Compendium of Mathematics & Physics

Nicolas Dewolf

August 10, 2022

Contents

Chapter 1

Probability Theory

The majority of this chapter uses the language of measure theory. For an introduction see Chapter **??**.

1.1 Probability

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

Definition 1.1.1 (Kolmogorov axioms). A probability space (Ω, Σ, P) is a measure space **??** with normalized measure $P(X) = 1$. The set Ω is called the **sample space**.

Definition 1.1.2 (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 1.1.3 (σ-algebra of a random variable). Let X be a random variable defined on a probability space (Ω, Σ, P) and denote the Borel σ-algebra of $\mathbb{R}$ by $\mathcal{B}$. The following family of sets is a σ-algebra:

$$X^{-1}(\mathcal{B}) := \{S \in \Sigma \mid \exists B \in \mathcal{B} : S = X^{-1}(B)\}. \tag{1.1}$$

Notation 1.1.4. The σ-algebra generated by the random variable X is often denoted by $\mathcal{F}_X$, analogous to **??**.

Definition 1.1.5 (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 1.1.6 (Union). Let A, B be two events. The probability that at least one of them occurs is given by the following formula:

$$P(A \cup B) = P(A) + P(B) + P(A \cap B). \tag{1.2}$$

Definition 1.1.7 (Disjoint events). Two events A and B are said to be disjoint if they cannot happen at the same time:

$$P(A \cap B) = 0. \tag{1.3}$$

Corollary 1.1.8. If A and B are disjoint, the probability that both A and B occur is just the sum of their individual probabilities.

Formula 1.1.9 (Complement). Let A be an event. The probability of A being false is denoted as $P(\overline{A})$ and is given by

$$P(\overline{A}) = 1 - P(A). \tag{1.4}$$

Corollary 1.1.10. From the previous equation and de Morgan's laws (**??**) and (**??**), one can derive the following formula:

$$P(\overline{A} \cap \overline{B}) = 1 - P(A \cup B). \tag{1.5}$$

1.2 Conditional probability

Definition 1.2.1 (Conditional probability). Let A, B be two events. The probability of A given that B is true is denoted as $P(A|B)$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.6}$$

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

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1.7}$$

Formula 1.2.3. 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:

$$P(A) = \sum_{n=1}^{\infty} P(A|B_n)P(B_n). \tag{1.8}$$

Definition 1.2.4 (Independent events). Let A, B be two events. A and B are said to be independent if they satisfy the following relation:

$$P(A \cap B) = P(A)P(B). \tag{1.9}$$

Corollary 1.2.5. If A and B are two independent events, Bayes's theorem simplifies to

$$P(A|B) = P(A). \tag{1.10}$$

The above definition can be generalized to multiple events:

Definition 1.2.6. 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 1.2.7 (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:

$$P(F_{i_1} \cap \cdots \cap F_{i_k}) = P(F_{i_1}) \cdots P(F_{i_k}). \tag{1.11}$$

Corollary 1.2.8. Let X, Y be two random variables. X and Y are independent if the σ-algebras generated by them are independent.

1.3 Probability distribution

Definition 1.3.1 (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 $\mathbb{R}$:

$$P_X(B) = P\big(X^{-1}(B)\big). \tag{1.12}$$

This measure is called the probability distribution of X.

Example 1.3.2 (Rademacher variable). A random variable on $\Omega = \{-1, 1\}$ with probability distribution $P_X(-1) = P_X(1) = \frac{1}{2}$.

Definition 1.3.3 (Density). Let $f \geq 0$ be an integrable function and recall Property **??**. The function f is called the density of the measure $P(A) := \int_A f \, d\lambda$ (with respect to the Lebesgue measure λ). If the measure is a probability measure, i.e. is normalized to 1, f is called a **probability density function**.

More generally, by the Radon-Nikodym theorem **??**, every absolutely continuous probability distribution P is of the form

$$P(A) = \int_A f \, d\lambda \tag{1.13}$$

for some integrable function f.

In the case where P is discrete, i.e. one works with respect to the counting measure, the Radon-Nikodym derivative is called the **probability mass function**. (In this compendium this function will also often be called the density function.)

Definition 1.3.4 (Cumulative distribution function). Consider a random variable X and its associated distribution P_X. The cumulative distribution function $F_X : \mathbb{R} \to [0, 1]$ is defined as follows:

$$F_X(a) := P_X(\{x \in \mathbb{R} \mid x \leq a\}). \tag{1.14}$$

Theorem 1.3.5 (Skorokhod's representation theorem). *Let $F : \mathbb{R} \to [0, 1]$ be a function that satisfies the following three properties:*

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

There exists a random variable $X : [0, 1] \to \mathbb{R}$ defined on the probability space $([0, 1], \mathcal{B}_{[0,1]}, \lambda_{[0,1]})$ such that $F = F_X$, where $\mathcal{B}_{[0,1]}$ is the Borel σ-algebra of $[0, 1]$ with its Euclidean topology.

The following theorem is a specific instance of the more general change-of-variables formula:

Theorem 1.3.6 (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 $g \in L^1(\mathbb{R})$:*

$$\int_\Omega g \circ X \, dP = \int_{\mathbb{R}} g \, dP_X. \tag{1.15}$$

Remark 1.3.7. 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 1.3.8. Consider an absolutely continuous probability function P 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 P.

$$\int_{\mathbb{R}^n} g\, dP = \int_{\mathbb{R}^n} f(x)g(x)\, dx \tag{1.16}$$

Corollary 1.3.9. The previous formula together with Theorem 1.3.6 gives rise to

$$\int_{\Omega} g \circ X\, dP = \int_{\mathbb{R}^n} f_X(x)g(x)\, dx. \tag{1.17}$$

Formula 1.3.10. 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

$$f_g(y) = f(g^{-1}(y))\left|\frac{\mathrm{d}g^{-1}}{\mathrm{d}y}(y)\right|. \tag{1.18}$$

Weak convergence of measures **??** induces a notion for convergence of random variables:

Definition 1.3.11 (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 cumulative 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. This is equivalent to requiring that the associated probability measures P_{X_n} converge weakly to P_X (Definition **??**).

Notation 1.3.12. 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 1.3.13 (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:*

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

Definition 1.3.14 (Convergence in probability). A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables on a metric space (Ω, d) is said to converge in probability to a random variable Y if for all $\varepsilon > 0$ the following statement holds:

$$\lim_{n\to\infty} \Pr(d(X_n, X) > \varepsilon) = 0. \tag{1.19}$$

Convergence in probability implies convergence in distribution.

Definition 1.3.15 (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:

$$\eta_X(x) := \delta_x. \tag{1.20}$$

The multiplication map is defined as follows:

$$\mu_X(Q)(U) := \int_{P\in\mathbb{P}X} \mathrm{ev}_U(P)\, dQ. \tag{1.21}$$

1.4 Moments

1.4.1 Expectation value

Definition 1.4.1 (Expectation value). Let X be random variable defined on a probability space (Ω, Σ, P).

$$\mathrm{E}[X] := \int_\Omega X \, dP \tag{1.22}$$

Notation 1.4.2. Other common notations are $\langle X \rangle$ and μ_X. However, the latter might be confused with a general measure on the space X and will, therefore, not be used here.

Property 1.4.3 (Markov's inequality). Let X be a random variable. For every constant $a > 0$ the following inequality holds:

$$\Pr(X \geq a) \leq \frac{\mathrm{E}[X]}{a}. \tag{1.23}$$

Definition 1.4.4 (Moment of order r). The moment of order r is defined as the expectation value of the r^{th} power of X. By Equation (1.17) this becomes

$$\mathrm{E}[X^r] = \int_{\mathbb{R}} x^r f_X(x) \, dx. \tag{1.24}$$

Definition 1.4.5 (Central moment of order r).

$$\mathrm{E}[(X - \mu)^r] = \int_{\mathbb{R}} (x - \mu)^r f_X(x) \, dx \tag{1.25}$$

Remark 1.4.6. 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 1.4.7 (Variance). The central moment of order 2 is called the variance:

$$\mathrm{Var}[X] := \mathrm{E}\big[(X - \mu)^2\big]. \tag{1.26}$$

Definition 1.4.8 (Standard deviation).

$$\sigma_X := \sqrt{V[X]} \tag{1.27}$$

Property 1.4.9. If $\mathrm{E}[|X|^n]$ is finite for some $n > 0$, then $\mathrm{E}\big[X^k\big]$ exists and is finite for all $k \leq n$.

Property 1.4.10 (Chebyshev's inequality). Let X be a nonnegative random variable. For every constant $a > 0$ the following inequality holds:

$$\Pr(|X - \mathrm{E}[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}. \tag{1.28}$$

Definition 1.4.11 (Moment generating function).

$$M_X(t) := \mathrm{E}\big[e^{tX}\big] = \int_{\mathbb{R}} e^{tx} f_X(x) \, dx \tag{1.29}$$

Property 1.4.12. If the moment generating function exists, the moments $\mathrm{E}[X^n]$ can be expressed in terms of M_X:

$$\mathrm{E}[X^n] = \left. \frac{\mathrm{d}^n M_X(t)}{\mathrm{d}t^n} \right|_{t=0}. \tag{1.30}$$

Method 1.4.13 (Chernoff bound). The Chernoff bound for a random variable gives a bound on the tail probabilities. For all constants $\lambda > 0$, the Markov inequality implies the following statement:

$$\Pr(X \geq a) = \Pr(e^{\lambda X} \geq e^{\lambda a}) \leq \frac{\mathrm{E}\big[e^{\lambda X}\big]}{e^{\lambda a}}. \tag{1.31}$$

If one has more information about the moment generating function, the Chernoff bound can be used to obtain improved concentration inequalities by optimizing over λ.

Property 1.4.14 (Hoeffding's inequalities). Consider a collection of bounded, independent random variables $X_1, \ldots, X_n$. Without loss of generality one can assume that they are bounded by the unit intevral, i.e. $0 \leq X_i \leq 1$. For every constant $\lambda \geq 0$ the following inequality holds:

$$\Pr\big(\overline{X} - \mathrm{E}\big[\overline{X}\big] \geq \lambda\big) \leq \exp(-2n\lambda^2). \tag{1.32}$$

If one can sharpen the bounds for the variables such that $X_i \in [a_i, b_i]$, then

$$\Pr\big(\overline{X} - \mathrm{E}\big[\overline{X}\big] \geq \lambda\big) \leq \exp\left(-\frac{2n^2\lambda^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{1.33}$$

Definition 1.4.15 (Characteristic function).

$$\varphi_X(t) := \mathrm{E}\big[e^{itX}\big] \tag{1.34}$$

Property 1.4.16. 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 1.4.17. If $\varphi_X(t)$ is k times continuously differentiable, then X has a finite k^{th} moment and

$$\mathrm{E}\Big[X^k\Big] = \frac{1}{i^k}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\varphi_X(0). \tag{1.35}$$

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

Formula 1.4.18 (Inversion formula). Let X be a random variable. If the CDF of X is continuous at $a, b \in \mathbb{R}$, then

$$F_X(b) - F_X(a) = \lim_{c\to\infty}\frac{1}{2\pi}\int_{-c}^{c}\frac{e^{-ita} - e^{-itb}}{it}\varphi_X(t)\,dt. \tag{1.36}$$

Formula 1.4.19. If $\varphi_X(t)$ is integrable, the CDF is given by:

$$f_X(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\varphi_X(t)\,dt. \tag{1.37}$$

Remark 1.4.20. This formula implies that the density function and the characteristic function form a Fourier transform pair.

1.4.2 Correlation

Property 1.4.21. 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 measurable bounded functions f, g.

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

Definition 1.4.22 (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 1.4.23 (Covariance). The covariance of two random variables X, Y is defined as follows:

$$\mathrm{cov}(X, Y) := \langle X_c \mid Y_c \rangle = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]. \tag{1.38}$$

Some basic math gives

$$\mathrm{cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]. \tag{1.39}$$

Definition 1.4.24 (Correlation). The correlation of two random variables X, Y is defined as the cosine of the angle between X_c and Y_c:

$$\rho_{XY} := \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{1.40}$$

Corollary 1.4.25. From Theorem 1.4.21 it follows that independent random variables are uncorrelated.

Corollary 1.4.26. If the random variables X and Y are uncorrelated, they satisfy $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

Formula 1.4.27 (Bienaymé formula). Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of independent (or uncorrelated) random variables. Their variances satisfy the following equation:

$$\mathrm{Var}\left[\sum_{i=1}^{\infty} X_i\right] = \sum_{i=1}^{\infty} \mathrm{Var}[X_i]. \tag{1.41}$$

1.4.3 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 **??** 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:

$$\forall Z \in L^2(\mathcal{G}) : \langle X - Y \mid Z \rangle \equiv \int_\Omega (X - Y) Z \, dP = 0. \tag{1.42}$$

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

$$\int_G X \, dP = \int_G Y \, dP \tag{1.43}$$

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

Definition 1.4.28 (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 (1.43) for every $G \in \mathcal{G}$. This variable Y is called the conditional expectation of X given $\mathcal{G}$ and it is denoted by $\mathrm{E}[X \mid \mathcal{G}]$:

$$\int_G \mathrm{E}[X \mid \mathcal{G}]\, dP = \int_G X\, dP. \tag{1.44}$$

Remark 1.4.29. Although this construction was based on orthogonal projections, one could as well have used the (signed) Radon-Nikodym theorem **??** since $G \mapsto \int_G X\, dP$ is absolutely continuous with respect to $P|_{\mathcal{G}}$.

Property 1.4.30. 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

$$\mathrm{E}[X \mid \mathcal{G}] = X \text{ a.s.} \tag{1.45}$$

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

$$\mathrm{E}[X \mid \mathcal{G}] = \mathrm{E}[X] \text{ a.s.} \tag{1.46}$$

1.5 Joint distributions

Definition 1.5.1 (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

$$P_{(X,Y)}(B) = P((X, Y)^{-1}(B)). \tag{1.47}$$

Definition 1.5.2 (Joint density). If the probability measure from the previous definition can be written as

$$P_{(X,Y)}(B) = \int_B f_{(X,Y)}(x, y)\, dxdy \tag{1.48}$$

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

Definition 1.5.3 (Marginal distribution). The distributions of the one-dimensional random variables is determined by the joint distribution:

$$P_X(A) = P_{(X,Y)}(A \times \mathbb{R})\,, \tag{1.49}$$
$$P_Y(A) = P_{(X,Y)}(\mathbb{R} \times A). \tag{1.50}$$

Corollary 1.5.4. If the joint density exists, the marginal distributions are absolutely continuous and the associated density functions are given by

$$f_X(x) = \int_{\mathbb{R}} f_{(X,Y)}(x, y)\, dy\,, \tag{1.51}$$
$$f_Y(y) = \int_{\mathbb{R}} f_{(X,Y)}(x, y)\, dx. \tag{1.52}$$

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

Property 1.5.5 (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:

$$P_{(X,Y)} = P_X \otimes P_Y. \tag{1.53}$$

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

Formula 1.5.6 (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:

$$f_Z(z) := f * g(z) = \int_{\mathbb{R}} g(x)h(z-x)\,dx = \int_{\mathbb{R}} g(z-y)h(y)\,dy\,, \tag{1.54}$$

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

Formula 1.5.7 (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

$$f_Z(z) = \int_{\mathbb{R}} g(x)h(z/x)\,\frac{dx}{|x|} = \int_{\mathbb{R}} g(z/y)h(y)\,\frac{dy}{|y|}\,, \tag{1.55}$$

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

Corollary 1.5.8. 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:

$$\mathcal{M}\{f\} = \mathcal{M}\{g\}\mathcal{M}\{h\}. \tag{1.56}$$

Formula 1.5.9 (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

$$h(y \mid X \in A) = \frac{\int_A f_{(X,Y)}(x,y)\,dx}{\int_A f_X(x)\,dx}. \tag{1.57}$$

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

$$h(y \mid A = a) := \frac{f_{(X,Y)}(a,y)}{f_X(a)}\,, \tag{1.58}$$

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:

$$P(Y \in B \mid X = a) := \int_B h(y \mid X = a)\,dy. \tag{1.59}$$

Formula 1.5.10 (Conditional expectation).

$$\mathrm{E}[Y \mid X](\omega) = \int_{\mathbb{R}} yh(y \mid X(\omega))\,dy \tag{1.60}$$

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:

$$\int_A \mathrm{E}[Y \mid X]\,dP = \int_A Y\,dP. \tag{1.61}$$

This implies that the conditional expectation $\mathrm{E}[Y \mid X]$ on $\mathcal{F}_X$ coincides with Definition 1.4.28.

Applying Property 1.4.30 to the case $\mathcal{G} = \mathcal{F}_X$ gives the law of total expectation:

Property 1.5.11 (Law of total expectation[1]).

$$\mathrm{E}[\mathrm{E}[Y \mid X]] = \mathrm{E}[Y] \tag{1.62}$$

Theorem 1.5.12 (Bayes's theorem). *The conditional density can be computed without prior knowledge of the joint density:*

$$g(x \mid y) = \frac{h(y \mid x) f_X(x)}{f_Y(y)}. \tag{1.63}$$

1.6 Stochastic calculus

Definition 1.6.1 (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 1.6.2 (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 1.6.3 (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 1.6.4 (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 1.6.5 (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

$$\{\tau \leq t\} \in \mathbb{F}_t \tag{1.64}$$

for all t. The stopping time is a ''time indicator'' that only depends on the knowledge of the process up to time $t \in T$.

1.6.1 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 1.6.6 (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**).

[1] Also called the **tower property**.

Example 1.6.7 (Doob martingale). Consider an integrable random variable X and a filtration $\mathbb{F}$. The associated Doob martingale (a martingale with respect to $\mathbb{F}$) is given by

$$Y_t := \mathrm{E}[X \mid \mathbb{F}_t]. \tag{1.65}$$

Property 1.6.8 (Doob-Ville inequality). Consider a càdlàg submartinagle $(X_t)_{t\in T}$.

$$\Pr\left(\sup_{t\leq\tau} X_t \geq C\right) \leq \frac{\mathrm{E}[\max(0, X_\tau)]}{C} \tag{1.66}$$

for all $C \geq 1$ and $\tau \in T$.

