1 - \xi_i & \forall 1 \leq i \leq n. \end{cases} \tag{5.56}$$

The resulting KKT conditions are as follows:

$$0 \leq \beta_i \leq C \tag{5.57}$$

and

$$\beta_i = 0 \implies y_i(w^Tx_i + b) \geq 1 \tag{5.58}$$

$$\beta_i = C \implies y_i(w^Tx_i + b) \leq 1 \tag{5.59}$$

$$\beta_i \in \,]0, C[\, \implies y_i(w^Tx_i + b) = 1 \tag{5.60}$$

?? COMPLETE (e.g. geometry)??

5.5 Time series analysis

Definition 5.5.1 (Time series). A $\mathbb{N}$- or $\mathbb{Z}$-indexed stochastic process. Since $\mathbb{N}$ and $\mathbb{Z}$ are isomorphic in a very simple way, the two conventions for time series will be used interchangeably.

5.5.1 Stationarity

Definition 5.5.2 (Strict stationarity). A time series $(X_n)_{n\in\mathbb{N}}$ is (strictly) stationary if for any two integers $r, s \in \mathbb{N}$, the joint distribution satisfies the following condition:

$$P(X_{t_1}, \ldots, X_{t_r}) = P(X_{t_1+s}, \ldots, X_{t_r+s}). \tag{5.61}$$

Definition 5.5.3 (Weak stationarity). A time series $(X_n)_{n\in\mathbb{N}}$ is said to be weakly (or **covariance**) stationary if it satisfies the following conditions:

1. **Mean-stationary**: $\mathrm{E}[X_n] = \mathrm{E}[X_0]$ for all $n \in \mathbb{N}$.
2. **Finite covariance**: $\mathrm{cov}(X_i, X_j) < \infty$ for all $i, j \in \mathbb{N}$.
3. **Covariance-stationary**: $\mathrm{cov}(X_i, X_{i+j}) = \mathrm{cov}(X_0, X_j)$ for all $i, j \in \mathbb{N}$.

The following definition is a reformulation of Birkhoff ergodicity 2.2.26:

Definition 5.5.4 (Ergodicity). A time series $\{X_t\}_{t\in\mathbb{Z}}$ is ergodic if for every measurable function f the following equation holds for all $t \in \mathbb{Z}$:

$$\lim_{T\to\infty} \frac{1}{2T+1} \sum_{k=-T}^{T} f(X_k) = \mathrm{E}[f(X_t)]. \tag{5.62}$$

Intuitively this means that state space averages can be evaluated as time averages.

5.5.2 Correlation

Definition 5.5.5 (Autocorrelation function). Consider a time series $(X_n)_{n\in\mathbb{N}}$. The autocovariance (resp. autocorrelation) function of this time series is defined as the covariance (resp. autocorrelation) function of the random variables $(X_n)_{n\in\mathbb{N}}$.

Definition 5.5.6 (Spectral density). Consider a (weakly) stationary time series $(X_n)_{n\in\mathbb{N}}$. If the associated autocovariance is in ℓ^1, one can define the spectral density as the discrete Fourier transform of the autocovariance function:

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{i\omega k}, \tag{5.63}$$

where $\gamma(k)$ is the autocovariance function at lag k.

Under the assumption that the spectral density exists, the time series is said to have **short memory** if $f(0)$ is finite. Otherwise the series is said to have **long memory**.

Definition 5.5.7 (Lag operator[7]). The lag operator sends a variable in a time series to the preceding value:

$$BX_t = X_{t-1}. \tag{5.64}$$

An important concept, especially in the context of autoregressive models, is that of a **lag polynomial** (the notation for these is not completely fixed in the literature, but the θ-notation is a common choice):

$$\theta(B) = 1 + \sum_{i=1}^{k} \theta_i B^i \tag{5.65}$$

$$\varphi(B) = 1 - \sum_{i=1}^{k} \varphi_i B^i. \tag{5.66}$$

Notation 5.5.8 (Difference operator). The difference operator Δ is defined as follows:

$$\Delta = 1 - B. \tag{5.67}$$

In a similar way one can define the **seasonal** difference operator:

$$\Delta_s = 1 - B^s. \tag{5.68}$$

Method 5.5.9 (Ljung-Box test). A test to see if a given set of autocorrelations of a time series is different from zero. Consider a time series of n elements and let $\{\rho_i\}_{1\leq i\leq k}$ be the first k lagged autocorrelation functions. The test statistic is defined as

$$Q = n(n+2) \sum_{i=1}^{k} \frac{\rho_k}{n-k}. \tag{5.69}$$

If the null hypothesis "there is no correlation" is true, the Q-statistic will asymptotically follow a χ^2-distribution with k degrees of freedom.

[7] Also called the **backshift operator**.

Method 5.5.10 (Augmented Dickey-Fuller test). Consider a time series $(X_t)_{t\in T}$. The (augmented) Dickey-Fuller test checks if the time series is (trend) stationary. For this test one considers the following regression model (similar to the ARIMA-*models* discussed in the next section):

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \sum_{i=1}^{p-1} \theta_i \Delta X_{t-i} + \varepsilon_t. \tag{5.70}$$

The test statistic is

$$\mathrm{DF} = \frac{\gamma}{\mathrm{SE}(\gamma)}, \tag{5.71}$$

where SE denotes the standard error. The null hypothesis states that $\gamma = 0$, i.e. there is a *unit root* $(1 - B)$ present in the model. Comparing the test statistic to tabulated critical values will give an indication whether to reject the hypothesis or not (the more negative the statistic, the more significant the result).

5.5.3 Autoregressive models

Definition 5.5.11 (AR(p)-model). Consider a time series $(X_t)_{t\in T}$. The autogressive model of order p is defined as the multiple linear regression model of X_t with respect to the first p lagged values $X_{t-1}, \dots, X_{t-p}$ of the time series:

$$X_t = \beta_0 + \beta_1 X_{t-1} + \cdots + \beta_p X_{t-p} + \varepsilon_t. \tag{5.72}$$

Definition 5.5.12 (Partial autocorrelation function). The p^{th} autocorrelation function is defined as the p^{th} coefficient in the AR(p)-model.

Remark 5.5.13. The optimal order p of an autoregressive model is the one for which all higher partial autocorrelation functions (almost) vanish.

Definition 5.5.14 (MA(p)-model). Consider a time series $(X_t)_{t\in T}$ where every X_t contains a white noise contribution $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. The moving average model of order p is defined as the multiple linear regression model of X_t with respect to the first p lagged values $\varepsilon_{t-1}, \dots, \varepsilon_{t-p}$ of the error term:

$$X_t = \beta_0 + \beta_1 \varepsilon_{t-1} + \cdots + \beta_p \varepsilon_{t-p} + \varepsilon_t. \tag{5.73}$$

Since the error terms are assumed to have mean zero, one can see that the intercept term β_0 gives the mean of the time series.