The following property generalizes the Hoeffding inequalities 1.4.14:

Property 1.6.9 (Hoeffding-Azuma inequality). Let $(X_n)_{n\in\mathbb{N}}$ be a (super)martingale with bounded differences, i.e. there exist constants $c_k > 0$ such that

$$|X_k - X_{k-1}| \leq c_k. \tag{1.67}$$

The following inequality holds for all $\lambda \geq 0$:

$$\Pr(X_N - X_0 \geq \lambda) \leq \exp\left(-\frac{\lambda^2}{2\sum_{i=1}^N c_i^2}\right). \tag{1.68}$$

A symmetric result for the lower tail holds for (sub)martingales. Moreover, if there exist predictable processes $(A_n)_{n\in\mathbb{N}}, (B_n)_{n\in\mathbb{N}}$ such that

$$A_k \leq X_k - X_{k-1} \leq B_k \tag{1.69}$$

and

$$B_k - A_k \leq c_k \tag{1.70}$$

for all $k \in \mathbb{N}$, the inequality can be sharpened:

$$\Pr(X_N - X_0 \geq \lambda) \leq \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^N c_i^2}\right). \tag{1.71}$$

Now, consider a function $f : \Omega^n \to \mathbb{R}$ such that

$$\sup_{x_1,\ldots,x_n,x_k'} |f(x_1, \ldots, x_k, \ldots, x_n) - f(x_1, \ldots, x_k', \ldots, x_n)| \leq c_k \tag{1.72}$$

for all $k \in \mathbb{N}$. By applying the above inequalities to the Doob martingale

$$Z_m := \mathrm{E}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_m], \tag{1.73}$$

one obtains the following inequality:

$$\Pr(f(X_1, \ldots, X_n) - \mathrm{E}[f] \geq \lambda) \leq \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n c_i^2}\right). \tag{1.74}$$

This inequality is sometimes called the **McDiarmid inequality**.

Theorem 1.6.10 (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:*

$$A_0 = 0 \qquad M_0 = 0 \tag{1.75}$$

$$\Delta A_t = \mathrm{E}[\Delta X_t \mid \mathbb{F}_{t-1}] \qquad \Delta M_t = \Delta X_t - \Delta A_t. \tag{1.76}$$

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

Corollary 1.6.11. Consider the special case $X = Y^2$ for some martingale Y. One can show the following property:

$$\Delta A_t = \mathrm{E}\big[(\Delta Y_t)^2 \mid \mathbb{F}_{t-1}\big] \qquad \forall t \in \mathbb{R}_+. \tag{1.77}$$

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 1.6.12 (Discrete stochastic integral[2]). 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:

$$(X \cdot M)_t(\omega) := \sum_{i=1}^{t} X(\omega)_i \Delta M_i(\omega)\,, \tag{1.78}$$

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

Property 1.6.13. If the process $(X_n)_{n\in\mathbb{N}}$ is bounded, the stochastic integral defines a martingale.

Property 1.6.14 (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 $n \geq 0$:

$$\mathrm{E}\Big[(X \cdot M)_n^2\Big] = \mathrm{E}\big[(X^2 \cdot [M])_n\big]. \tag{1.79}$$

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).

1.6.2 Markov processes

Definition 1.6.15 (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

$$P(X_t \mid \mathbb{F}_s) = P(X_t \mid X_s) \tag{1.80}$$

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

$$P(X_t \mid X_{t-1}, \ldots, X_{t-r}) = P(X_t \mid X_{t-1}) \tag{1.81}$$

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

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

1.7 Information theory

Definition 1.7.1 (Self-information). The self-information of an event x described by a distribution P is defined as follows:

$$I(x) := -\ln P(x). \tag{1.82}$$

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: $I(x) = 0$.[3]
- Events that are very rare contain a lot of information.
- Independent events contribute additively to the information.

Definition 1.7.2 (Shannon entropy). The amount of uncertainty in a discrete distribution P is characterized by its (Shannon) entropy

$$H(P) := \mathrm{E}[I(X)] = -\sum_i P_i \ln(P_i). \tag{1.83}$$

Definition 1.7.3 (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:

$$D_{\mathrm{KL}}(P\|Q) := \int_\Omega \log\left(\frac{P}{Q}\right) dP. \tag{1.84}$$

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 1.7.4 (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:

$$D_{\mathrm{KL}}(P\|Q) \geq 0. \tag{1.85}$$

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

1.8 Extreme value theory

Definition 1.8.1 (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:

$$F_u(y) = \Pr(X - u \leq y \mid X > u) = \frac{P(u+y) - P(u)}{1 - P(u)}. \tag{1.86}$$

Definition 1.8.2 (Extreme value distribution). The extreme value distribution is given by the following formula:

$$F(x;\xi) = \exp\Big(-(1+x\xi)^{-1/\xi}\Big). \tag{1.87}$$

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

$$F(x;0) = \exp(-e^{-x}). \tag{1.88}$$

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

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

Definition 1.8.3 (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 $F^n(a_n x + b_n) \longrightarrow H(x)$.

Theorem 1.8.4 (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 1.8.5 (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 $u \longrightarrow \infty$.*

1.9 Copulas

Property 1.9.1 (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 1.9.2 (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 1.9.3 (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:

$$\int_B dC := \sum_{\mathbf{z}\in\prod_i\{a_i,b_i\}} (-1)^{N_b(\mathbf{z})} C(\mathbf{z}) \geq 0\,, \tag{1.89}$$

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

Theorem 1.9.4 (Sklar). *For every joint distribution function H with marginals F_i there exists a unique copula C such that*

$$H(x_1, \ldots, x_d) = C\big(F_1(x_1), \ldots, F_d(x_d)\big). \tag{1.90}$$

Property 1.9.5 (Fréchet-Hoeffding bounds). Every copula $C : [0,1]^d \to [0,1]$ is bounded in the following way:

$$\max\left(\sum_{i=1}^{d} u_i - d + 1, 0\right) \leq C(u_1, \ldots, u_d) \leq \min_i u_i \tag{1.91}$$

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.[4]

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

Definition 1.9.6 (Extreme value copula). A copula C for which there exists a copula $\widetilde{C}$ such that

$$\left[\widetilde{C}(u_1^{1/n}, \ldots, u_d^{1/n})\right]^n \longrightarrow C(u_1, \ldots, u_d) \tag{1.92}$$

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

Property 1.9.7. A copula C is an extreme value copula if and only if it is stable in the following sense:

$$C(u_1, \ldots, u_d) = \left[C(u_1^{1/n}, \ldots, u_d^{1/n})\right]^n \tag{1.93}$$

for all $n \geq 1$.

1.10 Randomness ♣

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

Definition 1.10.1 (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

$$C(\kappa) := \min\{|\sigma| \mid \sigma \text{ is finite} \wedge U(\sigma) = \kappa\}. \tag{1.94}$$

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 1.10.2. 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 1.10.3. 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.

1.11 Optimal transport

In this section a new notion of atomicity of measures will be used:

Definition 1.11.1. A measure on $\mathbb{R}^n$ is said to **give mass to small sets** if there exists a subset of *Hausdorff dimension* $n-1$ (or smaller) that has nonzero measure.

1.11.1 Kantorovich duality

The problem of optimal transport constitutes the search of the most cost efficient transportation scheme that connects a set of producers to a set of consumers. Assume that these are described by the probability spaces (X, Σ_X, μ_X) and (Y, Σ_Y, μ_Y), respectively.

Definition 1.11.2 (Cost function). A measurable function $X \times Y \to \overline{\mathbb{R}}$.

Definition 1.11.3 (Transportation scheme). A transportation scheme or **transference plan** is a joint distribution $\pi \in \mathbb{P}(X \times Y)$ whose marginals coincide with μ_X and μ_Y.

Definition 1.11.4 (Monge-Kantorovich problem). The optimal transportation scheme for a given cost function according to *Kantorovich* is the solution of the following optimization problem:

$$\inf_{\pi\in\mathbb{P}(X\times Y)} \mathrm{E}_\pi[c] = \inf_{\pi\in\mathbb{P}(X\times Y)} \int_{X\times Y} c(x,y)\, d\pi(x,y). \tag{1.95}$$

The original problem of optimal transportation was considered by *Monge*. However, he studied a restricted problem, where every producer only delivers to a unique consumer. In this case the joint distributions have a specific form, namely

$$\int_{X\times Y} c(x,y)\, d\pi(x,y) = \int_X c(x,T(x))\, d\mu_X(x) \tag{1.96}$$

for some measurable function $T : X \to Y$ such that $T_*\mu_X = \mu_Y$.

Example 1.11.5 (Finite state spaces). Consider the case where both X and Y are finite of the same size and are both equipped with the uniform distribution. In this case the joint distributions π can be represented by *bistochastic matrices*, i.e. matrices with nonnegative entries such that every column and every row sums to one. This also implies that the optimization problem reduces to a linear problem on a convex, compact subset. This allows one to use Choquet's theorem **??** to restrict the attention to the extremal points, which in this case are given by permutation matrices. So, the optimal solution is given by the optimal one-to-one pairing of producers and consumers.

Property 1.11.6 (Kantorovich duality). Let X, Y be Polish spaces **??** and consider a lower semicontinuous function $c : X \times Y \to \overline{\mathbb{R}^+}$ (Definition **??**). Denote by $\mathbb{P}_{\mathrm{Borel}}(\mu,\nu)$ the space of Borel measures on $X \times Y$ whose marginals are given by μ_X and μ_Y. Moreover, denote by $\Phi_c \subseteq \mathcal{L}^1(X) \times \mathcal{L}^1(Y)$ the space of pairs of integrable functions satisfying

$$c_X(x) + c_Y(y) \le c(x,y) \tag{1.97}$$

for μ_X-almost all $x \in X$ and μ_Y-almost all $y \in Y$. Then

$$\inf_{\pi\in\mathbb{P}_{\mathrm{Borel}}(\mu,\nu)} \int_{X\times Y} c\, d\pi = \sup_{(c_X,c_Y)\in\Phi_c} \int_X c_X\, d\mu_X + \int_Y c_Y\, d\mu_Y \tag{1.98}$$

and the this problem admits a solution. Moreover, one can restrict the space of would-be solutions on the right-hand side to those that are also bounded and continuous without changing the solution.

Definition 1.11.7 (Kantorovich distance). Let X be a Polish space and consider a lower semicontinuous metric d on X. The Kantorovich(-Rubinstein) distance $\mathcal{T}_d$ between two Borel probability measures μ, ν on X is defined as the optimal transport cost between them:

$$\mathcal{T}_d(\mu,\nu) := \inf_{\pi\in\mathbb{P}_{\mathrm{Borel}}(X\times X)} \int_{X\times X} d(x,x')\, d\pi(x,x'). \tag{1.99}$$

If the metric d is the one inducing the topology on X, one obtains the definition of the **Wasserstein 1-metric**.

Theorem 1.11.8 (Kantorovich-Rubinstein). *If $X = Y$ and c is equal to some metric d on X, the Kantorovich distance is given by*

$$\mathcal{T}_d(\mu,\nu) = \sup\left\{ \int_X \varphi\, d\mu - \int_X \varphi\, d\nu \;\middle|\; \varphi \in \mathrm{Lip}(X,d) \cap \mathcal{L}^1(\mu) \cap \mathcal{L}^1(\nu) \wedge \|\varphi\|_{\mathrm{Lip}} \le 1 \right\}, \tag{1.100}$$

where

$$\|\varphi\|_{\mathrm{Lip}} := \sup_{x\neq x'\in X} \frac{|\varphi(x) - \varphi(x')|}{d(x,x')} \tag{1.101}$$

is the ***Lipschitz norm****.*

Property 1.11.9 (Translation invariance). The Kantorovich distance is invariant under translations by finite measures.

Property 1.11.10. When $X = Y = \mathbb{R}^n$ with d the Euclidean metric, the Kantorovich distance admits yet another description. In this case the Lipschitz norm is equal to the supremum norm of the gradient. This gives

$$\mathcal{T}_d(\mu, \nu) = \inf\{\|\sigma\|_1 \,|\, \nabla \cdot \sigma = \mu - \nu\}\,, \tag{1.102}$$

where the condition on σ makes sense by the Riesz-Markov theorem **??**.

1.11.2 Convex costs

In this section cost functions of the form

$$c(x, y) = h(x - y) \tag{1.103}$$

for some convex function $h : \mathbb{R}^n \to \mathbb{R}$ are considered. Moreover, the function h will be assumed to be at least differentiable with locally Lipschitz gradient.

Definition 1.11.11 (c-concave function). A function $f : \mathbb{R}^n \to [-\infty, +\infty[$, not identically $-\infty$, is said to be c-concave if there exists a set $A \subset \mathbb{R}^n \times \mathbb{R}$ such that

$$f(x) = \inf_{(x',\lambda)\in A} c(x, x') + \lambda. \tag{1.104}$$

Theorem 1.11.12 (Gangbo-McCann). *If c is strictly convex and μ does not give mass to small sets, the Monge-Kantorovich problem has a a.s. unique minimizer $\pi = (\mathbb{1} \times T)_*\mu$ with*

$$T(x) = x - (\nabla h)^{-1}\big(\nabla\psi(x)\big) \tag{1.105}$$

for some h-concave function $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$.

Remark 1.11.13. If h is a strictly convex function of the distance $\|x - y\|$, the theorem has to be modified:

- If $\mu \perp \nu$, the theorem still holds.
- If the measures are not singular, one has to restrict to transportation schemes that fix the shared mass. In effect, one removes the shared mass from the problem to recover the previous case.

Note that if h is sufficiently differentiable, the inverse ∇h^{-1} is equal to the gradient of the Legendre transform by Property **??**.

1.11.3 Concave costs

In this section cost functions of the form

$$c(x, y) = g(\|x - y\|) \tag{1.106}$$

for some concave function $g : \mathbb{R} \to \mathbb{R}$ are considered.

Property 1.11.14. Let c be strictly concave. If the transportation cost is not everywhere infinite and if μ does not give mass to small sets, then:

- If $\mu \perp \nu$, there exists a unique optimal transport scheme such that $\nu = T_*\mu$ with

  $$T(x) = x - (\nabla g)^{-1}\nabla\varphi(x) \tag{1.107}$$

  for some *c-concave function* φ.
- If the measures are not singular, there still exists a unique optimum by restricting to those schemes that fix shared mass.

1.11.4 Densities

Property 1.11.15 (Continuity equation). Let X be a complete smooth manifold and consider a family $(T_t)_{0\leq t\leq 1}$ of locally Lipschitz diffeomorphism on X such that $T_0 = \mathbb{1}_X$ with associated vector fields v_t. If μ is a probability measure on X, the family $(\mu_t := T_{t,*}\mu)_{0\leq t\leq 1}$ uniquely satisfies the **continuity equation**:

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mu_t v_t) = 0\,, \tag{1.108}$$

where the divergence of a measure is defined by duality.

Let $v : \mathbb{R}^n \to \mathbb{R}^n$ be an almost everywhere smooth vector field. This induces a linear, constant velocity flow as follows:

$$T_t(x) := x - tv(x). \tag{1.109}$$

If all T_t are diffeomorphisms, the Eulerian velocity field $v_t(x) := T_t^{-1}\big(v(x)\big)$ satisfies the Eulerian continuity equation:

$$\frac{\partial v_t}{\partial t} + (v_t \cdot \nabla)v_t = 0. \tag{1.110}$$

Formula 1.11.16. Given a solution of the continuity equation, the associated flow determines an optimal transport scheme for a cost function c if and only if

$$v_0 = -(\nabla c)^{-1}\nabla\psi \tag{1.111}$$

for some c-concave function ψ. Moreover, if $v_t = (\nabla c)^{-1}\nabla u$ for some function $u(t,x)$, then u satisfies the **Hamilton-Jacobi equation** with Hamiltonian c^*:

$$\frac{\partial u}{\partial t} + c^*(\nabla u) = 0. \tag{1.112}$$

In this section one considers absolutely continuous measures with respect to the Lebesgue measure on $\mathbb{R}^n$:

$$d\mu_X = \rho_0 dx \qquad\qquad d\mu_Y = \rho_1 dx. \tag{1.113}$$

The transport cost in the Monge problem can then be rewritten as

$$\int_{\mathbb{R}^n} c(x, T(x))\rho_0(x)\,dx \tag{1.114}$$

with

$$\int_{T^{-1}(A)} \rho_0(x)\,dx = \int_A \rho_1(x)\,dx \tag{1.115}$$

for all measurable $A \subset \mathbb{R}^n$. By the change-of-variables formula this (weak) integral equation is equivalent to the Jacobian equation for

$$\det(DT(x))\rho_1(T(x)) = \rho_0(x). \tag{1.116}$$

Example 1.11.17 (Euclidean metric). If the cost function c is the square of the Euclidean distance, the optimal transport mapping T, called the **Brenier map**, is given by the gradient of a convex potential:

$$T(x) = \nabla\varphi(x)\,, \tag{1.117}$$

and the optimal cost is equal to the square of the **Wasserstein 2-metric**:

$$\mathcal{T}_{\|\cdot\|_2^2}(\rho_0,\rho_1) = \inf_{\pi\in\mathbb{P}_{\text{Borel}}(\rho_0,\rho_1)} \int_{\mathbb{R}^n} \|x-x'\|\, d\pi(x,x') = W_2^2(\rho_0,\rho_1). \tag{1.118}$$

Moreover, this minimum is unique a.e.

It can also be shown that the flow acts affinely:

$$\sigma_t(x) = t\nabla\Phi(x) + (1-t)x. \tag{1.119}$$

In fact, the affinity of the flow can be shown more generally:

Property 1.11.18. Consider the time-dependent Monge-Kantorovich problem. If the differential cost $c : \mathbb{R}^n \to \mathbb{R}$ is strictly convex, the flows are given by straight lines:

$$x_t = x + t(x' - x). \tag{1.120}$$

This situation can be generalized to (complete) smooth manifolds, where the minimizers of ℓ^p-costs are geodesics with arc length parametrization.

It is possible to relate optimal transport to mechanics (Section **??**) in the following way:

Method 1.11.19 (Benamou-Brenier formulation). Let ρ_0 and ρ_1 describe the density of particles in a system at time steps $t = 0$ and $t = 1$. Assume that there exists a time-dependent velocity field $v : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. These are related by the *continuity equation* **??**:

$$\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho v) = 0. \tag{1.121}$$

The optimization problem now becomes minimizing the *action* or *kinetic energy*:

$$K(\rho, v) := \frac{1}{2}\int_{\mathbb{R}^n}\int_0^T \rho(t,x)\|v(t,x)\|^2\, dtdx. \tag{1.122}$$

By making the change of variables $(\rho, v) \longrightarrow (\rho, m := \rho v)$, one obtains a convex problem with a linear constraint (the continuity equation).

Property 1.11.20. The infimum of the Benamou-Brenier action is equal (up to constant factors) to the square of the Wasserstein 2-metric and, hence, gives an equivalent characterization of the Monge-Kantorovich problem for the Euclidean distance.

Chapter 2

Information Geometry ♣

The main reference for this chapter is [6]. For more information on differential geometry, see Chapter **??** and onwards.

2.1 Statistical manifolds

In this section an important subclass of Riemannian manifolds that admit two related flat connections will be introduced. These manifolds will formalize the geometric backbone of many statistical concepts and methods.

Definition 2.1.1 (Conjugate connections). Consider a Riemannian manifold (M, g) with an affine connection ∇. The conjugate (or dual) connection $\widetilde{\nabla}$ is uniquely defined by the following equation:

$$X(g(Y,Z)) = g(\nabla_X Y, Z) + g(Y, \widetilde{\nabla}_X Z)\,, \tag{2.1}$$

where $X, Y, Z \in TM$. Moreover, this construction is involutive:

$$\widetilde{\widetilde{\nabla}} = \nabla. \tag{2.2}$$

Property 2.1.2. Consider a pair of conjugate connections $\nabla, \widetilde{\nabla}$ on a Riemannian manifold (M, g) and denote their parallel transport maps by $\mathcal{P}$ and $\mathcal{P}'$, respectively. Although the metric is in general not preserved under either $\mathcal{P}$ or $\mathcal{P}'$, it is preserved under conjugate (or dual) transport:

$$g(v, w) = g\Big(\mathcal{P}_\gamma v, \widetilde{\mathcal{P}}_\gamma w\Big) \tag{2.3}$$

for every smooth path γ.

Property 2.1.3. Consider two conjugate connections $\nabla, \widetilde{\nabla}$ on a Riemannian manifold (M, g). The connection

$$\overline{\nabla} := \frac{\nabla + \widetilde{\nabla}}{2} \tag{2.4}$$

is metric(-preserving), i.e. $\overline{\nabla} g = 0$. Furthermore, if both ∇ and $\widetilde{\nabla}$ are torsion-free, then $\overline{\nabla}$ necessarily coincides with the Levi-Civita connection of g by Theorem **??**.

The above properties lead to the following definition:

Definition 2.1.4 (Statistical manifold). A Riemannian manifold (M, g) equipped with an affine connection that satisfies the **Codazzi condition**

$$\nabla_X g(Y, Z) = \nabla_Y g(X, Z)\,, \tag{2.5}$$

i.e. ∇g is totally symmetric. The Codazzi condition implies vanishing torsion and vice versa. The rank-3 tensor $T := \nabla g$ is sometimes called the **cubic tensor** or **Amari-Chentsov** tensor. In local coordinates the cubic tensor gives the difference between the Christoffel symbols of ∇ and $\widetilde{\nabla}$:

$$T_{ijk} = \widetilde{\Gamma}_{ijk} - \Gamma_{ijk}. \tag{2.6}$$

In the case where ∇ has nonvanishing torsion, one can generalize this definition by also relaxing the Codazzi equation:

$$\nabla_X g(Y, Z) - \nabla_Y g(X, Z) = -g\big(T^{\nabla}(X, Y), Z\big). \tag{2.7}$$

If this equation is satisfied for all X, Y and $Z \in TM$, the dual connection is torsion-free and the tuple (M, g, ∇) is called a **statistical manifold admitting torsion**.[1] The existence of a torsion-free connection is sufficient to turn a (pseudo-)Riemannian manifold into a statistical manifold admitting torsion.

Remark 2.1.5. One can show that the above definition is equivalent to that of a Riemannian manifold admitting a totally symmetric rank-3 tensor.

Definition 2.1.6 (Dually flat manifold). Consider a statistical manifold (M, g, ∇). If ∇ is flat, its conjugate $\widetilde{\nabla}$ is also flat and the tuple $(M, g, \nabla, \widetilde{\nabla})$ is called a dually flat manifold.

Because the affine connections are flat, they endow the manifold with an *affine structure*, i.e. there exist coordinate charts such that the coordinate-induced vector fields satisfy

$$\nabla_{\partial_i} \partial_j = 0 \tag{2.8}$$

for all $i, j \leq n$ and such that the transition functions are affine transformations. It can be shown that the conjugate connection induces a similar $\widetilde{\nabla}$-affine coordinate chart such that the coordinate-induced vector fields satisfy the following orthonormality condition:

$$g\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial y_j}\right) = \delta_i^j. \tag{2.9}$$

This coordinate system is called the **dual (coordinate) system**.

2.1.1 Divergences

Definition 2.1.7 (Divergence). Let M be a set. A smooth function $D(\cdot\|\cdot) : M \times M \to \mathbb{R}$ with the following properties is called a divergence (measure) on M:

1. **Positivity**: $D(p\|q) \geq 0$ for all $p, q \in M$, and
2. **Nondegeneracy**: $D(p\|q) = 0$ if and only if $p = q$.

The **dual divergence** D^* is defined by

$$D^*(p\|q) := D(q\|p) \tag{2.10}$$

for all $p, q \in M$.

[1] This situation arises in the study of quantum systems and density operators.

Property 2.1.8 (Induced metric). An interesting feature of these functions is that one can use their Hessians (with respect to either of the two arguments) to construct a Riemannian metric if M is a smooth manifold:

$$g_{ij}(\theta) := \frac{\partial^2 D}{\partial p^i \partial p^j}(p||q)\bigg|_{p=q=\theta} = \frac{\partial^2 D}{\partial q^i \partial q^j}(p||q)\bigg|_{p=q=\theta} = -\frac{\partial^2 D}{\partial p^i \partial q^j}(p||q)\bigg|_{p=q=\theta}. \tag{2.11}$$

Example 2.1.9 (f-divergences and α-divergences). Let f be a smooth convex function such that $f(1) = 0$ and let p, q be two probability distributions such that p is absolutely continuous with respect to q. The f-divergence is defined as follows:

$$D_f(p\|q) := \int_\Omega f\left(\frac{\mathrm{d}p}{\mathrm{d}q}\right) dq. \tag{2.12}$$

In the case where both p and q are absolutely continuous with respect to some given measure μ (with Radon-Nikodym derivatives g, h), one can rewrite the above formula as

$$D_f(p\|q) = \int_\Omega h(x) f\left(\frac{g(x)}{h(x)}\right) d\mu(x). \tag{2.13}$$

It is not hard to see that f-divergences remain invariant under affine transformations of the form

$$f(x) \longrightarrow f(x) + c(x-1)\,, \tag{2.14}$$

where the shift $x - 1$ is necessary to preserve the condition $f(1) = 0$. This implies that one can always choose $f'(1) = 0$ without loss of generality. Moreover, one can also easily see that these divergences transform linearly under scale transformations and, hence, one can also always choose $f''(1) = 1$. f-divergences with these properties are said to be **standard**.

A particular class of f-divergences are the α-divergences (in the sense of *Csiszár*[2]) where

$$f_\alpha(x) = \frac{1 - x^\alpha}{\alpha(1-\alpha)}. \tag{2.15}$$

Note that some authors replace $(1 - x^a)$ by $(x - x^a)$ since this does not make any difference when calculating the divergence for normalized distributions. For the cases $\alpha = 0, 1$ one can use a workaround. For $\alpha = 0$ one can take the limit of the above definition to obtain $f_0(x) = -\ln x$. This results in $D_0(p\|q) = D_{\mathrm{KL}}(q\|p)$. For $\alpha = 1$ one can look at the general expression of D_α and notice that it is invariant under the simultaneous exchanges $(\alpha \leftrightarrow 1-\alpha)$ and $(p \leftrightarrow q)$. Using this trick one can see that $D_1(p\|q) = D_{\mathrm{KL}}(p\|q)$.

Definition 2.1.10 (Bregman divergence). Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Because the function is convex, at every point $q \in \mathbb{R}^n$ the tangent plane to the graph of F is a supporting hyperplane, i.e. it lies underneath the graph everywhere and it touches the graph only at q. Using this hyperplane one can define the Bregman divergence as follows:

$$D_F(p\|q) := F(p) - F(q) - \vec{\nabla}F(q) \cdot (p - q)\,, \tag{2.16}$$

where the gradient is denoted by $\vec{\nabla}$ to avoid confusion with further occurences of the ∇-symbol for affine connections. This function gives the difference in "height" between the function value at p and the position of the tangent plane (defined by q) at p. Because the gradient of a convex function is monotonic, this difference will always increase the further the points are spread apart. (For convex functions this will in general only be nondecreasing. However, in the remainder of this chapter strict convexity will almost always be assumed.)

[2] *Tsallis* and *Rényi* introduced different divergences/entropies with the same name.

Example 2.1.11 (Kullback-Leibler divergence). The Kullback-Leibler divergence 1.7.3 can be obtained as the Bregman divergence associated to the (negative) Shannon entropy $F(\rho) := \sum_{i=1}^{n} \rho_i \ln \rho_i$. It is also equal to the f-divergence associated to the choice $f = x\ln(x)$.

A Bregman divergence can also be used to endow the underlying manifold with further structure. By restricting to strictly convex functions, i.e. requiring that the Hessian is positive-definite, one can perform a Legendre transformation **??** to obtain a new function:

$$\widetilde{F}(x^*) := x^* \cdot y - F(y)\,, \tag{2.17}$$

where $x^* = \vec{\nabla} F(y)$.[3] It can be shown that $\widetilde{F}$ is again (strictly) convex and, hence, also defines a Bregman divergence. This second Bregman divergence coincides with the dual divergence D_F^*:

$$D_F(p\|q) = D_{\widetilde{F}}(q^*\|p^*). \tag{2.18}$$

Using this relation one can also rewrite the original expression for the Bregman divergence:

$$D_F(p\|q) = F(p) + \widetilde{F}(q) - x^i(p)y_i(q). \tag{2.19}$$

Now, the two convex functions $F, \widetilde{F}$ define two coordinate systems that are related as follows:

$$y = \vec{\nabla} F \qquad \text{and} \qquad x = \vec{\nabla}\widetilde{F}(y). \tag{2.20}$$

However, convexity of functions is not preserved under arbitrary coordinate transformations and, hence, one should restrict the class of allowed coordinate transformations. To preserve convexity only affine transformations are allowed. In affine coordinates one can express any geodesic, i.e. any path γ such that $\nabla_{\dot\gamma}\dot\gamma = 0$, as a straight line:

$$\gamma_{q\to p}(t) \equiv tx(p) + (1-t)x(q). \tag{2.21}$$

Geodesics for the conjugate connection are called **dual geodesics**. It is important to note that the Legendre transform that maps the primary coordinates to the dual coordinates is in general not an affine transformation and, hence, does not preserve the dual structure. Moreover, it can be shown that neither of the parallel transport maps, although completely trivial due to the affine structure, are metric-preserving. However, parallel transporting one vector by ∇ and the other by $\widetilde{\nabla}$ does preserve the metric. The metric structures induced by the Hessians are also intertwined:

$$g_{ij} = \frac{\partial^2 F}{\partial x^i \partial x^j} \qquad \text{and} \qquad \widetilde{g}^{ij} = \frac{\partial^2 \widetilde{F}}{\partial y_i \partial y_j} \tag{2.22}$$

are mutual inverses. It can be concluded that a Bregman divergence endows a set with the structure of a dually flat manifold 2.1.6.

Example 2.1.12 (Euclidean distance). On $\mathbb{R}^n$ the most common choice of divergence measure is the Euclidean distance:

$$D_{\text{eucl}}(p\|q) := \frac{1}{2}\|p-q\|. \tag{2.23}$$

It is not hard to show that this function is in fact self-dual with respect to Legendre transformations. This also implies that the primary and dual structures on $\mathbb{R}^n$ (with respect to the Euclidean distance) coincide. The associated connections are equal to the (trivial) Levi-Civita connection.

[3] For general convex functions this relation is not necessarily invertible.

Property 2.1.13 (Bregman divergence). The dually flat structure on a dually flat manifold $(M, g, \nabla, \widetilde{\nabla})$ enables one to define two affine coordinate systems through two functions ψ, ϕ (called **potentials**). Because the connection ∇ is torsion-free and the metric is symmetric by definition, a function ψ can (locally) be found such that

$$g_{ij} = \partial_i \partial_j \psi. \tag{2.24}$$

The positive-definiteness of g further implies that ψ is convex. This implies that ψ can be used to define a Bregman divergence. The induced dually flat structure is exactly $(M, g, \nabla, \widetilde{\nabla})$.

2.1.2 Exponential families

The primary and dual affine geodesics are often given the names *e-geodesic* and *m-geodesic* in the literature. In this and the following section, this terminology is explained.

Definition 2.1.14 (Exponential family). Let $X : \Omega \to \mathbb{R}^k$ be a random variable. For every integer $n \in \mathbb{N}$, every collection of smooth functions $\{h_i : \mathbb{R}^k \to \mathbb{R}\}_{1 \leq i \leq n}$ and any smooth function $\lambda : \mathbb{R}^k \to \mathbb{R}$ one can define the following family of distributions indexed by some parameter $\theta \in \mathbb{R}^n$:

$$p(X; \theta) := \exp\big(h_i(X)\theta^i + \lambda(X) - \psi(\theta)\big). \tag{2.25}$$

The function $\psi(\theta)$ is introduced as a normalization function:

$$\psi(\theta) := \ln \int \exp(h_i(X)\theta^i) e^{\lambda(X)} \, d\mu(X). \tag{2.26}$$

The function λ can be removed through a redefinition of the measure:

$$d\mu(X) \longrightarrow d\nu(X) := \exp^{\lambda(X)} \, d\mu(X).$$

Remark 2.1.15. The function ψ is actually the cumulant-generating function (or free energy in physics terminology) of the sufficient statistics $h_i(X)$.

Such a family of exponential distributions forms a manifold with affine coordinates θ^i (these are also called the **natural parameters**). The dual coordinates $\nabla\psi(\theta)$ are the expectation values $\mathrm{E}[X]$ and the associated dual convex function ϕ is the Shannon entropy. Accordingly, the Bregman divergence associated to ψ is given by the dual Kullback-Leibler divergence:

$$D_\psi(\theta \| \theta') = D_{\mathrm{KL}}(\theta' \| \theta). \tag{2.27}$$

The metric induced by this structure is the Fisher information:

$$g_{ij} = \mathcal{I}_{ij}[X; \theta] := \mathrm{E}\left[\left(\frac{\partial}{\partial \theta^i} \ln p(X; \theta)\right)\left(\frac{\partial}{\partial \theta^j} \ln p(X; \theta)\right)\right]. \tag{2.28}$$

Now, consider an affine combination of natural parameters, hence an affine geodesic in the manifold of an exponential family:

$$\theta(t) := t\theta_2 + (1 - t)\theta_1.$$

The probability distributions along this path can themselves be interpreted as constituting an exponential family with natural parameter t and, therefore, one calls the primary geodesic $\theta(t)$ an **e-geodesic** ('e' for exponential).

2.1.3 Mixtures

Another important class of probability distributions is given by the mixture families:

Definition 2.1.16 (Mixture). Consider a collection of probability distributions $\{p_i(X)\}_{i\leq n}$. For every point $\eta \equiv (\eta_0, \ldots, \eta_n)$ in the probability simplex Δ^n, one can define the distribution

$$p(X;\eta) := \sum_{i=0}^{n} \eta_i p_i(X). \tag{2.29}$$

This mixture family forms a manifold with affine coordinates $\{\eta_i\}_{1\leq i\leq n}$ (note that η_0 can be calculated from the other weights and is therefore not an independent coordinate).

The (negative) Shannon entropy of a mixture defines a convex function φ and, as noted before, it induces the Kullback-Leibler divergence:

$$D_\varphi(\eta\|\eta') = D_{\mathrm{KL}}(\eta\|\eta'). \tag{2.30}$$

Example 2.1.17 (Discrete distribution). An interesting example of mixtures is given by the class of discrete (or categorical) distributions:

$$p(X;\eta) = \sum_{i=0}^{n} \eta_i \delta_i(X)\,, \tag{2.31}$$

where $\eta \in \Delta^n$. At the same time these models can be considered as distributions in an exponential family with affine coordinates $\theta^i := \ln \frac{p_i}{p_0}$. For these models the primary coordinates $\overline{\theta}$, dual to η, coincide with the natural parameters θ.

Consider two points with dual coordinates η_1, η_2. The dual geodesic connecting these points is of the form

$$\eta(t) = t\eta_2 + (1-y)\eta_1.$$

In the case of discrete distributions, where the dual coordinates are given by elements of the probability simplex Δ^n, one can see that such a geodesic induces a linear interpolation of distributions and accordingly defines a mixture family. For this reason one generally calls a dual geodesic an **m-geodesic** ('m' for mixture).[4]

Remark 2.1.18. The reader should be aware of an important source of confusion. The above sections would point to the following naming convention:

e-geodesic	$\leftrightarrow$	primary geodesic
m-geodesic	$\leftrightarrow$	dual geodesic

However, because the Kullback-Leibler divergence is the most widely used divergence measure, Equation (2.27) where the KL-divergence is the dual divergence, made it possible that the above convention got reversed in the bulk of the literature. This reversal also leads one to interchange "primary" and "dual" in most statements such as the Pythagorean and projections theorems.

To prevent confusion it is important that one pays attention to which divergence is used. In this text a distinction has been made (as much as possible) between the e/m-convention and the primary/dual convention. The second convention is the main choice here since this one is uniquely determined once one knows the divergence.

[4]For arbitrary families the dual geodesic does not necessarily induce a mixture of distributions.

2.1.4 Compatible divergences

The question to be answered in this section is the following one: ''Given a dually flat manifold, which divergences are compatible with this structure?''. In the previous sections it was shown that exponential families and mixture families naturally give rise to the Kullback-Leibler divergence. However, not all dually flat manifolds are induced by such families.

A basic requirement, as is generally the case with geometric structures, is the requirement that the divergences are invariant under coordinate transformations. To this end one needs the *monotonicity criterion of Chentsov*. This criterion states that no transformation should increase the divergence between two points (this corresponds to the idea that transformations can never increase the amount of information). Moreover, there exists a class of (noninvertible) transformations that leave the divergence invariant:

Definition 2.1.19 (Sufficient statistic). Consider a random variable X following the distribution $p(X;\theta)$. A transformation $\xi(X)$ is said to be sufficient (with respect to θ) if the distribution of X conditioned on $\xi(X)$ is independent of θ. The **Fisher-Neyman factorization theorem** states that this is equivalent to the existence of the following decomposition

$$p(X;\theta) = f(X)g_\theta(\xi(X)) \tag{2.32}$$

for some nonnegative functions f, g_θ.

The invariance criterion states that these transformations are the only transformations that leave the divergence invariant:

Axiom 2.1 (Invariance criterion). Consider a dually flat manifold M. Compatible divergences should satisfy the following inequality for all transformations $\overline{x} := \xi(x)$:

$$\overline{D}(\theta\|\theta') \leq D(\theta\|\theta'). \tag{2.33}$$

The equality holds if and only if the transformed variable $\overline{x}$ is a sufficient statistic.

Example 2.1.20 (f-divergences). An important class of invariant divergences on the manifold Δ^n is given by the f-divergences introduced in the beginning of this chapter. These also have the additional property that they are **decomposable**

$$D_f(p\|q) = \sum_{i=0}^{n} d(p_i, q_i) \tag{2.34}$$

for some nonnegative function d.

The following result gives a partial characterization of invariant divergences on the manifold of discrete distributions Δ^n:

Property 2.1.21. A divergence D on Δ^n ($n > 1$) is invariant and decomposable if and only if it it an f-divergence. If the induced geometric structure is required to be flat, then necessarily $D = D_{\mathrm{KL}}$. When extended to $\mathbb{R}^n_+$ (the discrete positive measures), the collection of all α-divergences is recovered.