Remark 5.5.15. The optimal order p of an autoregressive model is the one for which all higher autocorrelation functions (almost) vanish.

Definition 5.5.16 (Invertibility). An MA(q)-model is said to be invertible if all roots of its associated lag polynomial $\theta(B)$ lie outside the unit circle. This condition implies that the polynomial is invertible, i.e. $1/\theta(B)$ can be written as a convergent series in the operator B. This in turn implies[8] that one can write the MA(q)-model as an AR(p)-model, where possibly $p = \infty$. The analogous property for AR(p)-models leads to a definition of **stationarity**.

In practice it is not always possible to describe a data set using either an autoregressive or a moving average model. However, these two types of models can be combined:

[8] Sometimes this is used as a definition of invertibility.

Definition 5.5.17 (ARMA(p, q)**-model).**

$$X_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} + \varepsilon_t \tag{5.74}$$

As above, one can find the optimal values for p and q by analyzing the autocorrelation and partial autocorrelation functions.

Using the lag polynomials one can rewrite the ARMA(p, q)-model as follows:

$$\varphi(B)X_t = \alpha_0 + \theta(B)\varepsilon_t. \tag{5.75}$$

By considering the special case where the polynomial $\mathcal{B}_\alpha^-$ has a unit root $1 - B$ with multiplicity d, one can obtain a generalization of the model:

$$\varphi(B)(1 - B)^d X_t = \alpha_0 + \theta(B)\varepsilon_t. \tag{5.76}$$

The interpretation of this additional factor $(1 - B)^d$ is related to the stationarity of the time series. The operator $1 - B$ is a finite difference operator:

$$\begin{aligned}
(1 - B)X_t &= X_t - X_{t-1} \\
(1 - B)^2 X_t &= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\
&\dots
\end{aligned}$$

By successive applications, one can obtain a stationary time series from a nonstationary time series. This combination of differencing, autoregression and moving averages is called the **ARIMA**-model[9].

Remark 5.5.18. Including so-called *exogenous* variables, i.e. external predictors, leads to an **ARIMAX**-model.

Remark 5.5.19 (Fitting AR- and MA-models). As is clear from the definition of an AR(p)-model, the parameters θ_i can easily be found using standard techniques for multivariate linear regression such as ordinary least squares. However, in contrast to AR-models where the predictors are known, the estimation of coefficients in MA-models is harder since the error terms ε_t are by definition unknown.

To estimate the coefficients in a MA-model, people have introduced multiple techniques (see for example [103]). One of the most famous ones is the method by *Durbin*:

Method 5.5.20 (Durbin). By restricting to invertible MA(q)-models (or by approximating a noninvertible model by an invertible one), one can first fit an AR(p)-model with $p > q$ to obtain estimates for the errors ε_t and then, in a second step, use a least squares-method to solve for the coefficients in the MA-model.

As a last modification one can introduce seasonal components. Simple trends such as a linear growth are easily removed from the time series by detrending or differencing. However, a periodic pattern is harder to remove and, in general, ARIMA-models are not suited to accompany this type of features. Luckily one can easily modify the ARIMA-model to incorporate seasonal variations. The multiplicative SARIMA-model is obtained by inserting operators similar to the ones of the ordinary ARIMA-model, where the lag operator B is replaced by the seasonal lag operator B^s (where s is the period of the seasonal variation):

Definition 5.5.21 (ARIMA$(p, q, d)(P, Q, D)_s$**-model).**

$$\Phi(B^s)\varphi(B)\Delta_s^D \Delta^d X_t = \theta(B)\Theta(B^s)\varepsilon_t \tag{5.77}$$

[9] The 'I' stands for "integrated".

5.5.4 Causality

Definition 5.5.22 (Granger causality). Consider two time series $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$. The time series X_n is said to Granger-cause Y_n if past values of X_n help to predict future values of Y_n. More formally this can be stated as follows:

$$P[Y_{t+k} \in A \mid \Omega(t)] \neq P[Y_{t+k} \in A \mid \Omega\backslash X(t)] \tag{5.78}$$

for some k, where $\Omega(t)$ and $\Omega\backslash X(t)$ denote the available information at time t with and without removing the variable X from the universe.

This formulation of causality was introduced by *Granger* under the following two assumptions:

- The cause always happens prior to the effect.
- The cause carries unique information about the effect.

Remark 5.5.23. A slightly different but for computational purposes often more useful[10] notion of Granger-causality is as follows. A time series $(X_n)_{n\in\mathbb{N}}$ is said to **Granger-cause** a time series $(Y_n)_{n\in\mathbb{N}}$ if the variance of predictions of Y_n becomes smaller when the information contained in X_n is taken into account.

Remark 5.5.24. Assume that two uncorrelated models giving predictions of a time series are given. One way to check if they have the same accuracy is the *Diebold-Mariano test*. However, when testing for Granger-causality one should pay attention. This test is not valid for nested models and, hence, is not applicable to two models that only differ by a set of extra predictors (in this case an external time series).

5.6 Uncertainty modelling

5.6.1 Prediction regions

One of the simplest ways to express uncertainty about predictions or parameter estimates is to give a set of possible values instead of a single value. However, to be meaningful, these sets should satisfy some conditions.

Definition 5.6.1 (Validity). Consider a region predictor Γ and let P be the joint distribution on the instance space $Z \equiv X \times Y$. Γ is said to be valid (at **significance/confidence level** $\alpha \in [0,1]$) if satisfies

$$P(y \in \Gamma^\alpha(x)) \geq 1 - \alpha. \tag{5.79}$$

One sometimes also distinguishes between exact validity and **conservative** validity, where the former is a subcase of the latter for which the inequality becomes an equality.

In practice one will always prefer models that are as exact as possible.

Remark 5.6.2 (Confidence regions). The definition of valid confidence predictors above is similar to the definition of confidence regions (Section 4.6). However, in contrast to confidence regions of population parameters, the size of confidence regions for predictive distributions does not go towards to zero in the infinite data limit. This follows from the fact that in general all observations are subject to noise and, hence, even with perfect knowledge about the data generating distribution, an exact prediction is impossible.[11]

[10] In fact this was the original definition by *Granger*.

[11] Note that when observations are not sampled according to a distribution, but are perfectly predictable, this is of course not true.