Corollary 2.1.22. Because every Bregman divergence is flat, one can see that D_{KL} is the only Bregman divergence that is also an f-divergence.

One can also go a step further and ask which metrics can arise on such invariant structures. The answer is quite simple (at least for discrete distributions):

Theorem 2.1.23 (Chentsov). *Up to scaling, the only invariant metric that exists on Δ^n is the Fisher information metric. Extending this result to the manifold $\mathbb{R}^n_+$ of discrete positive measures, the only invariant metric on $\mathbb{R}^n_+$ is the Euclidean metric.*

Remark. Extensions to other families/manifolds of distributions can be found in the literature. However, most of these theorems have to make additional assumptions.

2.1.5 Flat structures

In this paragraph the flat structures on $\mathbb{R}^n_+$ are considered. By one of the invariance results above, these are exactly the α-divergences. Following *Amari*, the transformation $\alpha = 2q - 1$ is performed (this maps the *Csiszár* divergences to the *Tsallis* divergences). Given a discrete positive measure with coefficients m_i, the affine coordinates are defined as follows:

$$\theta^i \equiv h_\alpha(m_i) := m_i^{\frac{1-\alpha}{2}}. \tag{2.35}$$

It is not hard to see that the inverse function $\theta^i \mapsto m_i$ is convex (for $|\alpha| \leq 1$) and, hence, one can define a potential as follows:

$$\psi_\alpha(\theta) := \frac{1-\alpha}{2} \sum_{i=0}^{n} m_i = \frac{1-\alpha}{2} \sum_{i=0}^{n} \left(\theta^i\right)^{\frac{2}{1-\alpha}}. \tag{2.36}$$

The dual coordinates are given by

$$\eta_i \equiv h_{-\alpha(m_i)} = m_i^{\frac{1+\alpha}{2}}. \tag{2.37}$$

It should be noted that the normalization constraint $\sum_{i=0}^{n} m_i = 1$ that embeds Δ^n in $\mathbb{R}^n_+$ is a nonlinear constraint on the affine coordinates except for $\alpha = -1$ (for the dual coordinates this happens for $\alpha = 1$). This recovers the fact that the Kullback-Leibler divergence is the only flat, invariant and decomposable structure on Δ^n.

For any monotonic function h the so-called h**-representation** of x is defined as $h(x)$. Using this representation, one can define the h**-mean** as follows:

$$m_h(x, y) := h^{-1}\left(\frac{h(x) + h(y)}{2}\right). \tag{2.38}$$

The α-representations are exactly the ones inducing linearly scaling h-means:

$$m_\alpha(\lambda x, \lambda y) = \lambda m_\alpha(x, y). \tag{2.39}$$

It is not hard to see that all well-known means, such as the ordinary, geometric and harmonic means, are examples of α-means. Given an α-representation, one can define an α-mixture of distributions as the α-mean of the distributions (up to normalization). By allowing for weighted sums, the so-called α**-family** or α**-integration of distributions** with coordinate system $\{w_i\}_{1 \leq i \leq N}$ is obtained. The cases $\alpha = -1$ and $\alpha = 1$ can be seen to correspond to mixtures and exponential families, respectively.

2.2 Projections

The following theorem generalizes the classic Pythagorean theorem on $\mathbb{R}^n$ (and reduces to it when one chooses the Euclidean distance as the divergence measure)

Theorem 2.2.1 (Pythagoras). *Consider a triangle PQR on a dually flat manifold M with canonical divergence D. If the geodesic PQ and the dual geodesic QR are orthogonal, the following equation holds:*

$$D(P\|R) = D(P\|Q) + D(Q\|R). \tag{2.40}$$

One obtains a conjugate statement by dualizing all quantities.

In Euclidean space (and in general Hilbert spaces) one of the most powerful theorems is the projection theorem **??**, which states that the shortest path from a point to a subspace is given by orthogonal projection. This can be generalized to dually flat manifolds.

Definition 2.2.2 (Orthogonal projection). Consider a point $p \in M$ and a submanifold S of a dually flat manifold M such that $p \notin S$. A geodesic (orthogonal) projection of p on S is a point $p^* \in \partial S$ such that the affine geodesic connecting p and p^* is orthogonal to all of $T_{p^*}S$. One can obtain the notion of a dual geodesic projection in a simlar way.

Because in general there exist multiple projections, a strict minimality theorem cannot be formulated:

Theorem 2.2.3 (Projection theorem). *The extremal points of the map $s \mapsto D(p\|s)$ are geodesic projections of p onto S. The dual statement also holds.*

The strict projection theorem for Hilbert spaces only holds for affine subspaces. In the manifold setting this is reflected by a flatness condition:

Property 2.2.4. If the submanifold S is $\widetilde{\nabla}$-flat, the geodesic projection of $p \notin S$ is unique and it minimizes the map $s \mapsto D(p\|s)$. The dual statement also holds.

The e/m-terminology also exists for projections:

Definition 2.2.5 (Projections). The e- and m-projections are defined as follows:

- e-projection: $\pi_e(p) := \arg\min_{q \in S} D_{\mathrm{KL}}(q\|p)$, and
- m-projection: $\pi_m(p) := \arg\min_{q \in S} D_{\mathrm{KL}}(p\|q)$.

Chapter 3

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 1. For some sections the language of information geometry will be used as introduced in the previous chapter.

3.1 Data samples

3.1.1 Moment estimators

Formula 3.1.1 (r^{th} sample moment).

$$\overline{x^r} := \frac{1}{N}\sum_{i=1}^{N} x_i^r \tag{3.1}$$

Example 3.1.2 (Arithmetic mean). The arithmetic mean is used to average out differences between measurements. It is defined as the first sample moment:

$$\overline{x} := \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{3.2}$$

Theorem 3.1.3 (Weak law of large numbers[1]). *Assume that the sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is i.i.d. The sample average converges in probability 1.3.14 to the expectation value:*

$$\lim_{n\to\infty} \Pr\big(|\overline{X}_n - \mu| > \varepsilon\big) = 0 \tag{3.3}$$

for all $\varepsilon > 0$.

Theorem 3.1.4 (Strong law of large numbers[2]). *Assume that the sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is i.i.d. The sample average converges almost surely* **??** *to the expectation value:*

$$\Pr\Big(\lim_{n\to\infty} \overline{X}_n = \mu\Big) = 0. \tag{3.4}$$

In fact, the i.i.d. assumption can be weakened. The convergence

$$\overline{X}_n \xrightarrow{a.s.} \mathrm{E}\big[\overline{X}_n\big] \tag{3.5}$$

[1] Also called **Khinchin's law**.
[2] Also called **Kolmogorov's law**.

holds as long as the random variables are independent, have finite second moment and satisfy

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \mathrm{Var}[X_k] < +\infty. \tag{3.6}$$

Formula 3.1.5 (r^{th} central sample moment).

$$m_r := \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^r \tag{3.7}$$

Definition 3.1.6 (Weighted mean). Let $f : \mathbb{R} \to \mathbb{R}^+$ be a weight function. The weighted mean is given by:

$$\overline{x} := \frac{\sum_i f(x_i) x_i}{\sum_i f(x_i)}. \tag{3.8}$$

Example 3.1.7 (Binned mean). If the data has been grouped in bins, the weight function is given by the number of elements in each bin.

$$\overline{x} = \frac{1}{N} \sum_{i=1} n_i x_i. \tag{3.9}$$

Remark 3.1.8. In the above definitions, the measurements x_i can be replaced by function values $f(x_i)$ to calculate the mean of the function $f(x)$. This follows from Theorem 1.3.6. 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 3.1.9 (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 *normalized* measurements, i.e. ratios with respect to a reference value.

$$g := \left(\prod_{i=1}^{N} x_i \right)^{1/N} \tag{3.10}$$

The following relation exists between the arithmetic and geometic mean:

$$\ln g = \overline{\ln x}. \tag{3.11}$$

Definition 3.1.10 (Harmonic mean).

$$h := \left(\frac{1}{N} \sum_{i=1}^{N} x_i^{-1} \right)^{-1} \tag{3.12}$$

The following relation exists between the arithmetic and harmonic mean:

$$\frac{1}{h} = \overline{x^{-1}}. \tag{3.13}$$

Property 3.1.11. Let $\{x_i\}$ be a data set taking values in $\mathbb{R}_+$.

$$h \leq g \leq \overline{x} \tag{3.14}$$

The equalities only hold when all x_i are equal.

Definition 3.1.12 (Mode). The most occurring value in a data set.

Definition 3.1.13 (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.

3.1.2 Dispersion

Definition 3.1.14 (Range). The simplest indicator for statistical dispersion:

$$R := x_{\max} - x_{\min}. \tag{3.15}$$

However, it is very sensitive for outliers.

Definition 3.1.15 (Mean absolute difference).

$$\mathrm{MD} := \frac{1}{N}\sum_{i=1}^{N} |x_i - \overline{x}| \tag{3.16}$$

Definition 3.1.16 (Sample variance).

$$\mathrm{Var}(x) := \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{3.17}$$

Formula 3.1.17. The variance can also be rewritten in the following way:

$$\mathrm{Var}(x) = \overline{x^2} - \overline{x}^2. \tag{3.18}$$

Remark 3.1.18 (Bessel corection). A better estimator for the variance of a sample is given by the following formula:

$$\hat{s}^2 := \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{x})^2. \tag{3.19}$$

See Remark 3.4.9 for more information.

Definition 3.1.19 (Skewness). The skewness γ describes the asymmetry of a distribution. It is defined as the proportionality constant relating the third central moment and the standard deviation:

$$m_3 = \gamma\sigma^3. \tag{3.20}$$

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 3.1.20 (Pearson's mode skewness).

$$\gamma_P := \frac{\overline{x} - \mathrm{mode}}{\sigma} \tag{3.21}$$

Definition 3.1.21 (Kurtosis). The kurtosis c is an indicator for the ''tailedness''. It is defined as the proportionality constant relating the fourth central moment and the standard deviation:

$$m_4 = c\sigma^4. \tag{3.22}$$

Definition 3.1.22 (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 3.1.23 (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 3.1.24 (Interquartile range). The difference between the upper and lower quartile (75- and 25-percentiles respectively).

Definition 3.1.25 (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 3.1.26. For Gaussian distributions the following relation exists between the FWHM and the standard deviation:

$$\text{FWHM} = 2.35\sigma. \tag{3.23}$$

3.1.3 Multivariate data sets

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

Definition 3.1.27 (Covariance). The covariance of two data sequences is defined as follows:

$$\operatorname{cov}(x, y) := \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \overline{xy} - \overline{x}\ \overline{y}. \tag{3.24}$$

The covariance is also often denoted by σ_{xy} because of the next property:

Property 3.1.28. The covariance and standard deviation are related by the following equality:

$$\sigma_x^2 = \sigma_{xx}. \tag{3.25}$$

Definition 3.1.29 (Correlation coefficent).

$$\rho_{xy} := \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} \tag{3.26}$$

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

$$V_{ij} = \operatorname{cov}(x_{(i)}, x_{(j)}), \tag{3.27}$$

$$\rho_{ij} = \rho_{(i)(j)}. \tag{3.28}$$

3.2 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 3.2.1 (Percentile). The p-percentile c_p of a distribution F is defined as:

$$c_p = F^{-1}(p). \tag{3.29}$$

Definition 3.2.2 (Parametric family). A family of probability densities indexed by one or more parameters θ.

Example 3.2.3 (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 $\mathcal{P}$:

$$\left\{ \sum_{i=1}^{n} w_i P_i \;\middle|\; w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\}. \tag{3.30}$$

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

3.2.1 Empirical distribution

Definition 3.2.4 (Empirical distribution function). The (discrete) empirical probability distribution function is defined as the uniform mixture distribution with Dirac measures at the observations:

$$F_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}. \tag{3.31}$$

Theorem 3.2.5 (Borel's law of large numbers). *If the sample size approaches infinity, the observed frequencies approach the theoretical propabilities.*

Corollary 3.2.6 (Frequentist probability[3]).

$$\Pr(x) := \lim_{n\to\infty} \frac{f_n(x)}{n} \tag{3.32}$$

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

Theorem 3.2.7 (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*

$$\sup_{x\in\Omega} |F(x) - F_n(x)| \xrightarrow{a.s.} 0. \tag{3.33}$$

Remark 3.2.8. 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 3.2.9 (Kolmogorov-Smirnov statistic). Let F be a given cumulative distribution function. The n^{th} Kolmogorov-Smirnov statistic is defined as follows:

$$D_n := \sup_{x\in\Omega} |F_n(x) - F(x)|. \tag{3.34}$$

Definition 3.2.10 (Kolmogorov distribution).

$$F_{\mathrm{Kol}}(x) := 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty} e^{-(2i-1)^2\pi^2/(8x^2)} \tag{3.35}$$

Property 3.2.11 (Kolmogorov-Smirnov test). Let the null hypothesis H_0 state that a given data sample is described by a cumulative distribution function F. The null hypothesis is rejected at significance level α if

$$\sqrt{n}D_n > K_\alpha\,, \tag{3.36}$$

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

Definition 3.2.12 (Glivenko-Cantelli class). Consider a set of measurable functions $\mathcal{F}$ on a measurable space (Ω, Σ). For every probability measure P on Ω, one can define the $\mathcal{F}$-norm as follows:

$$\|P\|_{\mathcal{F}} := \sup\{\mathrm{E}_P[f] \mid f \in \mathcal{F}\}. \tag{3.37}$$

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

A class $\mathcal{F}$ of measurable functions is said to be Glivenko-Cantelli with respect to a probability measure P if it satisfies

$$\|P_n - P\|_{\mathcal{F}} \xrightarrow{\text{a.s.}} 0\,, \tag{3.38}$$

where P_n is the empirical measure.[4] The Glivenko-Cantelli theorem 3.2.7 says that the indicator functions of the sets $]-\infty, x]$ form a Glivenko-Cantelli class.[5] In fact, they are **universally GC** because this theorem applies to all probability measures on Ω. A class is said to be **uniformally GC** if the convergence holds uniformly over all probability measures.

Remark 3.2.13. Note that by the law of large numbers every singleton class is Glivenko-Cantelli (also universally and uniformly). The above definition strengthens the convergence of all elements of $\mathcal{F}$ to uniform convergence.

Property 3.2.14 (Bracketing number). Consider a collection of measurable functions $\mathcal{F}$ on a meaurable space (Ω, Σ) and recall Definition **??** of the bracketing number. If the bracketing number $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_1)$ is finite for all $\varepsilon > 0$, then $\mathcal{F}$ is Glivenko-Cantelli (with respect to the probability measure that induces the L^1-norm $\|\cdot\|_1$).

Property 3.2.15 (Metric entropy). Consider a collection of measurable functions $\mathcal{F}$ on a meaurable space (Ω, Σ) and recall Definition **??** of the metric entropy. Moreover, assume that $\mathcal{F}$ admits an integrable envelope F. Let $\mathcal{F}_M$ denote the collection of functions $f\mathbb{1}_{F\leq M}$, where $f \in \mathcal{F}$. If

$$\frac{1}{n} \ln N_C(\varepsilon, \mathcal{F}_M, \|\cdot\|_1) \xrightarrow{d} 0\,, \tag{3.39}$$

where $\|\cdot\|_1$ is the L^1-norm associated to the empirical measure P_n, for all $\varepsilon > 0$ and $M > 0$, then $\mathcal{F}$ is Glivenko-Cantelli.

Property 3.2.16 (Symmetrized empirical measure). Let $\{\sigma_1, \ldots, \sigma_n\}$ be a set of i.i.d. Rademacher variables 1.3.2. The symmetrized empirical measure is defined as follows:

$$P_n^\sigma : f \mapsto \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i). \tag{3.40}$$

Given a collection $\mathcal{F}$ of measurable functions on (Ω, Σ, P), the following inequality holds:

$$\mathrm{E}[\|P_n - P\|_{\mathcal{F}}] \leq 2\mathrm{E}[\|P_n^\sigma\|_{\mathcal{F}}]. \tag{3.41}$$

Theorem 3.2.17 (Donsker). *Consider an empirical distribution function F_n and define its normalized empirical process as*

$$G_n := \sqrt{n}(F_n - F)\,, \tag{3.42}$$

where F is the cumulative distribution function of the random variables X_i. The central limit theorem says that the empirical process converges in distribution to a standard normal distribution for every $x \in \mathbb{R}$. This can be strengthened as follows:

$$G_n \xrightarrow{d} U \tag{3.43}$$

in $D(\mathbb{R}, \|\cdot\|_\infty)$, where U is a standard Brownian bridge and $D(\mathbb{R}, \|\cdot\|_\infty)$ denotes the space of càdlàg functions equipped with the supremum metric **??**.

[4] If the convergence only holds in probability, the class is said to be **weakly GC**.

[5] Because every indicator function is uniquely associated to a set, one can also speak of GC classes of measurable sets.

3.2.2 Common distributions

Definition 3.2.18 (Uniform distribution).

$$f(x;a,b) := \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{elsewhere} \end{cases} \tag{3.44}$$

$$\mathrm{E}[x] = \frac{a+b}{2} \tag{3.45}$$

$$\mathrm{Var}[x] = \frac{(b-a)^2}{12} \tag{3.46}$$

Definition 3.2.19 (Gaussian distribution). Let σ be a positive number.

$$\mathcal{N}(x;\mu,\sigma) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{3.47}$$

This distribution is also called a (univariate) **normal distribution** with **mean** μ and **standard deviation** σ (or **variance** σ^2).

This formula can also be generalized to multivariate distributions. Let Σ be a positive-definite matrix, the **covariance** matrix. The associated multivariate normal distribution is given by

$$\mathcal{N}(\vec{\boldsymbol{x}};\vec{\boldsymbol{\mu}},\Sigma) = \frac{1}{\sqrt{2\pi}\det(\Sigma)} \exp\left(-\frac{(\vec{\boldsymbol{x}}-\vec{\boldsymbol{\mu}})^T\Sigma^{-1}(\vec{\boldsymbol{x}}-\vec{\boldsymbol{\mu}})}{2}\right). \tag{3.48}$$

Definition 3.2.20 (Standard normal distribution).

$$\mathcal{N}(z) := \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{3.49}$$

The cumulative distribution of $\mathcal{N}$ is called the **error function** $\mathrm{erf}(x)$.

Remark 3.2.21. 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 3.2.22 (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:*

$$\sqrt{n}\left(\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} \mathcal{N}(0,\sigma^2). \tag{3.50}$$

Remark 3.2.23. If the random variables are not independent, the CLT will not hold. 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.)

Formula 3.2.24. The sum of any number of (independent) Gaussian random variables is again Gaussian with the sum of the means and variances as parameters:

$$\forall i \in I : X_i \sim \mathcal{N}(\mu_i,\sigma_i^2) \implies \sum_{i\in I} X_i \sim \mathcal{N}\left(\sum_{i\in I}\mu_i, \sum_{i\in I}\sigma_i^2\right). \tag{3.51}$$

Definition 3.2.25 (Exponential distribution).

$$f(x;\tau) := \frac{1}{\tau} e^{-\frac{x}{\tau}} \tag{3.52}$$

$$\mathrm{E}[x] = \tau \tag{3.53}$$

$$\mathrm{Var}[x] = \tau^2. \tag{3.54}$$

Property 3.2.26. The exponential distribution is **memoryless**:

$$\Pr(X > x_1 + x_2 \mid X > x_2) = \Pr(X > x_1). \tag{3.55}$$

Definition 3.2.27 (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

$$p(k;\rho) := \rho^k (1-\rho)^{1-k} \tag{3.56}$$

$$\mathrm{E}[k] = \rho \tag{3.57}$$

$$\mathrm{Var}[k] = \rho(1-\rho). \tag{3.58}$$

Definition 3.2.28 (Binomial distribution). A process with n i.i.d. Bernoulli trials with probability ρ, is described by a binomial distribution:

$$\mathrm{Binom}(k;\rho,n) := \binom{n}{k} \rho^k (1-\rho)^{n-k} \tag{3.59}$$

$$\mathrm{E}[k] = n\rho \tag{3.60}$$

$$\mathrm{Var}[k] = n\rho(1-\rho). \tag{3.61}$$

Definition 3.2.29 (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.

$$\mathrm{Poisson}(r;\lambda) := \frac{e^{-\lambda}\lambda^r}{r!} \tag{3.62}$$

$$\mathrm{E}[r] = \mathrm{Var}[r] = \lambda. \tag{3.63}$$

Formula 3.2.30. If multiple independent Poisson processes occur simultaneously, the probability of r events is also described by a Poisson distribution:

$$\forall i \in I : X_i \sim \mathrm{Poisson}(\lambda_i) \implies \sum_{i\in I} X_i \sim \mathrm{Poisson}\left(\sum_{i\in I} \lambda_i\right). \tag{3.64}$$

The number of events coming from the process described by $\lambda_i | r = \sum_{i\in I} \lambda_i$ is given by a binomial distribution $\mathrm{Binom}(r_i; \Lambda_i, r)$ with $\Lambda_i = \frac{\lambda_i}{\sum_{i\in I}\lambda_i}$.

Remark 3.2.31. For $\lambda \longrightarrow \infty$, the Poisson distribution $\mathrm{Poisson}(r;\lambda)$ can be approximated by a Gaussian distribution $\mathcal{N}(x;\lambda,\sqrt{\lambda})$.

Theorem 3.2.32 (Raikov). *If the sum of two independent random variables is Poisson, the individual random variables are also Poisson.*

Definition 3.2.33 (χ^2-distribution). The sum of k squared independent (standard) normally distributed random variables Y_i defines the random variable:

$$\chi^2_k := \sum_{i=1}^{k} Y_i^2 \,, \tag{3.65}$$

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

$$f(\chi^2; n) := \frac{\chi^{n-2} e^{-\frac{\chi^2}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}. \tag{3.66}$$

Property 3.2.34. Due to the CLT 3.2.22, the χ^2-distribution approximates a Gaussian distribution for large n:

$$f(\chi^2; n) \overset{n>30}{\longrightarrow} \mathcal{N}(\sqrt{2\chi^2}; \sqrt{2n-1}, 1). \tag{3.67}$$

Definition 3.2.35 (Student-t distribution). The Student-t distribution describes the difference between the true mean and a sample average with estimated standard deviation $\hat{\sigma}$:

$$f(t; n) := \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\, \Gamma\left(\frac{n}{2}\right) \left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}}} \,, \tag{3.68}$$

where

$$t := \frac{(x-\mu)/\sigma}{\hat{\sigma}/\sigma} = \frac{z}{\sqrt{\chi^2/n}}. \tag{3.69}$$

Definition 3.2.36 (Cauchy distribution[6]). The general density $f(x; x_0, \gamma)$ is given by

$$f(x; x_0, \gamma) := \frac{1}{\pi} \frac{\gamma}{(x-x_0)^2 + \gamma^2}. \tag{3.70}$$

The associated characteristic function is given by

$$\mathrm{E}\left[e^{itx}\right] = e^{ix_0 t - \gamma|t|}. \tag{3.71}$$

Remark 3.2.37. Both the mean and variance of the Cauchy distribution are undefined.

3.3 Errors

Definition 3.3.1 (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 3.3.2 (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 3.1.2 in a meaningful way because the measurements are not of the same type. Therefore, it is also impossible to apply the CLT 3.2.22.

[6] Also known, especially in particle physics, as the **Breit-Wigner** distribution.

These problems can be resolved by the using the weighted mean 3.1.6:

$$\overline{x} := \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}. \tag{3.72}$$

The variation of the weighted mean is given by

$$\mathrm{Var}(\overline{x}) := \frac{1}{\sum_i \sigma_i^{-2}}. \tag{3.73}$$

Formula 3.3.3 (Error propagation). Let X be a vector random variable such that functions of X admit a good first-order Taylor approximation around the mean (this usually means that the covariance is small). The variance of a general function of X is given by

$$\mathrm{Var}[f(X)] \approx \sum_{i=1}^{n} \left(\frac{\partial f}{\partial X_i}\right)^2 \mathrm{Var}[X_i] + \sum_{i \neq j} \left(\frac{\partial f}{\partial X_i}\right)\left(\frac{\partial f}{\partial X_j}\right) \mathrm{cov}[X_i, X_j]. \tag{3.74}$$

Corollary 3.3.4. The correlation coefficient 3.1.29 of a random variable X and a **linear** function of X is independent of σ_x and is always equal to ± 1.

Definition 3.3.5 (Fractional error). Let X, Y be two independent random variables. The standard deviation of $f(X, Y) = XY$ is given by the fractional error:

$$\left(\frac{\sigma_f}{f}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2. \tag{3.75}$$

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

Property 3.3.6 (Logarithm). Let X be a random variable. The error of the logarithm of X is equal to the fractional error of X.

Formula 3.3.7 (Covariance of functions).

$$\mathrm{cov}[f, g] \approx \sum_{i,j} \left(\frac{\partial f}{\partial X_i}\right)\left(\frac{\partial g}{\partial X_j}\right) \mathrm{cov}[X_i, X_j] \tag{3.76}$$

Corollary 3.3.8. Let $f \equiv (f_1, \ldots, f_k)$ be a vector-valued function. The covariance matrix is to first order given by

$$\mathrm{Var}[f(X)] \approx J \mathrm{Var}[X] J^T, \tag{3.77}$$

where J is the Jacobian matrix of f.

3.4 Parameter estimation

3.4.1 General properties

Definition 3.4.1 (Consistency). An estimator $\hat{a}$ is said to be consistent if it is asymptotically equal to the true parameter:

$$\lim_{N \to \infty} \hat{a} = a. \tag{3.78}$$

Definition 3.4.2 (Unbiased estimator). An estimator $\hat{a}$ is said to be unbiased if its expectation value is equal to the true parameter:

$$\langle \hat{a} \rangle = a. \tag{3.79}$$

Note that neither consistency, nor unbiasedness implies the other.

Definition 3.4.3 (Bias).

$$B(\hat{a}) := |\langle \hat{a} \rangle - a|. \tag{3.80}$$

Definition 3.4.4 (Mean squared error).

$$\mathrm{MSE}(\hat{a}) := B(\hat{a})^2 + \mathrm{Var}(\hat{a}). \tag{3.81}$$

Remark 3.4.5. If an estimator is unbiased, the MSE is equal to the variance of the estimator.

3.4.2 Common estimators

Property 3.4.6 (Unbiased mean). The CLT 3.2.22 implies that the sample mean 3.1.2 is a consistent and unbiased estimator of the population mean.

Formula 3.4.7 (Standard error of the mean). Using the Bienaymé formula 1.4.27 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:

$$\mathrm{Var}[\overline{x}] = \frac{\sigma^2}{N}. \tag{3.82}$$

Formula 3.4.8 (Variance estimator for known mean). If the true mean μ is known, a consistent and unbiased estimator for the variance is given by

$$\widehat{\mathrm{Var}[X]} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{3.83}$$

Formula 3.4.9 (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

$$\hat{s}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2. \tag{3.84}$$

The modified factor $\frac{1}{N-1}$ is called the **Bessel correction**. It corrects the bias of the estimator given by the sample variance 3.1.16. The consistency is guaranteed by the CLT.

Property 3.4.10 (Characterization of normal distributions). The class of normal distributions is uniquely characterized by those distributions for which the sample mean and sample variance are independent.

3.4.3 Estimation error

Formula 3.4.11 (Variance of the estimator of the variance).

$$\mathrm{Var}\Big(\widehat{\mathrm{Var}[X]}\Big) = \frac{(N-1)^2}{N^3}\langle (x - \langle x \rangle)^4 \rangle - \frac{(N-1)(N-3)}{N^3}\langle (x - \langle x \rangle)^2 \rangle^2 \tag{3.85}$$

Formula 3.4.12 (Variance of the estimator of the standard deviation).

$$\mathrm{Var}(\widehat{\sigma}) = \frac{1}{4\sigma^2}\mathrm{Var}\Big(\widehat{\mathrm{Var}[X]}\Big) \tag{3.86}$$

Remark 3.4.13. 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.

3.4.4 Likelihood function

Definition 3.4.14 (Likelihood). The likelihood $\mathcal{L}(a; \mathbf{x})$ is the joint density of a set of measurements $\mathbf{x} := \{x_1, \ldots, x_N\}$:

$$\mathcal{L}(a; \mathbf{x}) = \prod_{i=1}^{N} f(x_i; a). \tag{3.87}$$

Theorem 3.4.15 (Cramer-Rao bound). *The variance of an **unbiased** estimator has a lower bound called the Cramer-Rao bound or **minimum variance bound (MVB)**:*

$$\mathrm{Var}(\hat{a}) \geq \frac{1}{\left\langle \left(\frac{\mathrm{d} \ln \mathcal{L}}{\mathrm{d}a}\right)^2 \right\rangle}. \tag{3.88}$$

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

$$\mathrm{Var}(\hat{a}) \geq \frac{\left(1 + \frac{\mathrm{d}b}{\mathrm{d}a}\right)^2}{\left\langle \left(\frac{\mathrm{d} \ln \mathcal{L}}{\mathrm{d}a}\right)^2 \right\rangle}. \tag{3.89}$$

Remark 3.4.16.

$$\left\langle \left(\frac{\mathrm{d} \ln \mathcal{L}}{\mathrm{d}a}\right)^2 \right\rangle = -\left\langle \frac{\mathrm{d}^2 \ln \mathcal{L}}{\mathrm{d}a^2} \right\rangle \tag{3.90}$$

Definition 3.4.17 (Fisher information).

$$I_X(a) := \left\langle \left(\frac{\mathrm{d} \ln \mathcal{L}}{\mathrm{d}a}\right)^2 \right\rangle = N \int \left(\frac{\mathrm{d} \ln f}{\mathrm{d}a}\right)^2 f \, d\mu \tag{3.91}$$

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

$$\mathrm{Var}(\hat{a}) \geq I_X(a). \tag{3.92}$$

Definition 3.4.18 (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:

$$e(\hat{a}) := \frac{I_X(a)^{-1}}{\mathrm{Var}(\hat{a})}. \tag{3.93}$$

3.4.5 Maximum likelihood estimation

From Definition 3.4.14 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 $f(a \mid x) = f(x \mid a)\frac{f(a)}{f(x)}$. 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 $f(a \mid x)$ and $f(x \mid a)$ 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 3.4.19 (Maximum likelihood estimator). The maximum likelihood estimator $\hat{a}$ is obtained by solving the following equation:

$$\left.\frac{\mathrm{d} \ln \mathcal{L}}{\mathrm{d}a}\right|_{a=\hat{a}} = 0. \tag{3.94}$$

Remark 3.4.20. MLE estimators are mostly consistent but often biased.

Property 3.4.21. MLE estimators are invariant under parameter transformations.

Corollary 3.4.22. The invariance implies that the two estimators $\hat{a}$ and $\widehat{f(a)}$ cannot both be unbiased at the same time.

Property 3.4.23. Every consistent estimator asymptotically becomes unbiased and efficient.

Property 3.4.24 (Minimizing KL-divergence). It can be shown that maximizing the log-likelihood is equivalent to minimizing the Kullback-Leibler divergence 1.7.3 between the would-be density $f(x;\theta)$ and the true density $q(x)$:

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

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

3.4.6 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 3.4.25 (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):

$$\chi^2 := \sum_i \frac{d_i^2}{f_i}\,, \tag{3.95}$$

   where $f_i = f(x_i;a)$.
3. Find the most probable value of $\hat{a}$ by solving the equation

$$\frac{\mathrm{d}\chi^2}{\mathrm{d}a} = 0. \tag{3.96}$$

Property 3.4.26. The optimal χ^2-value is asymptotically distributed according to a χ^2-distribution with n degrees of freedom. The parameter n is equal to the number of events N minus the number of fitted parameters k. (See more in Section 3.8.1.)

Formula 3.4.27 (Linear fit). When all uncertainties σ_i are equal, the slope $\hat{m}$ and intercept $\hat{c}$ are given by the following formulas:

$$\hat{m} := \frac{\overline{xy} - \overline{x}\;\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{Var}[x]} \tag{3.97}$$

$$\hat{c} := \overline{y} - \hat{m}\overline{x} = \frac{\overline{x^2} - \overline{x}\;\overline{y}}{\overline{x^2} - \overline{x}^2}. \tag{3.98}$$

Remark 3.4.28. 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 3.4.29 (Errors of linear fit).

$$\mathrm{Var}[\hat{m}] = \frac{1}{N(\overline{x^2} - \overline{x}^2)}\sigma^2 \tag{3.99}$$

$$\mathrm{Var}[\hat{c}] = \frac{\overline{x^2}}{N(\overline{x^2} - \overline{x}^2)}\sigma^2 \tag{3.100}$$

$$\mathrm{cov}(\hat{m}, \hat{c}) = \frac{-\overline{x}}{N(\overline{x^2} - \overline{x}^2)}\sigma^2 \tag{3.101}$$

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

Method 3.4.30 (Binned least squares).

1. N i.i.d. events with density $f(X; a)$ 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: $f_j = NW_j f(x_j; a)$.
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

$$\chi^2 := \sum_i^{N_B} \frac{(n_i - f_i)^2}{f_i^2}. \tag{3.102}$$

3.4.7 Geometric approach

Consider a sample $\mathbf{x} := \{x_1, \ldots, x_n\}$ drawn from a distribution $f(x; \theta)$ in an exponential family. The likelihood 3.4.14 is given by

$$\mathcal{L}(\theta; \mathbf{x}) := \prod_{i=1}^{n} f(x_i; \theta).$$

The m-coordinates of the observed point are

$$\eta = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}. \tag{3.103}$$

The optimal value for θ can be found by maximizing the log-likelihood as before. In Property 3.4.24 it was shown that this is equivalent to minimizing the Kullback-Leibler divergence between the ''true'' distribution $p(x; \xi)$ and the variational solution $f(x; \theta)$. However, in practice the true distribution is not known. Luckily one can replace the true distribution by the empirical distribution in the proof of 3.4.24. Minimization then corresponds to m-projecting the observed point η on the submanifold S of ''admissible'' distributions.

Theorem 3.4.31 (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* (3.2.4)*. Further, let*

Γ be a collection of probability distributions such that $P_n \in \Gamma$. The joint distribution q^n satisfies the following inequality:

$$q^n(P_n \in \Gamma) \leq (n+1)^{|S|} 2^{-nD_{\mathrm{KL}}(p^* \| q)}\,, \tag{3.104}$$

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

$$\lim_{n\to\infty} \frac{1}{n} q^n(P_n \in \Gamma) = -D_{\mathrm{KL}}(p^* \| q). \tag{3.105}$$

3.5 Bayesian modelling

Definition 3.5.1 (Conjugate distributions). Consider a prior distribution $F(\theta)$ and a posterior distribution $F(\theta \mid X)$. 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 $F(X \mid \theta)$.

Example 3.5.2. The simplest example is the case of binomial distributions, where the conjugate prior is the *β-distribution*. This can be generalized to multi-class situations. The conjugate prior of a categorical (or even *multinomial*) distribution is the *Dirichlet distribution*.

3.6 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 3.6.1 (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.

3.6.1 Interval types

Definition 3.6.2 (Two-sided confidence interval).

$$\Pr(x_- \leq X \leq x_+) = \int_{x_-}^{x_+} f(x)\,dx = C \tag{3.106}$$

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 3.6.3. For Gaussian distributions these three definitions are equivalent.

Definition 3.6.4 (One-sided confidence interval).

$$\Pr(x \geq x_-) = \int_{x_-}^{\infty} f(x)\,dx = C \tag{3.107}$$

$$\Pr(x \leq x_+) = \int_{-\infty}^{x_+} f(x)\,dx = C \tag{3.108}$$

Definition 3.6.5 (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 $[x_-, x_+]$.

$$x_- = \arg\min_\theta \left[\frac{1-C}{2} - \sum_{x=0}^{\theta-1} p(x) \right] \tag{3.109}$$

$$x_+ = \arg\min_\theta \left[\frac{1-C}{2} - \sum_{x=\theta+1}^{\infty} p(x) \right] \tag{3.110}$$

3.6.2 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 3.6.6. 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.

3.6.3 Interval for a sample mean

Formula 3.6.7 (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

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \right]. \tag{3.111}$$

Remark 3.6.8. If the sample size is not sufficiently large, the measured quantity must follow a normal distribution.

Formula 3.6.9 (Interval with unknown variance). To account for the uncertainty of the estimated standard deviation $\hat{\sigma}$, the student-t distribution 3.2.35 is used instead of a Gaussian distribution to describe the sample mean $\overline{X}$. The α-level confidence interval is given by

$$\left[\overline{X} - t_{\alpha/2;(n-1)}\frac{s}{\sqrt{N}}, \overline{X} + t_{\alpha/2;(n-1)}\frac{s}{\sqrt{N}} \right], \tag{3.112}$$

where s is the estimated standard deviation 3.4.9.

Formula 3.6.10 (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

$$\left[\frac{(2N\hat{P}+z_{\alpha/2}^2)-z_{\alpha/2}\sqrt{z_{\alpha/2}^2+4N\hat{P}(1-\hat{P})}}{2(N+z_{\alpha/2}^2)},\frac{(2N\hat{P}+z_{\alpha/2}^2)+z_{\alpha/2}\sqrt{z_{\alpha/2}^2+4N\hat{P}(1-\hat{P})}}{2(N+z_{\alpha/2}^2)}\right]. \tag{3.113}$$

Remark. The expectation value and variance are these of a binomial distribution 3.2.28 with $r = X/N$.

3.7 Hypothesis testing

Definition 3.7.1 (Simple hypothesis). A hypothesis where the distribution is fully specified.