5.6.2 Conformal prediction

A very general framework for the construction of valid prediction intervals in a model-independent manner is given by the conformal prediction framework by *Vovk et al.* The main ingredients for the construction are randomization and *conformity measures.*

The first step will be studying the behaviour under randomization of the existing data (be it measurements or past predictions). To ensure that the procedure satisfies the required confidence (or probability) bounds, one has to make some assumptions. One of the main benefits of this framework is that one can relax the condition of the data being i.i.d. to it being exchangeable:

Definition 5.6.3 (Exchangeability). Consider an ordered data sample $\{z_i\}_{1\leq i\leq N}$. The joint distribution $P(z_1,\dots,z_N)$ is said to be exchangeable if it is invariant under any permutation of the data points. A distribution Q is said to be exchangeable if Q^n is exchangeable for all $n\in\mathbb{N}$.

This definition can be restated in a purely combinatorial way. First, define the notion of a **bag** obtained from a (possibly ordered) data sample $\{z_i\}_{1\leq i\leq N}$ as the (unordered) set $\mathcal{B}$ containing these elements. The joint distribution P is then said to be exchangeable if the probability of finding any sequence of data points is equal to the probability of drawing this same sequence from the bag of these elements. Since this probability is purely combinatorial and, hence, completely independent of the ordering, it should be clear that this coincides with the first definition. The set of bags in a space X is sometimes denoted by X^∞.

Definition 5.6.4 (Nonconformity measure). Consider a bag $\mathcal{B}$ together with a new element z^* in a probability space (Z,Σ,P). A noncomformity measure $f:Z^\infty\times Z\rightarrow\mathbb{R}$ is a measurable function that gives a number indicating how different z^* is from the content of $\mathcal{B}$.

Remark. One could restate all statements in this section in terms of "conformity measures" and, hence, look at similarities instead of dissimilarities. It will become clear that the procedure is invariant under monotone transformations and, hence, everything can be multiplied by -1.

Example 5.6.5 (Point predictors). A general class of nonconformity measures is obtained from point predictors for a metric space (Y,d) . Given a point predictor $\rho:X\rightarrow Y$ trained on a bag $\mathcal{B}$, one can define a nonconformity measure as follows:

$$A_\rho(\mathcal{B},(x,y)) := d(\rho(x),y). \tag{5.80}$$

Example 5.6.6 (Interval predictors). For every model $C\equiv(l,u):X\rightarrow\mathbb{R}^2$ that predicts intervals, i.e. $\forall x\in X: l(x)\leq u(x)$, that is trained on a bag $\mathcal{B}$, one can define a nonconformity measure as follows:

$$A_C(\mathcal{B},(x,y)) := \max\Big(l(x)-y, y-u(x)\Big). \tag{5.81}$$

Example 5.6.7 (Nested interval predictors). Let T be a totally ordered set **??**. For every model $(C_t)_{t\in T}\equiv(l_t,u_t)_{t\in T}:X\rightarrow\mathbb{R}^2$ that predicts a sequence of nested intervals, i.e. $s\leq t\implies l_s(x)\geq l_t(x)\wedge u_s(x)\leq u_t(x)$, that is trained on a bag $\mathcal{B}$, one can define a nonconformity measure as follows:

$$A_T(\mathcal{B},(x,y)) := \inf\{t\in T\mid y\in C_t(x)\}. \tag{5.82}$$

Construction 5.6.8 (Conformal predictor). Consider a data sample given as a bag $\mathcal{B}\in P(Z)$ together with a nonconformity measure A and let α denote the confidence level of the prediction region to be constructed. For any new element $z^*\in Z$ the algorithm proceeds as follows:

1. Denote the nonconformity score $A(\mathcal{B},z^*)$ by μ_{z^*}.

2. For every element z in $\mathcal{B}$, define μ_z by replacing z by z^* in the bag and calculating the nonconformity score as in the previous step.
3. Calculate the conformal p-value as the fraction of elements $z \in \mathcal{B} \cup \{z^*\}$ for which $\mu_z \geq \mu_{z^*}$:
$$p^* := \frac{|\{z \in \mathcal{B} \cup \{z^*\} \mid \mu_z \geq \mu_{z^*}\}|}{|\mathcal{B}| + 1}. \tag{5.83}$$
4. Include an element z^* in the prediction region C^α if and only if $p^* > \alpha$.

It should be noted that, in general, the construction of these regions can be quite time-consuming because if the nonconformity measure A depends on model that has to be trained on $\mathcal{B}$, this training has to be reperformed for every element $z \in \mathcal{B}$ in step 2. For low-dimensional regions it can often be achieved by solving inequalities derived from the specific form of the nonconformity measure.

Property 5.6.9 (Optimality). A conformal predictor satisfies the following conditions:

- Regions are (conservatively) valid, i.e. $P(p^* \leq \alpha) \leq \alpha$ and thus also $P(y^* \in \Gamma^\alpha(x)) \geq 1 - \alpha$.
- Regions are nested, i.e. $\alpha \leq \beta \implies \forall x \in X : \Gamma^\alpha(x) \subseteq \Gamma^\beta(x)$.

Given any region predictor satisfying these three properties below, there exists a conformal predictor that is more efficient, i.e. produces smaller prediction regions.

Property 5.6.10 (Smooth conformal predictors). One can modify the above construction in such a way that the resulting conformal predictors are not only conservatively valid but are also exactly valid. To this end one replaces the conformal p-value (5.83) by
$$p^*(\theta) := \frac{|\{z \in \mathcal{B} \cup \{z^*\} \mid \mu_z > \mu_{z^*}\}| + \theta\,|\{z \in \mathcal{B} \cup \{z^*\} \mid \mu_z = \mu_{z^*}\}|}{|\mathcal{B}| + 1}, \tag{5.84}$$
where θ is uniformly sampled from the unit interval $[0, 1]$. Exact validity is then obtained by also marginalizing over the random variable θ.

Now, one could wonder if the assumption of exchangeability is a realistic assumption. Obviously if one applies the procedure to independent observations, everything is fine since i.i.d. sequences are clearly exchangeable. However, some important classes of data sequences are clearly not exchangeable, e.g. time series. This kind of data often contains intrinsic correlation and, hence, the exchangeability assumption is almost always violated. However, a solution exists. One can restate the construction above using an explicit randomization as is done in [79]. There, one replaces the nonconformity measure by a function that acts on ordinary sequences instead of unordered bags. The fraction p^* can then be expressed as follows:
$$p^* = \frac{1}{|S_{N+1}|} \sum_{\sigma \in S_{N+1}} \mathbb{1}(A(\sigma \cdot \vec{z}) \geq A(\vec{z})), \tag{5.85}$$
where $\vec{z} \equiv (z_1, \dots, z_N, z^*)$. Using these explicit permutations, one can generalize the construction of conformal predictors to arbitrary randomization schemes, i.e. to subgroups of S_{N+1}. However, in general this will ruin the validity of the procedure.