Definition 3.7.2 (Composite hypothesis). A hypothesis where the distribution is given relative to some parameter values.

3.7.1 Testing

Definition 3.7.3 (Type I error). Rejection of a true null hypothesis.

Definition 3.7.4 (Type II error). Acceptance of a false null hypothesis.

Definition 3.7.5 (Significance). The probability of making a type-I error.

Property 3.7.6. Let $\alpha_1 > \alpha_2$. An α_2-level test is also significant at the α_1-level.

Remark 3.7.7. 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 3.7.8 (Power). The probability of not making a type-II error.

Remark 3.7.9. 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 3.7.10 (Shapiro-Wilk test). After obtaining a data sample it is often interesting to see if the data is distributed normally, since many other tests and methods assume a normal distribution. The Shapiro-Wilk test considers the following test statistic:

$$W := \frac{\sum_{i=1}^{n}(a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\,, \tag{3.114}$$

where

- $x_{(i)}$ are the order statistics,
- $(a_i,\ldots,a_n) := \frac{m^T V}{\|V^{-1}m\|}$,
- $m \equiv (m_1,\ldots,m_n)$ are the expectation values of the order statistics of i.i.d. standard normal distributions, and
- V is the covariance matrix of m.

The test statistic does not follow a known distribution and all critical values are calculated with Monte-Carlo simulations.

Definition 3.7.11 (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:

$$\Lambda(x) = \frac{\mathcal{L}(\theta_0 \mid x)}{\mathcal{L}(\theta_1 \mid x)} \leq \eta\,, \tag{3.115}$$

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

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

Theorem 3.7.12 (Neyman-Pearson lemma). *The likelihood ratio test is the most powerful test at significance level α.*

Definition 3.7.13 (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 3.7.14 (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:

$$\alpha \longrightarrow \frac{\alpha}{n}. \tag{3.116}$$

3.7.2 Comparison tests

Definition 3.7.15 (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

(3.117)

The null hypothesis of the McNemar test is that there is no significant difference between the predictive power of the models, 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 class i. In fact it is easy to see that the diagonal values are irrelevant for this hypothesis:

$$H_0 : b = c\,,$$
$$H_1 : b \neq c.$$

The test statistic is the McNemar chi-squared statistic:

$$\chi^2 = \frac{(b-c)^2}{b+c}. \tag{3.118}$$

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

$$\chi^2 := \frac{(|b-c|-1)^2}{b+c}. \tag{3.119}$$

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):

$$p = 2\sum_{i=b}^{b+c} \binom{b+c}{i} 0.5^i (1-0.5)^{b+c-i}. \tag{3.120}$$

Definition 3.7.17 (Wilcoxon signed-rank test). Consider a paired data sample, i.e. two dependent data samples for which the entries are uniquely paired. This test checks if the population means (more generally, the location parameters) are different.

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 the smallest of these:

$$T := \min(R_+, R_-). \tag{3.121}$$

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 3.7.18. The main 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.

Remark 3.7.19 (Independent samples). There exists a similar rank-based test for unpaired data samples. This is the **Wilcoxon rank-sum test** or **Mann-Whitney U-test**.

Definition 3.7.20 (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

$$\chi_F^2 := \frac{12N}{k(k+1)}\left(\sum_{i\leq k} R_i^2 - \frac{k(k+1)^2}{4}\right) \tag{3.122}$$

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

$$F := \frac{(N-1)\chi_F^2}{N(k+1) - \chi_F^2}. \tag{3.123}$$

This follows an F-distribution with $k-1$ and $(N-1)(k-1)$ degrees of freedom. 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).

3.7.3 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 3.7.22 (Nemenyi test). Consider the average ranks R_i from the Friedman test. As a test statistic one uses

$$z := \frac{R_i - R_j}{\sqrt{\frac{k(k+1)}{6N}}}, \tag{3.124}$$

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.

Remark 3.7.23 (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 3.7.24 (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 3.7.25 (Power). 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.

3.8 Goodness of fit

Definition 3.8.1 (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:

$$\mathrm{AIC} := 2k - 2\ln(\mathcal{L}_0). \tag{3.125}$$

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.

3.8.1 χ^2-test

Property 3.8.2. If there are $N-n$ fitted parameters one has:

$$\int_{\chi^2}^{\infty} f_{\chi^2}(x \mid n)\, dx \approx 1 \implies \begin{cases} \bullet \text{ good fit} \\ \bullet \text{ errors were overestimated} \\ \bullet \text{ selected measurements} \\ \bullet \text{ lucky shot} \end{cases} \tag{3.126}$$

Property 3.8.3 (Reduced χ^2). The reduced chi-squared statistic is defined as follows:

$$\chi^2_{\text{red}} := \chi^2/n, \tag{3.127}$$

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.

3.8.2 Runs test

A good χ^2-test does not mean that the fit is good. As mentioned in Property 3.8.2, 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 3.8.5.

Remark 3.8.4. 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 3.8.5 (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):

$$P(r_{\text{even}}) = 2\frac{\binom{N_+-1}{\frac{r}{2}-1}\binom{N_--1}{\frac{r}{2}-1}}{\binom{N}{N_+}} \qquad P(r_{\text{odd}}) = \frac{\binom{N_+-1}{\frac{r-3}{2}}\binom{N_--1}{\frac{r-1}{2}} + \binom{N_--1}{\frac{r-3}{2}}\binom{N_+-1}{\frac{r-1}{2}}}{\binom{N}{N_+}}, \tag{3.128}$$

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

$$\mathrm{E}[r] = 1 + 2\frac{N_+N_-}{N}, \tag{3.129}$$

$$\mathrm{Var}[r] = 2\frac{N_+N_-}{N}\frac{2N_+N_- - N}{N(N-1)}. \tag{3.130}$$

Remark 3.8.6. For $r > 15$, the runs distribution approximates a Gaussian distribution.

Chapter 4

Data Analysis

The main reference for the sections on optimization problems is [1]. For the geometry of clustering methods, see [2]. The main references for the section on *conformal prediction* are [3,4]. 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 **??**).

4.1 Data sampling

4.1.1 Inverse CDF sampling

Although one of the most straightforward sampling algorithms, this approach makes the strong assumption that the cumulative distribution function 1.3.4 is invertible.

Method 4.1.1. Sample a point λ uniformly from the unit interval $[0, 1]$. This number gives the cumulative distribution of the point to be sampled. (The CDF $F_X(X)$ is itself uniformly distributed.) The new point x' is simply given by $F_X^{-1}(\lambda)$.

In the case where F_X is discrete λ might not lie in the image of F_X and the inverse might not admit an algorithmically useful expression, so one should use a different approach. Given a point $x \in \mathbb{R}$ and its associated cumulative probability $F_X(x)$ one can sample a new point x' as follows. One increases (or decreases) x until the unique point x' is found such that $F_X(x' - 1) < \lambda \leq F_X(x')$.

4.1.2 Uniform rejection sampling

This method again uses the fact that the value of the cumulative distribution function is itself uniformly distributed on the unit interval $[0, 1]$. The CDF does not have to be invertible for this method, but the probability density should be compactly supported.

Method 4.1.2. Consider an interval $[a, b]$ such that f_X vanishes outside this interval and let q_0 be an upper bound for f_X. Now, sample a point x' uniformly on $[a, b]$ and sample a point q uniformly on $[0, q_0]$. If $f_X(x') \geq q$, then x' is a good sample. If not, repeat this procedure.

The proof of that this algorithm works is quite easy and mainly depends on the fact that

$$\Pr(Q \leq f_X(X)) = \frac{1}{q_0(b-a)}.$$

This value is often called the **acceptance probability**. From this expression it is clear that if one chooses q_0 too large, the acceptance probability becomes very small and the algorithm will take a long time to produce a sample.

4.1.3 Monte Carlo sampling

The general idea of (Markov chain) Monte Carlo methods is to construct a sequence of points such that (starting from a given index) all points represent good samples and such that the sequence forms a Markov chain.

The first Monte Carlo algorithm uses an acceptance threshold:

Method 4.1.3 (Metropolis-Hastings). Assume that a normal distribution $\mathcal{N}(\mu, \sigma^2)$ is given. The first element is given by $x_0 = \mu$. Subsequent points are constructed as follows:

- Sample a point x' from the normal distribution.
- Calculate the **acceptance ratio**

$$\lambda := \frac{f_X(x')}{f_X(x_{i-1})}.$$

- If $\lambda \geq 1$, $x_i = x'$.
- If not, sample a point q uniformly on $[0, 1]$. If $\lambda \geq q$, $x_i = x'$, else $x_i = x_{i-1}$.[1]

To obtain an efficient algorithm, it is helpful to choose $\mu = \mathrm{E}[X]$ and $\sigma^2 = \mathrm{Var}[X]$. This ensures that the points are sampled in a region that resembles the form of f_X.

In fact, this method can be drastically generalized. First of all it is possible to replace the normal distribution by any symmetric transition probability:

$$g(x' \mid x) = g(x \mid x'). \tag{4.1}$$

In case the transition probability is not symmetric, the acceptance ratio needs to be modified:

$$\lambda := \frac{f_X(x')g(x_{i-1} \mid x')}{f_X(x_{i-1})g(x' \mid x_{i-1})}. \tag{4.2}$$

Remark 4.1.4. It is clear from the definition of the acceptance ratio that one does not need f_X to be normalized. This avoids costly calculations of the normalization factor.

4.2 Optimization

4.2.1 Linear equations

Method 4.2.1 (Normal equation). Given the equation

$$Ax = b$$

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

$$\hat{x} := \arg\min_x (Ax - b)^T (Ax - b). \tag{4.3}$$

This leads to the so-called normal equation[2]

$$A^T Ax = A^T b. \tag{4.4}$$

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.

[1] This step could be merged with the previous one since $\lambda \geq 1$ always implies $\lambda \geq q$.

[2] 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 4.2.2. 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 4.2.3 (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 3, 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:

$$\|Ax - b\|^2 + \|\Gamma x\|^2 , \tag{4.5}$$

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

Remark 4.2.4. 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 4.2.5 (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

$$\lambda_0 + \lambda_1 X_1 + \cdots + \lambda_n X_n = 0. \tag{4.6}$$

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 $X^T X$ might have an ill-conditioned inverse which might render the algorithms unstable.

Definition 4.2.6 (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:

$$\mathrm{VIF}_i := \frac{1}{1 - R_i^2} , \tag{4.7}$$

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.

4.2.2 Gradient descent

The gradient descent algorithm is first introduced in the case of quadratic forms:

Method 4.2.7 (Steepest descent). Consider the quadratic form

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c.$$

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 $\nabla f(x_0) = Ax_0 - b =: -r_0$:

$$x_{i+1} := x_i + \alpha r_i. \tag{4.8}$$

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:

$$\alpha_i = \frac{r_i^T r_i}{r_i^T A r_i}. \tag{4.9}$$

This choice forces the descent direction to be orthogonal to the previous one since

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} f(x_i) = -\nabla f(x_i) \cdot \nabla f(x_{i-1}).$$

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.

4.2.3 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 4.2.8 (Conjugate vectors). Consider a symmetric positive-definite matrix A. Any such matrix induces an inner product as follows:

$$\langle v|w\rangle_A := v^T A w. \tag{4.10}$$

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 (4.10). 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:

$$r_{i+1} = r_i - \alpha_i A d_i \,, \tag{4.11}$$

where the step size α_i is defined similar to the step size for ordinary steepest descent:

$$\alpha_i = \frac{d_i^T r_i}{d_i^T A d_i}. \tag{4.12}$$

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 $r_{i+1} \in \mathcal{D}_i + A d_i$ leads to the following expression because of the above recursion relation:

$$\mathcal{D}_i = \operatorname{span}\big\{r_0, A r_0, \ldots, A^{i-1} r_0\big\}. \tag{4.13}$$

Because of their prominence in the literature on numeric optimization techniques, these subspaces have earned their own name:

Definition 4.2.9 (Krylov subspace). A vector space $\mathcal{K}$ of the form

$$\mathcal{K} := \operatorname{span}\{v, Av, \ldots, A^n v\} \tag{4.14}$$

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 4.2.10 (Conjugate gradient). Let x_0 be the initial guess with the associated residual $r_0 := b - A x_0$ 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 $Ax = b$:

$$\alpha_i := \frac{r_i^T r_i}{d_i^T A d_i} \tag{4.15}$$

$$x_{i+1} := x_i + \alpha_i d_i \tag{4.16}$$

$$r_{i+1} := r_i - \alpha_i A d_i \tag{4.17}$$

$$d_{i+1} := r_{i+1} + \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i} d_i. \tag{4.18}$$

Remark 4.2.11. 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 (4.17), but more importantly[3] 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 (4.4) using the same methods, since A^TA 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^TA will be nonsingular and the numerical methods can find an exact solution. However, the condition number of A^TA 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^TA, but the individual matrices are used A, A^T directly. This way not one Krylov space is generated, but two dual ''copies'' are constructed:

$$\mathcal{D}_i := \operatorname{span}\{r_0, Ar_0, \dots, A^{i-1}r_0\},$$
$$\widetilde{\mathcal{D}}_i := \operatorname{span}\{\widetilde{r}_0, A^T\widetilde{r}_0, \dots, (A^T)^{i-1}\widetilde{r}_0\},$$

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:

$$\langle r_i \mid r_j \rangle = \|r_i\|^2 \delta_{ij}. \tag{4.19}$$

As a consequence the search directions also form biconjugate bases:

$$\langle d_i \mid d_j \rangle_A = \|d_i\|_A^2 \delta_{ij}. \tag{4.20}$$

4.2.4 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 (4.18). 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 4.2.12 (Nonlinear CG). Since there is no linear equation related to the minimization problem, the residuals are always defined as $r_i := -\nabla f(x_i)$. The algorithm consists of the following iterations:

$$\alpha_i := \arg\min_{\alpha} f(x_i + \alpha d_i) \tag{4.21}$$
$$x_{i+1} := x_i + \alpha_i d_i \tag{4.22}$$
$$r_{i+1} := -\nabla f(x_i) \tag{4.23}$$
$$d_{i+1} := r_{i+1} + \beta_{i+1} d_i\,, \tag{4.24}$$

[3] 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 β_{i+1} is computed by one of the following formulas:

- **Fletcher-Reeves formula**:

$$\beta_{i+1} := \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i}. \tag{4.25}$$

- **Polak-Ribière formula**:

$$\beta_{i+1} := \max\left\{\frac{r_{i+1}^T (r_{i+1} - r_i)}{r_i^T r_i}, 0\right\}. \tag{4.26}$$

Some general remarks have to be made concerning the nonlinear CG algorithm:

Remark 4.2.13. 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.[4]

For linear CG a simple formula for finding the optimal value of α_i was obtained. However, for nonlinear CG one cannot solve Equation (4.21) 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:[5]

$$\alpha_i = \frac{\nabla f(x_i)^T d_i}{d_i^T \mathrm{Hess} f(x_i) d_i}. \tag{4.27}$$

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.

4.2.5 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

$$x_i = x_0 + \sum_{k=0}^{i-1} r_k = x_0 + \sum_{k=0}^{i-1} (\mathbb{1} - A)^k r_0. \tag{4.28}$$

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 $x_i - x_0 = \sum_{k=0}^{i-1} P_k(A) r_0$.

[4] The max operation in Equation (4.26) is already a form of restarting, due to the fact that the Polak-Ribière version of nonlinear CG sometimes results in cyclic behaviour.

[5] Another common method is the *secant method*.

Method 4.2.14 (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:

$$x_{i+1} := D^{-1}(b - Ex_i). \tag{4.29}$$

A sufficient condition for convergence is strict diagonal dominance, i.e. $|D_{ii}| > \sum_{j \neq i} |E_{ij}|$.

?? COMPLETE (e.g. Lanczos)??

4.3 Constrained optimization

4.3.1 Lagrange multipliers

A common generalization of the above optimization problems is the addition of constraints involving equalities:

$$\arg\min_x f(x) \qquad \text{such that} \qquad g_i(x) = 0 \qquad \forall 1 \leq i \leq n. \tag{4.30}$$

The general approach to solving such constrained problems is by extending the optimization loss:

Method 4.3.1 (Lagrange multipliers). Given a constrained optimization problem of the form (4.30), one can construct the enhanced loss function

$$\mathcal{L}(x, \lambda_1, \ldots, \lambda_n) := f(x) + \sum_{i=1}^{n} \lambda_i g_i(x). \tag{4.31}$$

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):

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial x} = 0 \\ \\ \dfrac{\partial \mathcal{L}}{\partial \lambda_i} = 0 \quad \forall 1 \leq i \leq n. \end{cases} \tag{4.32}$$

The situation becomes even more interesting when one also allows constraints involving inequalities:

$$\arg\min_x f(x) \qquad \text{such that} \qquad \begin{cases} g_i(x) = 0 & \forall 1 \leq i \leq m \\ h_j(x) \leq 0 & \forall 1 \leq j \leq n. \end{cases} \tag{4.33}$$

Problems of this form are called **primal optimization problems**. By defining an enhanced loss using Lagrange multipliers as before

$$\mathcal{L}(x, \alpha, \beta) := f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{i=1}^{n} \beta_i h_i(x)\,, \tag{4.34}$$

it is not hard to see that

$$\max_{\alpha, \beta; \beta_j \geq 0} \mathcal{L}(x, \alpha, \beta) = \begin{cases} \infty & \text{if a constraint is violated} \\ f(x) & \text{if all constraints are satisfied.} \end{cases} \tag{4.35}$$

Definition 4.3.2 (Primal optimization problem). Denote the maximum of $\mathcal{L}(x, \alpha, \beta)$ by $\theta_P(x)$.

$$p^* := \min_x \theta_P(x) = \min_x \max_{\alpha,\beta;\beta_i \geq 0} \mathcal{L}(x, \alpha, \beta). \tag{4.36}$$

By interchanging the max and min operators in the primal formulation, another problem is obtained:

Definition 4.3.3 (Dual optimization problem).

$$d^* := \max_{\alpha,\beta;\beta_i \geq 0} \theta_D(\alpha, \beta) = \max_{\alpha,\beta;\beta_i \geq 0} \min_x \mathcal{L}(x, \alpha, \beta). \tag{4.37}$$

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 4.3.4 (Slater conditions). Consider a convex optimization problem, i.e. a problem of the form (4.33) 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 4.3.5 (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 4.3.6 (Karush-Kuhn-Tucker conditions). If there exist x, α and β such that strong duality holds, the following conditions are satisfied:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial x} = 0 \\ \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \end{cases} \quad \forall 1 \leq i \leq m \qquad \text{and} \qquad \begin{cases} \beta_j h_j(x) = 0 \\ h_j(x) \leq 0 \\ \beta_j \geq 0 \end{cases} \quad \forall 1 \leq j \leq n. \tag{4.38}$$

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 4.3.7 (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 4.3.8. 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**.