?? FINISH ??

At last a computationally efficient modification of the original CP algorithm is introduced. For most applications, especially those in machine learning and big data, the computational inefficiency of conformal predictors would make them hard to use. To overcome this issue *Papadopoulos et al.* introduced the following modification:

Construction 5.6.11 (Inductive CP). Consider a data set $\mathcal{D} \subset Z$ and a nonconformity measure A based on an underlying predictor. First, split $\mathcal{D}$ into a training set $\mathcal{T}$ and a calibration set $\mathcal{C}$. Using $\mathcal{T}$, train the underlying predictor of A. Then, for every point $z \equiv (x, y) \in \mathcal{C}$, construct the nonconformity score $\mu_z := A(z)$. As before, for every new element $z^* \in Z$ the conformal p-value is defined as the fraction of elements in $\mathcal{C}$ for which the nonconformity measure is larger than the one for z^*:

$$p^* := \frac{|\{z \in \mathcal{C} \cap \{z^*\} \mid \mu_z \geq \mu_{z^*}\}|}{|\mathcal{C}| + 1}. \tag{5.86}$$

As in the original CP algorithm, a new observation z^* is included in the prediction region if and only if $p^* > \alpha$:

$$\Gamma^\alpha(x) := \{y \in Y \mid p(x, y) > \alpha\}. \tag{5.87}$$

It should be clear that the underlying predictor only needs to be trained once using this scheme.

Remark 5.6.12 (Terminology). The name "inductive CP" stems from the fact that the general behaviour is deduced from a small subset of all observations. For this reason one sometimes calls the original algorithm a "transductive" method.

Property 5.6.13 (Validity). Although the ICP algorithm is computationally much more efficient than its transductive counterpart, one does not have to give up any of its properties. If the data is exchangeable (and the resulting nonconformity score are too) it is not hard to see that the rank of any new nonconformity score among the calibration scores $A(\mathcal{C}) := \{\mu_z \in \mathbb{R} \mid z \in \mathcal{C}\}$ is uniformly distributed in $\{1, \ldots, |\mathcal{C}| + 1\}$. So, given a quantile level α, the probability of finding a new nonconformity score smaller or equal than the α-quantile of $A(\mathcal{C})$ is

$$P\Big(\mu_{z^*} \leq q_\alpha(A(\mathcal{C}))\Big) = \frac{\lceil \alpha |\mathcal{C}| \rceil}{|\mathcal{C}| + 1}. \tag{5.88}$$

By comparison with Equation (5.86) one immediately gets the following result:

$$P(y^* \in \Gamma^\alpha(x, \mathcal{C})) = \frac{\lceil (1 - \alpha)|\mathcal{C}| \rceil}{|\mathcal{C}| + 1}. \tag{5.89}$$

However, this is where the TCP and ICP algorithms differ. When using Equation (5.83) for the transductive algorithm, the above probabilities would contain a factor $|\mathcal{C}| + 1$ in both the numerator and the denominator, so the coverage condition is satisfied. However, for the inductive algorithm, a minimal modification also leads to validity:

$$\Gamma^\alpha(x) \longrightarrow \Gamma^\alpha(x) := \big\{y \in Y \mid A(x, y) \leq q_{(1+1/|\mathcal{C}|)(1-\alpha)}(A(\mathcal{C}))\big\}, \tag{5.90}$$

i.e. the empirical quantiles are replaced by "inflated" quantiles. The smoothing of Property 5.6.10 can also be implied in this case. If one does not use a smoothened conformal predictor but assumes that all calibration scores are distinct, the following property is obtained:

$$P(y^* \in \Gamma^\alpha(x, \mathcal{C})) \geq 1 - \alpha + \frac{1}{|\mathcal{C}|}. \tag{5.91}$$

So, in the limit of large calibration sets, the exact validity is recovered.

Remark 5.6.14. One does have to pay attention when interpreting the above statement. The validity property holds in probability with respect to both the new instance (x^*, y^*) and the calibration set $\mathcal{C}$. However, this does not not mean that for a fixed calibration set $\mathcal{C}$ the error fraction

$$\frac{|\{1 \leq i \leq k : y_i \in \Gamma^\alpha(x_i, \mathcal{C})\}|}{k} \tag{5.92}$$

is bounded by α for $k \longrightarrow \infty$. Because the events are not independent, the error can be much larger than α.

The above offline ICP algorithm can be generalized to an online algorithm:

Construction 5.6.15 (Online ICP). Consider an increasing sequence of positive integers $(m_n)_{n\in\mathbb{N}_0}$ of "update thresholds". The prediction region $\Gamma^\alpha(\mathcal{S})$ for the data sample $\mathcal{S} := \{(x_1, y_1), \dots, (x_n, y_n)\}$ is defined as follows:

- If $n \leq m_1$, use a fixed conformal predictor to construct $\Gamma^\alpha(\mathcal{S})$.
- If $m_k < n \leq m_{k+1}$, construct $\Gamma^\alpha(\mathcal{S})$ as follows:

$$\Gamma^\alpha(\mathcal{S}) := \left\{ y \in Y \;\middle|\; \frac{|\{m_k < j \leq n \mid \mu_{z_j} \geq \mu_{z_n}\}|}{n - m_k} > \varepsilon \right\}, \tag{5.93}$$

where

$$\begin{aligned} \mu_{z_j} &:= A(\mathcal{B}(z_1, \dots, z_{m_k}), (x_j, y_j)) \\ \mu_{z_n} &:= A(\mathcal{B}(z_1, \dots, z_{m_k}), (x_n, y)). \end{aligned}$$

It is clear that for $k \ll |\mathcal{C}|$ the offline ICP algorithm approximates the online version. The major benefit of the online algorithm is that one does not need to use the inflated quantiles to obtain valid prediction regions, because the calibration set grows every time new data is observed and, hence, the finite-size fluctuations get supressed.

?? COMPLETE ??

5.6.3 Distribution-shift

An important problem in the field of data science, in particular that of machine learning, is the change of the data generating process. Consider a classic train-test routine. If the model was trained on a data set sampled from the distribution P_0, but applied to a data set sampled from the distribution P_1, there is no reason the expect that the model will still give reasonable results. Furthermore, even if the crude point predictions are still somewhat sensible, this does not mean that their theoretical properties, such as the validity of confidence regions, is preserved.