4.3.2 Riemannian gradient descent

In many situations the full parameter space of an optimization problem is constrained in such a way that the resulting admissible subset admits the structure of a smooth manifold and, in particular that of a Riemannian manifold. When trying to extend gradient descent algorithms to this setting, one has to take into account that most manifolds are not linear spaces and, hence, that linear updates will often lead outside the manifold.

The first point that we have to treat is the occurence of the gradient in these algorithms. In ordinary Euclidean space, one simply takes the gradient to be the vector of partial derivatives. However, on general smooth manifolds this object is actually given by the de Rham differential, which is a covariant vector. However, a short proof shows that the Riemannian gradient from Remark **??** actually gives the direction of steepest ascent. So even on Riemannian manifolds the gradient is the correct direction to work with. However, as mentioned above, the form of the update will be a problem in general.

?? COMPLETE ??

4.4 Approximation theory

4.4.1 Bayes optimality

Definition 4.4.1 (Bayes risk). The minimal risk over all models:

$$R^* := \inf_{f:\mathcal{X}\to\mathcal{Y}} R(f). \tag{4.39}$$

Definition 4.4.2 (Bayes classifier). Given a joint probability distribution P over the instance space $\mathcal{X} \times \mathcal{Y}$ and a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the pointwise Bayes predictor is defined as follows:

$$f^* : x \mapsto \arg\min_{y\in\mathcal{Y}} \int_{\mathcal{Y}} l(y, y')\, dP(y' \mid x). \tag{4.40}$$

Property 4.4.3 (Bayes optimality). The risk of the pointwise Bayes predictor is minimal, i.e. $R(f^*) = R^*$. In practice, however, one cannot achieve Bayes optimality (through the pointwise Bayes predictor), since this would require the knowledge of the distribution.

Definition 4.4.4 (Approximation error). Given a data set $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}$, the empirical risk R_{emp} is defined as follows:

$$R_{\text{emp}} : \hat{y} \mapsto \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} l(\hat{y}(x), y). \tag{4.41}$$

Consider a hypothesis space $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ (e.g. selected by a choice of model architecture). The minimizer of the true risk in $\mathcal{H}$ is denoted by h^*, while the empirical risk minimizer is denoted by $\hat{h}$. The **approximation error/uncertainty** is defined as the difference $R(\hat{h}) - R(h^*)$, while the **model uncertainty** is defined as the difference $R(h^*) - R^*$. The total error made by $\hat{h}$ (with respect to the Bayes optimum) is then simply the sum of these two.

4.4.2 PAC theory and empirical risk minimization

Property 4.4.5. Note that by the (strong) law of large numbers 3.1.4, the empirical risk converges to the true risk almost surely whenever the data poitns are sampled i.i.d. However, in practice, the true question becomes how fast this convergence happens.

Definition 4.4.6 (Probably approximately correct). A model $\hat{f}$ is said to be ε-accurate at the confidence level $1 - \delta$ (with respect to the hypothesis space $\mathcal{H}$), or simply (ε, δ)-PAC, if it satisfies

$$\Pr\left(R(\hat{f}) - \inf_{f\in\mathcal{H}} R(f) > \varepsilon\right) < \delta. \tag{4.42}$$

Corollary 4.4.7. If one can prove a PAC bound, one gets the following bound for free:

$$R(\hat{f}) \leq \inf_{f \in \mathcal{H}} R(f) + 2\varepsilon \tag{4.43}$$

with probability $1 - \delta$, i.e. with high probability the true risk for the minimizer is only slightly higher than the empirical risk and, hence, the minimizer is a reasonbaly good estimate.

Property 4.4.8 (Bounded losses). Consider a hypothesis space $\mathcal{H}$ with a bounded loss function. The probability that for a given number of data points $n \in \mathbb{N}_0$ their exists at least one model $f \in \mathcal{H}$ for which the empirical risk deviates significantly from the true risk is bounded as follows:

$$\begin{aligned}
\Pr(\exists f \in \mathcal{H} : |R_{\text{emp}}(f) - R(f)| \geq \varepsilon) &= \Pr(\sup_{f \in \mathcal{H}} |R_{\text{emp}}(f) - R(f)| \geq \varepsilon) \\
&\leq \sum_{f \in \mathcal{H}} \Pr(|R_{\text{emp}}(f) - R(f)| \geq \varepsilon) \\
&\leq \sum_{f \in \mathcal{H}} 2e^{-2n\varepsilon^2} \\
&= 2|\mathcal{H}|e^{-2n\varepsilon^2}\,,
\end{aligned} \tag{4.44}$$

where the second inequality comes from Hoeffding's inequality 1.4.14 (the empirical risk is an average). In combination with a similar expression for the one-sided tails, using the one-sided Hoeffding inequality, the PAC bound becomes

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(1/\delta)}{2n}}\,, \tag{4.45}$$

with probability at least $1 - \delta$. Note that this bound is distribution-free, i.e. it does not depend on the data-generating distribution.

This bound on the empirical risk also implies a bound on the expected risk of the minimizer $\hat{f}$:

$$\mathrm{E}\Big[R(\hat{f})\Big] \leq \inf_{f \in \mathcal{H}} R(f) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(1/\delta)}{2n}} + \delta\,, \tag{4.46}$$

where $\delta > 0$ is now arbitrary.

The issue with the above two bounds is that they only apply to the case where the hypothesis space $\mathcal{H}$ is finite. To obtain useful bounds for countable hypothesis spaces, a new tool is required:

Definition 4.4.9 (Complexity regularizer). A function $c : X \to \mathbb{R}^+$ such that

$$\sum_{x \in X} e^{-c(x)} \leq 1. \tag{4.47}$$

A simple example would be the log-probabilities $c(x) := \ln(P(x))$ when a probability distribution P on X is given.

Using a complexity regularizer on $\mathcal{H}$, the following risk bound can be obtained with probability $1 - \delta$:

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{c(f) + \ln(1/\delta)}{2n}}. \tag{4.48}$$

4.5 Classification problems

4.5.1 Clustering

Probably the most well-known and simplest algorithm for clustering in the unsupervised setting is the k-means algorithm:

Method 4.5.1 (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:
$$i = \arg\min_{j \leq k} \|d - c_j\|^2. \tag{4.49}$$

2. Update the centroids c_i to represent the center of mass of the associated cluster C_i:
$$c_i \longleftarrow \frac{1}{|C_i|} \sum_{d \in C_i} d. \tag{4.50}$$

This algorithm optimizes the following global cost function with respect to the centroids c_i:

$$\mathcal{L}_{k\text{-means}}(c_1, \ldots, c_k) = \sum_{i=1}^{k} \sum_{d \in C_i} \|d - c_i\|^2. \tag{4.51}$$

Given the above idea, one could ask for a more general algorithm where clustering is performed with respect to a divergence function 2.1.7. In the case of Bregman divergences 2.1.10 it can be shown that all one needs to do is replace the Euclidean distance by the divergence D_f:

Property 4.5.2 (Centroid position). Let D_f be a Bregman divergence. The minimizer

$$\arg\min_{\kappa} \sum_{i=1}^{k} D_f(x_i \| \kappa) \tag{4.52}$$

is given by the arithmetic average

$$\kappa = \frac{1}{k} \sum_{i=1}^{k} x_i. \tag{4.53}$$

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 4.5.3 (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 geodesic hypersurface orthogonal to the dual geodesic connecting c_1 and c_2. This partitioning of the data manifold is a generalization of *Voronoi diagrams* to (Bregman) divergences.[6]

[6] See [5] for more information. This is also introduced in [6], but there the author has confusingly interchanged the affine and dual coordinates.

4.5.2 Nearest neighbour search

?? COMPLETE ??

4.6 Garden

?? ADD (e.g. trees, forests) ??

4.7 Support-vector machines

4.7.1 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 4.7.1 (Kernel[7]). 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 4.7.2 (Reproducing kernel Hilbert space). A Hilbert space $\mathcal{H} \subset \text{Map}(X, \mathbb{C})$ of functions over a 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 $\mathcal{H}$ as a function $K_x \in \mathcal{H}$. This allows for the introduction of a kernel on X:

Definition 4.7.3 (Reproducing kernel). Let $\mathcal{H}$ be an RKHS on a set X. The (reproducing) kernel k on X is defined as follows:

$$k(x, y) := \delta_x(K_y) \overset{\text{Riesz}}{=} \langle K_x \mid K_y \rangle_{\mathcal{H}}. \tag{4.54}$$

Because k is given by an inner product, it is not hard to see that the reproducing kernel is a kernel 4.7.1.

Starting from a kernel one can also characterize an RKHS as follows:

Alternative Definition 4.7.4 (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 $\delta_x(f) = \langle k(\cdot, x) \mid f \rangle_{\mathcal{H}}$.
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 4.7.5 (Convergence). In an RKHS, convergence in norm implies pointwise convergence.

Theorem 4.7.6 (Moore-Aronszajn). *There exists a bijection between RKHSs and kernels.*

[7] Also called a **Mercer kernel**. See Mercer's theorem below for more information.

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 the function $K_x := k(\cdot, x)$ for all $x \in X$. 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

$$\left\langle \sum_{x\in X} a_x K_x \middle| \sum_{y\in X} b_y K_y \right\rangle := \sum_{x,y\in X} \overline{a_x} b_y k(x,y). \tag{4.55}$$

Formula 4.7.7. 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

$$k(x,y) = \sum_{i=1}^{\dim(\mathcal{H})} e_i(x)\overline{e_i(y)}. \tag{4.56}$$

Remark 4.7.8. Note that one can use different conventions in the above definitions, e.g. the definition $k(x,y) = \langle K_y \mid K_x \rangle_{\mathcal{H}}$ is also valid.

Theorem 4.7.9 (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***

$$\iint_{X\times X} k(x,y)\overline{f(x)}f(y)\,dxdy \geq 0 \tag{4.57}$$

for all $f \in L^2(X,\mathbb{C})$, the Hilbert-Schmidt operator

$$T_k : L^2(X,\mathbb{C}) \to L^2(X,\mathbb{C}) : f \mapsto \int_X k(\cdot,x)f(x)\,dx \tag{4.58}$$

admits a countable orthonormal basis $\{e_i\}_{i\in\mathbb{N}}$ with nonnegative eigenvalues $\{\lambda_i\}_{i\in\mathbb{N}}$ such that

$$k(x,y) = \sum_{i=1}^{\infty} \lambda_i e_i(x)\overline{e_i(y)}. \tag{4.59}$$

Theorem 4.7.10 (Bochner). *A continuous function satisfies the Mercer condition if and only if it is a kernel.*

Alternative Definition 4.7.11 (Kernel). Consider a set X. A function $k : X \times X \to \mathbb{C}$ is called a (Mercer) kernel on X if there exists a Hilbert space $\mathcal{H}$ together with a function $\phi : X \to \mathcal{H}$ such that

$$k(x,y) = \langle \phi(x) \mid \phi(y) \rangle_{\mathcal{H}}. \tag{4.60}$$

When using Mercer's theorem, the feature maps are given by

$$\phi_i : x \mapsto \sqrt{\lambda_i} e_i(x). \tag{4.61}$$

Remark 4.7.12. The kernel expressions in the Mercer and Moore-Aronszajn theorems are related by the fact that the RKHSs induced by kernels satisfying the assumptions of the Mercer theorem are of the form

$$\mathcal{H} = \left\{ f \in L^2(X,\mathbb{C}) \middle| \sum_{i=1}^{\infty} \frac{\langle f, e_i\rangle_{L^2}^2}{\lambda_i} < +\infty \right\}. \tag{4.62}$$

Remark 4.7.13 (Vector-valued functions). Much of this section can be generalized to the setting of vector-valued functions $f : X \to \mathbb{C}^d$. In this case the kernels $k : X \times X \to \mathbb{C}$ are generalized to a matrix-valued functions $k : X \times X \to \mathbb{C}^{d\times d}$.

4.7.2 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

$$\mathrm{sgn}(y) = \begin{cases} 1 & y > 0 \\ -1 & y < 0 \end{cases} \tag{4.63}$$

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

$$d(x) = \frac{w^T x + b}{\|w\|}. \tag{4.64}$$

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

$$\max_w \frac{\gamma}{\|w\|} \qquad \text{such that} \qquad y_i(w^T x_i + b) \geq \gamma \|w\| \qquad \forall 1 \leq i \leq n\,, \tag{4.65}$$

where $\gamma = \min_{i \in \{1,\ldots,n\}} \gamma(x_i)$ 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:

$$\min_w \|w\|^2 \qquad \text{such that} \qquad y_i(w^T x_i + b) \geq 1 \qquad \forall 1 \leq i \leq n. \tag{4.66}$$

The KKT conditions for this problem give the following results:

$$w = \sum_{i=1}^{n} \beta_i y_i x_i \tag{4.67}$$

and

$$\sum_{i=1}^{n} \beta_i y_i = 0\,, \tag{4.68}$$

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:

$$y \equiv w^T x + b = \sum_{i=1}^{n} \beta_i y_i \langle x_i \mid x \rangle + b. \tag{4.69}$$

Two observations can be made at this point. First of all, complementary slackness 4.3.7 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:

$$\min_w \|w\|^2 + C\sum_{i=1}^{n} \xi \qquad \text{such that} \qquad \begin{cases} \xi_i \geq 0 & \forall 1 \leq i \leq n \\ y_i(w^T x_i + b) \geq 1 - \xi_i & \forall 1 \leq i \leq n. \end{cases} \tag{4.70}$$

The resulting KKT conditions are as follows:

$$0 \leq \beta_i \leq C \tag{4.71}$$

and

$$\beta_i = 0 \implies y_i(w^T x_i + b) \geq 1 \tag{4.72}$$

$$\beta_i = C \implies y_i(w^T x_i + b) \leq 1 \tag{4.73}$$

$$\beta_i \in]0, C[\implies y_i(w^T x_i + b) = 1. \tag{4.74}$$

?? COMPLETE (e.g. geometry)??

4.8 Vapnik-Chervonenkis theory

4.8.1 VC dimension

Definition 4.8.1 (Shatter coefficient). Let Ω denote the universe of discours and consider a set $C \subset P(\Omega)$. C **shatters** a set $A \subset \Omega$ if for every subset $a \subseteq A$ there exists a subset $c \in C$ such that

$$a = c \cap A. \tag{4.75}$$

The shatter(ing) coefficients of C are defined as follows:

$$S_n(C) := \max_{x_1,\ldots,x_n \in \Omega} \big|\big\{\{x_1, \ldots, x_n\} \cap c \,\big|\, c \in C\big\}\big|. \tag{4.76}$$

It should be clear that every shatter coefficient S_n is bounded above by 2^n. If $S_n(C) = 2^n$, then C shatters some set of cardinality n.

Given a collection of binary functions $\mathcal{F} \subseteq \{0,1\}^X$, the shatter(ing) coefficients (or **growth functions**) of $\mathcal{F}$ are given by:

$$S_n(\mathcal{F}) := \max_{x_1,\ldots,x_n \in X} \big|\big\{\{f(x_1), \ldots, f(x_n)\} \,\big|\, f \in \mathcal{F}\big\}\big|. \tag{4.77}$$

The shatter coefficients give a notion of the effective size of $\mathcal{F}$, since they say how many different results the functions can produce given a data set of n points.

Definition 4.8.2 (Vapnik-Chervonenkis dimension). The Vapnik-Chervonenkis dimension of a collection of functions $\mathcal{F}$ is defined as follows:

$$\mathrm{VC}(\mathcal{F}) := \max\{k \in \mathbb{N} \mid S_k(\mathcal{F}) = 2^k\}. \tag{4.78}$$

Note that if a collection shatters n points, it also necessarily shatters a subset of these points, therefore, one also has

$$S_n(\mathcal{F}) = 2^n \tag{4.79}$$

for all $n < \mathrm{VC}(\mathcal{F})$.

A collection is called a **Vapnik-Chervonenkis class** if its VC dimension is finite.

Property 4.8.3 (Sauer's lemma[8]). The VC dimension bounds shatter coefficients in the following way:

$$S_n(\mathcal{F}) \leq \sum_{k=0}^{\mathrm{VC}(\mathcal{F})} \binom{n}{k} \tag{4.80}$$

for all $n \in \mathbb{N}$. For $n \leq \mathrm{VC}(\mathcal{F})$, the right-hand side is just the full binomial expansion for 2^n and, accordingly, the shatter coefficients grow exponentially. For $n \geq d$, the binomial series is truncated and polynomial behaviour is obtained:

$$\forall n \geq \mathrm{VC}(\mathcal{F}) : S_n(\mathcal{F}) \leq \left(\frac{e\,n}{\mathrm{VC}(\mathcal{F})}\right)^{\mathrm{VC}(\mathcal{F})}, \tag{4.81}$$

where e is the Euler number.

The generalization bounds of Section 4.4.2 on empirical risk minimization can be extended to uncountable hypothesis spaces as follows:

Property 4.8.4 (Generalization bound). The expected risk of the empirical risk minimizer $\hat{f}$ is bounded as follows:

$$R(\hat{f}) \leq \inf_{f \in \mathcal{H}} R(f) + 4\sqrt{2\frac{\mathrm{VC}(\mathcal{H})\ln(e\,n/\mathrm{VC}(\mathcal{H})) + \ln(2/\delta)}{n}} \tag{4.82}$$

with probability at least $1 - \delta$.

?? CHECK ALL THESE BOUNDS ??

4.8.2 Rademacher complexity

Remark 4.8.5 (Real-valued functions). The VC dimension of a collection of arbitrary real-valued functions can be defined as the VC dimension of the corresponding collection of indicator functions (Heaviside functions):

$$\mathrm{VC}(\mathcal{F}) := \mathrm{VC}\big(\{\theta(f(x) - \lambda) \,\big|\, f \in \mathcal{F}, \lambda \in \mathbb{R}\}\big). \tag{4.83}$$

This definition is equivalent to the following one based on subgraphs:

$$\mathrm{VC}(\mathcal{F}) = \mathrm{VC}\big(\{C_f := \{(x, \lambda) \in X \times \mathbb{R} \mid \lambda < f(x)\} \,\big|\, f \in \mathcal{F}\}\big). \tag{4.84}$$

[8] Sometimes called the **Sauer-Shelah lemma**.

Example 4.8.6 (Linear spaces). Every vector space V of real-valued functions has VC dimension at most $\dim(V) + 1$.

Example 4.8.7 (Translations). The set of translations of a real-valued function has VC dimension 1.

Although this remark says that one can in theory extend ordinary VC theory to arbitrary (real-valued) functions, this does not mean that the VC bounds obtained before above make sense in this setting. To obtain useful bounds, it is important to introduce a new notion of effective size:

Definition 4.8.8 (Rademacher complexity). Consider a collection of functions $\mathcal{F} := \{f : X \to \mathbb{R}\}$. The Rademacher complexity is defined as follows:

$$\mathfrak{R}_n(\mathcal{F}) := \mathrm{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i)\right], \tag{4.85}$$

where the σ_i are Rademacher variables 1.3.2 and the expectation is taken over both the Rademacher variables and the sample $(X_1, \ldots, X_n)$.