A first step to resolve this problem is the detection of a possible distribution shift. Conformal prediction, as introduced in the previous section, can be used to detect such shifts in an online fashion. The main idea of the algorithm is to construct a martingale 3.6.6 from the p-values produced by a conformal predictor. For every (reasonable) martingale $(X_n)_{n\in\mathbb{N}}$ the Doob-Ville inequality 3.6.7 says that the growth is bounded. However, if the sequence $(X_n)_{n\in\mathbb{N}}$ is constructed in such a way that the martingale property is lost once the data distribution changes, the Doob-Ville inequality can be violated and the crossing of a given threshold can be regarded as evidence for this distribution shift [8].

The general expression of the martingale will be of the form

$$X_i \equiv \prod_{k=0}^{i} f_k(p_k), \tag{5.94}$$

where p_k is the p-value of the k^{th} data point as produced by some conformal predictor. The functions f_k are called the **betting functions**. For $(X_n)_{n\in\mathbb{N}}$ to be a martingale with respect to the natural filtration of the p_k's, the betting functions should satisfy

$$\int_0^1 f_k(p)dp = 1. \tag{5.95}$$

Method 5.6.16 (Power martingale). For every constant $\varepsilon \in [0,1]$ one defines the power martingale by

$$X_i^\varepsilon := \prod_{k=0}^{i} \varepsilon p_k^{\varepsilon-1}. \tag{5.96}$$

One can also construct a mixture martingale by integrating over ε:

$$X_i := \int_0^1 X_i^\varepsilon d\varepsilon. \tag{5.97}$$

Because $\varepsilon \leq 0$, the above martingales will start to become very large if the conformal predictor produces small p-values, i.e. when unlikely values are observed.

However, not all distribution shifts will give rise to such behaviour. For example, it is possible that, although the p-values are not distributed uniformly anymore (since the data is not exchangeable), they are concentrated in the upper half of the unit interval and, hence, do not let the ‘‘martingale'' grow strongly. For this reason it is convenient to construct betting functions that take into account the distribution of the p-values.

Method 5.6.17 (Plug-in martingale). Let $\hat{\rho}_i(p)$ denote an estimate of the probability density constructed using the p-values $\{p_1, \dots, p_i\}$. The plug-in martingale is defined by the betting functions

$$f_i := \hat{\rho}_{i-1}. \tag{5.98}$$

Now, if the empirical distribution functions of the p-values converge weakly 2.1.49 to an absolutely continuous distribution and $\log(\hat{\rho}_i(p)) \longrightarrow \log(\rho(p))$ uniformly, where ρ is the limit density of the empirical distributions, it can be shown that plug-in martingale grows quicker than the martingale associated to any other (continuous) betting function.

List of Symbols

The following symbols are used throughout the summary:

Abbreviations

AIC	Akaike information criterion
ARMA	autoregressive moving-average model
BCH	Baker-Campbell-Hausdorff
CCR	canonical commutation relation
CDF	cumulative distribution function
CFT	conformal field theory
CIS	completely integrable system
CP	completely positive
CPTP	completely positive trace-preserving
CR	Cauchy-Riemann
DGA	differential graded algebra
DGCA	differential graded-commutative algebra
EPR	Einstein-Podolsky-Rosen
ETCS	Elementary Theory of the Category of Sets
FWHM	full width at half maximum
GHZ	Greenberger-Horne-Zeilinger
GNS	Gel'fand-Naimark-Segal
HoTT	Homotopy Type Theory
KKT	Karush-Kuhn-Tucker
LIVF	left-invariant vector field
MPS	matrix product states
MTC	modular tensor category
NDR	neighbourhood deformation retract
OPE	operator product expansion
OZI	Okubo-Zweig-Iizuka
PVM	projection-valued measure
RKHS	reproducing kernel Hilbert space
SVM	support-vector machine
TQFT	topological quantum field theory
ZFC	Zermelo-Frenkel set theory with the axiom of choice

TVS	topological vector space
Operations	
Ad_g	adjoint representation of a Lie group G
ad_X	adjoint representation of a Lie algebra $\mathfrak{g}$
arg	argument of a complex number
$\Box$	d'Alembert operator
$\deg(f)$	degree of the polynomial f
e	identity element of a group
$\Gamma(E)$	set of global sections of a fibre bundle E
Im	imaginary part of a complex number
$\mathrm{Ind}_f(z)$	index of a point $z \in \mathbb{C}$ with respect to a function f
$\hookrightarrow$	injective function
$\cong$	is isomorphic to
Par_t^γ	parallel transport map with respect to the curve γ
Re	real part of a complex number
Res	residue of a complex function
$\twoheadrightarrow$	surjective function
$\{\cdot,\cdot\}$	Poisson bracket
∂X	boundary of a topological space X
$\overline{X}$	closure of a topological space X
$X^\circ, \mathring{X}$	interior of a topological space X
$\sphericalangle(\cdot,\cdot)$	angle between two vectors
$X \times Y$	cartesian product of the sets X and Y
$X + Y$	sum of the vector spaces X and Y
$X \oplus Y$	direct sum of the vector spaces X and Y
$V \otimes W$	tensor product of the vector spaces V and W
$\mathbb{1}_X$	identity morphism on the object X
$\approx$	is approximately equal to
$\hookrightarrow$	is included in
$\cong$	is isomorphic to
$\mapsto$	mapsto
Collections	
Ab	category of Abelian groups
$\mathrm{Aut}(X)$	automorphism group of an object X
$\mathcal{B}_0(V,W)$	space of compact bounded operators between the Banach spaces V and W
$\mathcal{B}(V,W)$	space of bounded linear maps from the space V to the space W
CartSp	the category of Euclidean spaces and "suitable" homomorphisms (e.g. linear maps, smooth maps, ...)
$C_\bullet$	chain complex
$\mathbf{Ch}(\mathbf{A})$	category of chain complexes with objects in the additive category $\mathbf{A}$

$\mathbf{C}^\infty$	category of smooth spaces
$C_p^\infty(M)$	ring of smooth functions $f : M \to \mathbb{R}$ on a neighbourhood of $p \in M$
$C^\omega(V)$	the set of all analytic functions defined on the set V
$\mathrm{Conf}(M)$	conformal group of (pseudo-)Riemannian manifold M
$C(X, Y)$	set of continuous functions between two topological spaces X and Y
$\mathbf{C^\infty Ring}, \mathbf{C^\infty Alg}$	category of smooth algebras
Diff	category of smooth manifolds
DiffSp	category of diffeological spaces and smooth maps
D^n	standard n-disk
$\mathrm{dom}(f)$	domain of a function f
$\mathrm{End}(X)$	endomorphism monoid of a an object X
$\mathcal{E}\mathrm{nd}$	endomorphism operad
$\mathbf{FormalCartSp_{diff}}$	category of infinitesimally thickened Euclidean spaces
$\mathrm{GL}(V)$	general linear group, the group of automorphisms of a vector space V
$\mathrm{GL}(n, K)$	general linear group: the group of all invertible $n \times n$-matrices over the field K
Grp	category of groups and group homomorphisms
Grpd	category of groupoids
$\mathrm{Hol}_p(\omega)$	holonomy group at the point p with respect to the principal connection ω
$\mathrm{Hom}_{\mathbf{C}}(V, W)$	set of homomorphisms from an object V to an object W in a category $\mathbf{C}$
hTop	homotopy category
$\mathrm{im}(f)$	image of a function f
$K^0(X)$	K-theory over a (compact Hausdorff) space X
Kan	category of Kan complexes
$\mathcal{K}_n(A, v)$	Krylov subspace of dimension n generated by the matrix A and the vector v
L^1	space of integrable functions
Law	category of Lawvere theories
Lie	category of Lie groups
$\mathfrak{Lie}$	category of Lie algebras
$\mathfrak{X}^L$	space of left-invariant vector fields on a Lie group
LX	free loop space on X
$\mathbf{Man}^p$	category of C^p-manifolds
Meas	category of measure spaces and measure-preserving functions
$N\mathbf{C}$	the simplicial nerve of a small category $\mathbf{C}$
$\mathbf{Open}(X)$	category of open subsets of a topological space X
$\mathrm{O}(n, K)$	group of $n \times n$ orthogonal matrices over a field K
$P(S), 2^S$	power set of S
$\mathrm{Pin}(V)$	pin group of the Clifford algebra $C\ell(V, Q)$
$\mathbf{Psh}(\mathbf{C}), \widehat{\mathbf{C}}$	category of presheaves on a (small) category $\mathbf{C}$
$\mathbf{Sh}(X)$	category of sheaves on a topological space X
$\mathbf{Sh}(\mathbf{C}, J)$	category of J-sheaves on a site $(\mathbf{C}, J)$

$\mathbf{\Delta}$	The simplex category.
$\mathrm{SL}_n(K)$	special linear group: group of all invertible n-dimensional matrices with unit determinant over the field K
S^n	standard n-sphere
$S^n(V)$	space of symmetric rank n tensors over a vector space V
$W^{m,p}(U)$	the Sobolov space in L^p of order m
$\mathbf{Span}(\mathbf{C})$	span category over $\mathbf{C}$
$\mathrm{Spec}(R)$	spectrum of a commutative ring R
$\mathrm{supp}(f)$	support of a function f
S_n	symmetric group of degree n
$\mathrm{Sym}(X)$	symmetric group on the set X
$\mathrm{Sp}(n, K)$	group of matrices preserving a canonical symplectic form over the field K
$\mathrm{Sp}(n)$	compact symplectic group
$\mathrm{TL}_n(\delta)$	Temperley-Lieb algebra with $n-1$ generators and parameter δ.
T^n	standard n-torus (the n-fold Cartesian product of S^1)
Top	category of topological spaces
Topos	the 2-category of (elementary) topoi and geometric morphisms
$U(\mathfrak{g})$	universal enveloping algebra of a Lie algebra $\mathfrak{g}$
$\mathrm{U}(n, K)$	group of $n \times n$ unitary matrices over a field K
$\mathbf{Vect}(X)$	category of vector bundles over a manifold X
$\mathbf{Vect}_K$	category of vector spaces and linear maps over a field K
Y^X	set of functions from a set X to a set Y
$\emptyset$	empty set
$\pi_n(X, x_0)$	n^{th} homotopy space over X with basepoint x_0
$[a, b]$	closed interval
$]a, b[$	open interval
$\Lambda^n(V)$	space of antisymmetric rank n tensors over a vector space V
ΩX	(based) loop space on X
$\Omega^k(M)$	$C^\infty(M)$-module of differential k-forms on the manifold M
$\rho(A)$	resolvent set of a bounded linear operator A
$\mathfrak{X}(M)$	$C^\infty(M)$-module of vector fields on the manifold M

Bibliography

[1] Thomas Augustin, Frank PA Coolen, Gert De Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities.* John Wiley & Sons, 2014.

[2] Tetsuji Miwa, Michio Jimbo, Michio Jimbo, and E Date. *Solitons: Differential Equations, Symmetries and Infinite-dimensional Algebras*, volume 135. Cambridge University Press, 2000.

[3] Vladimir I. Arnol'd. *Mathematical Methods of Classical Mechanics*, volume 60. Springer Science & Business Media, 2013.

[4] Edwin J. Beggs and Shahn Majid. *Quantum Riemannian Geometry.* Springer, 2020.

[5] Marc Henneaux and Claudio Teitelboim. *Quantization of Gauge Systems.* Princeton university press, 1992.

[6] Mark Hovey. *Model Categories.* Number 63. American Mathematical Soc., 2007.

[7] Gregory M. Kelly. *Basic Concepts of Enriched Category Theory*, volume 64. CUP Archive, 1982.

[8] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World.* Springer Science & Business Media, 2005.

[9] Mukund Rangamani and Tadashi Takayanagi. *Holographic Entanglement Entropy.* Springer, 2017.

[10] Saunders Mac Lane. *Categories for the Working Mathematician*, volume 5. Springer Science & Business Media, 2013.

[11] Peter T. Johnstone. *Topos Theory.* Dover Publications, 2014.

[12] Shun-ichi Amari. *Information Geometry and Its Applications.* Springer Publishing Company, Incorporated, 2016.

[13] Charles W. Misner, Kip S. Thorne, and John A. Wheeler. *Gravitation.* Princeton University Press, 2017.

[14] Carlo Rovelli and Francesca Vidotto. *Covariant Loop Quantum Gravity: An Elementary Introduction to Quantum Gravity and Spinfoam Theory.* Cambridge University Press, 2014.

[15] Richard W. Sharpe. *Differential Geometry: Cartan's Generalization of Klein's Erlangen Program*, volume 166. Springer Science & Business Media, 2000.

[16] John C. Baez, Irving E. Segal, and Zhengfang Zhou. *Introduction to Algebraic and Constructive Quantum Field Theory.* Princeton University Press, 2014.

[17] Raoul Bott and Loring W. Tu. *Differential Forms in Algebraic Topology.* Graduate Texts in Mathematics. Springer New York, 1995.

[18] The Univalent Foundations Program. *Homotopy Type Theory: Univalent Foundations of Mathematics.* `https://homotopytypetheory.org/book`, Institute for Advanced Study, 2013.

[19] Bruce Blackadar. *Operator Algebras: Theory of C∗-Algebras and von Neumann Algebras.* Springer, 2013.

[20] Marek Capinski and Peter E. Kopp. *Measure, Integral and Probability.* Springer Science & Business Media, 2013.

[21] Georgiev Svetlin. *Theory of Distributions.* Springer, 2015.

[22] Gerd Rudolph and Matthias Schmidt. *Differential Geometry and Mathematical Physics: Part II. Fibre Bundles, Topology and Gauge Fields.* Springer, 2017.

[23] Martin Schottenloher. *A Mathematical Introduction to Conformal Field Theory*, volume 759. 2008.

[24] Dusa McDuff and Deitmar Salamon. *Introduction to Symplectic Topology.* Oxford Graduate Texts in Mathematics. Oxford University Press, 2017.

[25] John C. Baez and Peter May. *Towards Higher Categories*, volume 152 of *IMA Volumes in Mathematics and its Applications.* Springer, 2009.

[26] Mikhail. M. Kapranov and Vladimir A. Voevodsky. *2-categories and Zamolodchikov Tetrahedra Equations*, volume 56 of *Proc. Sympos. Pure Math.* Amer. Math. Soc., Providence, RI, 1994.

[27] Geoffrey Compère. *Advanced Lectures on General Relativity*, volume 952. Springer, 2019.

[28] Pavel Etingof, Shlomo Gelaki, Dmitri Nikshych, and Victor Ostrik. *Tensor Categories*, volume 205. American Mathematical Soc., 2016.

[29] David Mumford. *The Red Book of Varieties and Schemes: Includes the Michigan Lectures (1974) on Curves and Their Jacobians*, volume 1358. Springer Science & Business Media, 1999.

[30] Charles A. Weibel. *An Introduction to Homological Algebra.* Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1994.

[31] Peter J. Hilton and Urs Stammbach. *A Course in Homological Algebra.* Springer.

[32] Jean-Luc Brylinski. *Loop Spaces, Characteristic Classes and Geometric Quantization.* Birkhauser.

[33] Antoine Van Proeyen and Daniel Freedman. *Supergravity.* Cambridge University Press.

[34] William S. Massey. *A Basic Course in Algebraic Topology.* Springer.

[35] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to Quantum Field Theory.* Westview Press.

[36] Nadir Jeevanjee. *An Introduction to Tensors and Group Theory for Physicists.* Birkhauser.

[37] Yvonne Choquet-Bruhat, Cecile DeWitt-Morette, and Margaret Dillard-Bleick. *Analysis, Manifolds and Physics, Part 1: Basics.* North-Holland.

[38] Yvonne Choquet-Bruhat and Cecile DeWitt-Morette. *Analysis, Manifolds and Physics, Part 2.* North-Holland.

[39] Herbet Goldstein, John L. Safko, and Charles P. Poole. *Classical Mechanics*. Pearson.

[40] Franco Cardin. *Elementary Symplectic Topology and Mechanics*. Springer.

[41] Walter Greiner and Joachim Reinhardt. *Field Quantization*. Springer.

[42] Walter Greiner. *Quantum Mechanics*. Springer.

[43] B. H. Bransden and Charles J. Joachain. *Quantum Mechanics*. Prentice Hall.

[44] Heydar Radjavi and Peter Rosenthal. *Invariant Subspaces*. Dover Publications.

[45] Max Karoubi. *K-Theory: An Introduction*. Springer.

[46] Damien Calaque and Thomas Strobl. *Mathematical Aspects of Quantum Field Theories*. Springer, 2015.

[47] Ivan Kolar, Peter W. Michor, and Jan Slovak. *Normal Operations in Differential Geometry*. Springer.

[48] Stephen B. Sontz. *Principal Bundles: The Classical Case*. Springer.

[49] Stephen B. Sontz. *Principal Bundles: The Quantum Case*. Springer.

[50] William Fulton and Joe Harris. *Representation Theory: A First Course*. Springer.

[51] Peter Petersen. *Riemannian Geometry*. Springer.

[52] Charles Nash and Siddharta Sen. *Topology and Geometry for Physicists*. Dover Publications.

[53] Ian M. Anderson. *The Variational Bicomplex*.

[54] Edward Witten. Global anomalies in string theory. In *Symposium on Anomalies, Geometry, Topology*, 6 1985.

[55] F.A. Berezin and M.S. Marinov.

[56] Paul A. M. Dirac. Generalized Hamiltonian dynamics. *Canadian Journal of Mathematics*, 2:129–148, 1950.

[57] Angelo Vistoli. Notes on Grothendieck topologies, fibered categories and descent theory. *arXiv:math/0412512*, 2004.

[58] Emily Riehl and Dominic Verity. The theory and practice of Reedy categories. *Theory and Applications of Categories*, 29, 2013.

[59] Emily Riehl. Homotopical categories: From model categories to $(\infty, 1)$-categories. 2019. arXiv:1904.00886.

[60] Floris Takens. A global version of the inverse problem of the calculus of variations. *Journal of Differential Geometry*, 14(4):543--562, 1979.

[61] John Baez and Alexander Hoffnung. Convenient categories of smooth spaces. *Transactions of the American Mathematical Society*, 363(11):5789--5825, 2011.

[62] John C. Baez and Alissa S. Crans. Higher-dimensional algebra vi: Lie 2-algebras. 2003. arXiv:math/0307263.

[63] John C. Baez and Aaron D. Lauda. Higher-dimensional algebra v: 2-groups. 2003. arXiv:math/0307200.

[64] Edward Witten. Supersymmetry and Morse theory. *J. Diff. Geom*, 17(4):661--692, 1982.

[65] Urs Schreiber. *From Loop Space Mechanics to Nonabelian Strings*. PhD thesis, 2005.

[66] John C Baez and Urs Schreiber. Higher gauge theory. 2005. arXiv:math/0511710.

[67] Jade Master. Why is homology so powerful? 2020. arXiv:2001.00314.

[68] Marcus Berg, Cécile DeWitt-Morette, Shangjr Gwo, and Eric Kramer. The Pin groups in physics: C, P and T. *Reviews in Mathematical Physics*, 13(08):953--1034, 2001.

[69] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(Oct):1705--1749, 2005.

[70] Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. Bregman Voronoi diagrams. *Discrete & Computational Geometry*, 44(2):281--307, 2010.

[71] Richard Palais. The symmetries of solitons. *Bulletin of the American Mathematical Society*, 34(4):339--403, 1997.

[72] Michael F. Atiyah. Topological quantum field theory. *Publications Mathématiques de l'IHÉS*, 68:175--186, 1988.

[73] Jens Eisert, Christoph Simon, and Martin B Plenio. On the quantification of entanglement in infinite-dimensional quantum systems. *Journal of Physics A: Mathematical and General*, 35(17):3911--3923, 2002.

[74] Benoît Tuybens. Entanglement entropy of gauge theories. 2017.

[75] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338--353, 1965.

[76] John Baez, Alexander Hoffnung, and Christopher Rogers. Categorified symplectic geometry and the classical string. *Communications in Mathematical Physics*, 293:701--725, 2010.

[77] Charles Rezk. A model for the homotopy theory of homotopy theory. *Transactions of the American Mathematical Society*, 353(3):973--1007, 2001.

[78] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371-–421, 2008.

[79] Victor Chernozhukov, Kaspar Wüthrich, and Zhu Yinchu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 732--749. PMLR, 2018.

[80] Peter May. A note on the splitting principle. *Topology and Its Applications*, 153(4):605--609, 2005.

[81] Irina Markina. Group of diffeomorphisms of the unit circle as a principal $U(1)$-bundle.

[82] Sjoerd E. Crans. Localizations of transfors. 1998.

[83] Tom Leinster. Basic bicategories. 1998. arXiv:math/9810017.

[84] Alexander E. Hoffnung. Spans in 2-categories: A monoidal tricategory. 2011. arXiv:1112.0560.

[85] Eugenia Cheng and Nick Gurski. The periodic table of n-categories for low dimensions ii: Degenerate tricategories. 2007. arXiv:0706.2307.

[86] Mehmet B. Şahinoğlu, Dominic Williamson, Nick Bultinck, Michael Mariën, Jutho Haegeman, Norbert Schuch, and Frank Verstraete. Characterizing topological order with matrix product operators. 2014. arXiv:1409.2150.

[87] Dominic J. Williamson, Nick Bultinck, Michael Mariën, Mehmet B. Şahinoğlu, Jutho Haegeman, and Frank Verstraete. Matrix product operators for symmetry-protected topological phases: Gauging and edge theories. *Phys. Rev. B*, 94, 2016.

[88] Guifré Vidal. Efficient classical simulation of slightly entangled quantum computations. *Phys. Rev. Lett.*, 91, 2003.

[89] Aaron D. Lauda and Hendryk Pfeiffer. Open–closed strings: Two-dimensional extended TQFTs and Frobenius algebras. *Topology and its Applications*, 155(7):623--666, 2008.

[90] Domenico Fiorenza. An introduction to the Batalin-Vilkovisky formalism. 2004. arXiv:math/0402057v2.

[91] Stefan Cordes, Gregory Moore, and Sanjaye Ramgoolam. Lectures on 2d Yang-Mills theory, equivariant cohomology and topological field theories. arXiv:hep-th/9411210v2.

[92] Donald C. Ferguson. A theorem of Looman-Menchoff. `http://digitool.library.mcgill.ca/thesisfile111406.pdf`.

[93] Holger Lyre. Berry phase and quantum structure. arXiv:1408.6867.

[94] Florin Belgun. Gauge theory. `http://www.math.uni-hamburg.de/home/belgun/Gauge4.pdf`.

[95] Vladimir Itskov, Peter J. Olver, and Francis Valiquette. Lie completion of pseudogroups. *Transformation Groups*, 16:161--173, 2011.

[96] Richard Borcherds. Lie groups. `https://math.berkeley.edu/~reb/courses/261/`.

[97] Andrei Losev. From Berezin integral to Batalin-Vilkovisky formalism: A mathematical physicist's point of view. 2007.

[98] Edward Witten. Coadjoint orbits of the Virasoro group. *Comm. Math. Phys.*, 114(1):1--53, 1988.

[99] Sidney R. Coleman and Jeffrey E. Mandula. All possible symmetries of the S-matrix. *Phys. Rev.*, 159, 1967.

[100] Emily Riehl. Monoidal algebraic model structures. *Journal of Pure and Applied Algebra*, 217(6):1069--1104, 2013.

[101] Valter Moretti. Mathematical foundations of quantum mechanics: An advanced short course. *International Journal of Geometric Methods in Modern Physics*, 13, 2016.

[102] Antonio Michele Miti. Homotopy comomentum maps in multisymplectic geometry, 2021.

[103] Niclas Sandgren and Petre Stoica. On moving average parameter estimation. Technical Report 2006-022, Department of Information Technology, Uppsala University, 2006.

[104] John E. Roberts. Spontaneously broken gauge symmetries and superselection rules. 1974.

[105] Jean Gallier. Clifford algebras, Clifford groups, and a generalization of the quaternions, 2008. arXiv:0805.0311.

[106] Bozhidar Z. Iliev. Normal frames for general connections on differentiable fibre bundles. arXiv:math/0405004.

[107] Piotr Stachura. Short and biased introduction to groupoids. arXiv:1311.3866.

[108] Fosco Loregian. Coend calculus. arXiv:1501.02503.

[109] Frederic Schuller. Lectures on the geometric anatomy of theoretical physics. https://www.youtube.com/channel/UC6SaWe7xeOp31Vo8cQG1oXw.

[110] Nima Amini. Infinite-dimensional Lie algebras. https://people.kth.se/~namini/PartIIIEssay.pdf.

[111] Peter Selinger. Lecture notes on lambda calculus.

[112] Nigel Hitchin. Lectures on special Lagrangian submanifolds. https://arxiv.org/abs/math/9907034v1, 1999.

[113] Olivia Caramello. Lectures on topos theory at the university of Insubria. https://www.oliviacaramello.com/Teaching/Teaching.htm.

[114] Derek Sorensen. An introduction to characteristic classes. http://derekhsorensen.com/docs/sorensen-characteristic-classes.pdf, 2017.

[115] Arun Debray. Characteristic classes. https://web.ma.utexas.edu/users/a.debray/lecture_notes/u17_characteristic_classes.pdf.

[116] Jonathan R. Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. Technical report, 1994.

[117] Pascal Lambrechts.

[118] Chris Tiee. Contravariance, covariance, densities, and all that: An informal discussion on tensor calculus. https://ccom.ucsd.edu/~ctiee/notes/tensors.pdf, 2006.

[119] Emily Riehl. Homotopy (limits and) colimits. http://www.math.jhu.edu/~eriehl/hocolimits.pdf.

[120] Andreas Gathmann. Algebraic geometry. https://www.mathematik.uni-kl.de/~gathmann/class/alggeom-2019/alggeom-2019.pdf.

[121] Will J. Merry. Algebraic topology. https://www.merry.io/algebraic-topology.

[122] Stacks project. https://stacks.math.columbia.edu/.

[123] The nlab. https://ncatlab.org/nlab.

[124] Wikipedia. https://www.wikipedia.org/.

Index