Property 4.8.9 (Risk bound). Consider a collection of bounded functions $\mathcal{F} \subseteq [a, b]^{\mathcal{X}}$.

$$\mathrm{E}[R(f)] \leq R_{\mathrm{emp}}(f) + 2\mathfrak{R}_n(\mathcal{F}) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}} \tag{4.86}$$

with probability at least $1 - \delta$.

Property 4.8.10 (VC dimension). The shatter coefficient bounds the Rademacher complexity in the following way:

$$\mathfrak{R}_n(\mathcal{F}) \leq \sqrt{\frac{2\ln(S_n(\mathcal{F}))}{n}}. \tag{4.87}$$

4.8.3 Relation to Glivenko-Cantelli classes

Property 4.8.11. Recall Property 3.2.15. The empirical L^1-norm only depends on the values of the functions $f \in \mathcal{F}$ at the given data points and, therefore, the covering number of $\mathcal{F}$ is bounded above by the covering number of $\{(f(x_1), \ldots, f(x_n)) \mid f \in \mathcal{F}\}$. The latter is itself bounded above by the shatter coefficient $S_n(\mathcal{F})$. If $\mathcal{F}$ has finite VC dimension, Sauer's lemma 4.8.3 implies that this coefficient grows polynomial in n, so

$$\frac{1}{n}\ln N_C(\varepsilon, \mathcal{F}_M, \|\cdot\|_1) \xrightarrow{d} 0 \tag{4.88}$$

and, thus, $\mathcal{F}$ is Glivenko-Cantelli.

Theorem 4.8.12. *A class of sets is Vapnik-Chervonenkis if and only if it is Glivenko-Cantelli 3.2.12.*

4.9 Time series analysis

Definition 4.9.1 (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.

4.9.1 Stationarity

Definition 4.9.2 (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:

$$P(X_{t_1}, \ldots, X_{t_r}) = P(X_{t_1+s}, \ldots, X_{t_r+s}). \tag{4.89}$$

Definition 4.9.3 (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 **??**:

Definition 4.9.4 (Ergodicity). A time series $\{X_t\}_{t\in\mathbb{Z}}$ is ergodic if for every measurable function f the following equation holds for all $t \in \mathbb{Z}$:

$$\lim_{T\to\infty} \frac{1}{2T+1} \sum_{k=-T}^{T} f(X_k) = \mathrm{E}[f(X_t)]. \tag{4.90}$$

Intuitively this means that state space averages can be evaluated as time averages.

4.9.2 Correlation

Definition 4.9.5 (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 4.9.6 (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:

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{i\omega k}, \tag{4.91}$$

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 4.9.7 (Lag operator[9]). The lag operator sends a variable in a time series to the preceding value:

$$BX_t = X_{t-1}. \tag{4.92}$$

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):

$$\theta(B) = 1 + \sum_{i=1}^{k} \theta_i B^i\,, \tag{4.93}$$

$$\varphi(B) = 1 - \sum_{i=1}^{k} \varphi_i B^i. \tag{4.94}$$

[9] Also called the **backshift operator**.

Notation 4.9.8 (Difference operator). The difference operator Δ is defined as follows:

$$\Delta = 1 - B. \tag{4.95}$$

In a similar way one can define the **seasonal** difference operator:

$$\Delta_s = 1 - B^s. \tag{4.96}$$

Method 4.9.9 (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

$$Q = n(n+2)\sum_{i=1}^{k}\frac{\rho_k}{n-k}. \tag{4.97}$$

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.

Method 4.9.10 (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):

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \sum_{i=1}^{p-1}\theta_i \Delta X_{t-i} + \varepsilon_t. \tag{4.98}$$

The test statistic is

$$\mathrm{DF} := \frac{\gamma}{\mathrm{SE}(\gamma)}\,, \tag{4.99}$$

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).

4.9.3 Autoregressive models

Definition 4.9.11 (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 $X_{t-1},\ldots,X_{t-p}$ of the time series:

$$X_t = \beta_0 + \beta_1 X_{t-1} + \cdots + \beta_p X_{t-p} + \varepsilon_t. \tag{4.100}$$

Definition 4.9.12 (Partial autocorrelation function). The p^{th} autocorrelation function is defined as the p^{th} coefficient in the AR(p)-model.

Remark 4.9.13. The optimal order p of an autoregressive model is the one for which all higher partial autocorrelation functions (almost) vanish.

Definition 4.9.14 (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 $\varepsilon_{t-1},\ldots,\varepsilon_{t-p}$ of the error term:

$$X_t = \beta_0 + \beta_1 \varepsilon_{t-1} + \cdots + \beta_p \varepsilon_{t-p} + \varepsilon_t. \tag{4.101}$$

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 4.9.15. The optimal order p of an autoregressive model is the one for which all higher autocorrelation functions (almost) vanish.

Definition 4.9.16 (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[10] 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:

Definition 4.9.17 (ARMA(p, q)**-model).**

$$X_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} + \varepsilon_t \tag{4.102}$$

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:

$$\varphi(B)X_t = \alpha_0 + \theta(B)\varepsilon_t. \tag{4.103}$$

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:

$$\varphi(B)(1 - B)^d X_t = \alpha_0 + \theta(B)\varepsilon_t. \tag{4.104}$$

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[11].

Remark 4.9.18. Including so-called *exogenous* variables, i.e. external predictors, leads to an **ARIMAX**-model.

Remark 4.9.19 (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 [7]). One of the most famous ones is the method by *Durbin*:

[10] Sometimes this is used as a definition of invertibility.

[11] The 'I' stands for "integrated".

Method 4.9.20 (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 4.9.21 (ARIMA$(p,q,d)(P,Q,D)_s$**-model).**

$$\Phi(B^s)\varphi(B)\Delta_s^D\Delta^d X_t = \theta(B)\Theta(B^s)\varepsilon_t \tag{4.105}$$

4.9.4 Causality

Definition 4.9.22 (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:

$$P[Y_{t+k} \in A \mid \Omega(t)] \neq P[Y_{t+k} \in A \mid \Omega\backslash X(t)] \tag{4.106}$$

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 4.9.23. A slightly different but for computational purposes often more useful[12] 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 4.9.24. 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).

4.10 Uncertainty modelling

4.10.1 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.

[12] In fact this was the original definition by *Granger*.

Definition 4.10.1 (Validity). Consider a measurable function $\Gamma : X \to P(Y)$ and let P be the joint distribution on the instance space $Z \equiv X \times Y$. Γ is said to be valid (at **significance level** $\alpha \in [0, 1]$ or **confidence level** $1 - \alpha$) if it satisfies

$$P(y \in \Gamma^\alpha(x)) \geq 1 - \alpha. \tag{4.107}$$

One sometimes also distinguishes between exact validity and **conservative** validity, where the former is the subcase of the latter for which the inequality becomes an equality.

In fact, one can define two notions of validity: pointwise and asymptotic. Equation (4.107) characterizes pointwise validity in the sense that the probability of having an error is given by a Bernoulli process with parameter α. Asymptotic validity is a frequentist notion in the following sense:

$$\lim_{n\to\infty} \frac{\mathrm{Err}_n(\Gamma)}{n} \leq \alpha\,, \tag{4.108}$$

where $\mathrm{Err}_n(\Gamma)$ is the number of errors made by Γ after n trials. It should be clear that pointwise validity (both exact and conservative) implies asymptotic validity.

Remark 4.10.2 (Confidence regions). The definition of valid confidence predictors above is similar to the definition of confidence regions (Section 3.6). However, in contrast to confidence regions for population parameters, the size of confidence regions for predictive distributions does not go towards 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.[13]

4.10.2 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 4.10.3 (Exchangeability). Consider an ordered data sample $\{z_i\}_{1\leq i\leq N}$. The joint distribution $P(z_1, \ldots, z_N)$ 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) multiset $\mathcal{B}$ 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 4.10.4 (Nonconformity measure). Consider a bag $\mathcal{B}$ together with a new element z^* in a probability space (Z, Σ, P). A noncomformity measure $A : Z^\infty \times Z \to \mathbb{R}$ is a measurable function that gives a number indicating how different z^* is from the content of $\mathcal{B}$.

[13] Note that when observations are not sampled according to a distribution, but are perfectly predictable, this is of course not true.

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 4.10.5 (Point predictors). A general class of nonconformity measures is obtained from point predictors for a metric space (Y, d) . Given a point predictor $\rho : X \to Y$ trained on a bag $\mathcal{B}$, one can define a nonconformity measure as follows:

$$A_\rho\big(\mathcal{B}, (x, y)\big) := d\big(\rho(x), y\big). \tag{4.109}$$

Example 4.10.6 (Interval predictors). For every interval predictor $(l, u) : X \to \mathbb{R}^2$, i.e. a pair of functions $u, l : X \to \mathbb{R}$ such that $\forall x \in X : l(x) \leq u(x)$, trained on a bag $\mathcal{B}$, one can define a nonconformity measure as follows:

$$A_{u,l}\big(\mathcal{B}, (x, y)\big) := \max\big(l(x) - y, y - u(x)\big). \tag{4.110}$$

It should be noted that although common, nonconformity measures and, by extension, conformal prediction are also applicable to nonmetric spaces:

Example 4.10.7 (Nested region predictors). Let T be a totally ordered set **??**. For every model $(C_t)_{t\in T} : X \to P(Y)$ that predicts a sequence of nested regions, i.e.

$$s \leq t \implies C_s(x) \subseteq C_t(x)\,, \tag{4.111}$$

trained on a bag $\mathcal{B}$, one can define a nonconformity measure as follows:

$$A_T\big(\mathcal{B}, (x, y)\big) := \inf\{t \in T \mid y \in C_t(x)\}. \tag{4.112}$$

Construction 4.10.8 (Conformal predictor). Consider a data sample given as a bag $\mathcal{B} \in Z^\infty$ 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 ''nonconforming'' elements:[14]

$$p^* := \frac{|\{z \in \mathcal{B} \mid \mu_z \geq \mu_{z^*}\}| + 1}{|\mathcal{B}| + 1}. \tag{4.113}$$

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 4.10.9 (Optimality). A conformal predictor satisfies the following conditions:

[14]The reason for the $+1$ in this formula is that one considers the extended data set $\mathcal{B} \cup \{z^*\}$.

- Regions are (conservatively) valid, i.e.

$$P(p^* \leq \alpha) \leq \alpha\,, \tag{4.114}$$

  and thus also

$$P\big(y^* \in \Gamma^\alpha(x^*, \mathcal{B})\big) \geq 1 - \alpha\,, \tag{4.115}$$

  where the probability is taken over both the bag $\mathcal{B}$ and the point (x^*, y^*).

- 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 4.10.10 (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 (4.113) by

$$p^*(\theta) := \frac{|\{z \in \mathcal{B} \mid \mu_z > \mu_{z^*}\}| + \theta\,|\{z \in \mathcal{B} \cup \{z^*\} \mid \mu_z = \mu_{z^*}\}|}{|\mathcal{B}| + 1}\,, \tag{4.116}$$

where θ is independently and uniformly sampled from the unit interval $[0, 1]$. Exact validity is obtained by also marginalizing over 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 [8]. 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:

$$p^* = \frac{1}{|S_{N+1}|} \sum_{\sigma \in S_{N+1}} \mathbb{1}\big(A(\sigma \cdot \vec{\boldsymbol{z}}) \geq A(\vec{\boldsymbol{z}})\big)\,, \tag{4.117}$$

where $\vec{\boldsymbol{z}} := (z_1, \ldots, z_N, z^*)$. 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 (is this even relevant for this compendium) ??

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 4.10.11 (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 ''nonconforming'' elements:

$$p^* := \frac{|\{z \in \mathcal{C} \mid \mu_z \geq \mu_{z^*}\}| + 1}{|\mathcal{C}| + 1}. \tag{4.118}$$

As in the original CP algorithm, a new observation z^* is included in the prediction region if and only if $p^* > \alpha$:

$$\Gamma^\alpha(x) := \{y \in Y \mid p(x,y) > \alpha\}. \tag{4.119}$$

It should be clear that the underlying predictor only needs to be trained once using this scheme.

Remark 4.10.12 (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 4.10.13 (Validity). Although the above ICP algorithm is already computationally much more efficient than its transductive counterpart, one can go even further. However, in this case one needs to pay attention in order not to ruin the validity. To use Equation (4.118) one does not have to retrain the model every time, but one still needs to reevaluate the nonconformity score for possible $y^* \in Y$. It would be even better if one could extract the boundaries of the prediction region straight from the calibration data.

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 sampled from $\{1, \ldots, |\mathcal{C}|+1\}$. So, given a quantile level β, the probability of finding a new nonconformity score smaller than or equal to the β-quantile of $A(\mathcal{C})$ is

$$P\big(\mu_{z^*} \leq q_\beta(A(\mathcal{C}))\big) = \frac{\lceil \beta|\mathcal{C}| \rceil}{|\mathcal{C}|+1}. \tag{4.120}$$

If one constructs a prediction region by including all points $(x^*, y^*) \in Z$ such that the associated nonconformity score is smaller than the $(1-\alpha)$-quantile, one immediately obtains

$$P\big(y^* \in \Gamma^\alpha(x^*, \mathcal{C})\big) = \frac{\lceil (1-\alpha)|\mathcal{C}| \rceil}{|\mathcal{C}|+1}. \tag{4.121}$$

However, this is where the TCP and ICP algorithms differ. When using Equation (4.113) 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 this quantile-based reformulation of the inductive algorithm, a minimal modification also leads to validity:

$$\Gamma^\alpha(x) \longrightarrow \Gamma^\alpha(x) := \left\{y \in Y \mid A(x,y) \leq q_{(1+1/|\mathcal{C}|)(1-\alpha)}(A(\mathcal{C}))\right\}, \tag{4.122}$$

i.e. the empirical quantiles are replaced by ''inflated'' quantiles. When using Equations (4.118) and (4.119) to determine the ICP region, one is essentially including all points (x^*, y^*) such that the nonconformity score is smaller than or equal to the $(1-\alpha)$-quantile of the enhanced calibration curve $A(\mathcal{C}) \cup \{A(x^*, y^*)\}$. It is not hard to show that this quantile is equivalent to the inflated quantile of the ordinary calibration curve.

The smoothing of Property 4.10.10 can also be applied 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:

$$P\big(y^* \in \Gamma^\alpha(x, \mathcal{C})\big) \geq 1 - \alpha + \frac{1}{|\mathcal{C}|}. \tag{4.123}$$

So, in the limit of large calibration sets, the exact validity is recovered.

Remark 4.10.14. 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

$$\frac{|\{1 \leq i \leq k \mid y_i \in \Gamma^\alpha(x_i, \mathcal{C})\}|}{k} \tag{4.124}$$

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 4.10.15 (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 $\mathcal{S} := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ 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:

$$\Gamma^\alpha(\mathcal{S}) := \left\{ y \in Y \;\middle|\; \frac{|\{m_k < j \leq n \mid \mu_{z_j} \geq \mu_{z_n}\}|}{n - m_k} > \varepsilon \right\}, \tag{4.125}$$

  where

$$\mu_{z_j} := A\big(\mathcal{B}(z_1, \ldots, z_{m_k}), (x_j, y_j)\big)\,,$$
$$\mu_{z_n} := A\big(\mathcal{B}(z_1, \ldots, z_{m_k}), (x_n, y)\big).$$

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 ??

4.10.3 Classifier calibration

A specific instance of regression problems are classification tasks, where the one tries to model a function $f : X \to Y$ with Y finite (or discrete). For the case of probabilistic classifiers, the notion of validity 4.10.1 admits the following formulation:

Definition 4.10.16 (Calibration). A probabilistic (multiclass) classifier $\hat{P}(\cdot \mid \cdot) : Y \times X \to [0, 1]$ is said to be (well-)calibrated if

$$\Pr\Big(y \;\Big|\; \hat{P}(y \mid x) = p\Big) = p \tag{4.126}$$

for all $p \in [0, 1]$. Here, the confidence level $1 - \alpha$ is the output of the classifier $\hat{P}$, i.e. instead of asking for a region satisfying a given confidence level, the model yields the confidence level required to include a given class. A possible way to visually investigate the calibration of a probabilistic classifier is to draw **calibration plots** or **reliability diagrams** for all classes, where for every class label y and for a suitable partition $\mathcal{P} := \{0 = p_1, p_2, \ldots, p_n = 1\}$ of the interval $[0, 1]$ the i^{th} point is the proportion of instances for which y was predicted with probability between p_i and p_{i+1}. For a calibrated model, these points should lie on the diagonal.

Consider now the case of binary classification: $X \to \{0, 1\}$. Here one can easily apply the conformal prediction framework from the previous section. To this end, define the following nonconformity measure:

$$A_{\text{binary}}\big(\mathcal{B}, (x, y)\big) := 1 - \hat{P}(y \mid x; \mathcal{B})\,, \tag{4.127}$$

where the classifier is possibly estimated on $\mathcal{B}$. At significance α, the model predicts all classes such that the conformal p-value is greater than α. As before, one can adopt an inductive framework and use the $(1 - \alpha)$-quantile of the calibration scores as a threshold.

In fact, for classification tasks, where the codomain is finite (or at least discrete), it makes sense to slightly change the approach. Instead of fixing the significance level α, one could (or should) consider some specific values:

- **Confidence**: $\sup\{1 - \alpha \mid |\Gamma^\alpha(x)| \leq 1\}$, and
- **Credibility**: $\inf\{\alpha \mid |\Gamma^\alpha(x)| = 0\}$.

The former is the highest probability with which the model can predict a single class. If one wants a higher probability, more than one class has to be considered. The latter gives a measure of how ''credible'' the predictions are. The smaller this number, i.e. the greater the confidence with which the model predicts a vacuously false result (every sample necessarily belongs to a class), the less reliable the model is for a given sample.

Various othe approaches exist to calibrate existing probabilistic classifiers. In practice it has been observed that many models show sigmoidal distortions in their calibration plots. A possible approach is then to modify the final sigmoidal layer in these models (or add an extra sigmoidal layer if the model did not use one). This gives rise to **Platt scaling** and **temperature scaling**. For the former one takes the output $\hat{P}$ and fits a sigmoidal layer of the form

$$\hat{P}_{\text{Platt}}(y \mid x; A, B) = \frac{1}{1 + \exp\big(-A\hat{P}(y \mid x) - B\big)}. \tag{4.128}$$

For the latter one modifies the existing logits as follows for some parameter $T > 0$:

$$\hat{P}_{\text{temp}}(y \mid x; T) = \frac{1}{1 + \exp\big(-\hat{z}(y \mid x)/T\big)}\,, \tag{4.129}$$

where $\hat{z}$ represents the logits of the estimator $\hat{P}$.

Remark 4.10.17 (Accuracy). Because temperature scaling does not change the maximum of the softmax function, the eventual predictions do not change.

4.10.4 Normalizing flows

One of the main problems for obtaining valid uncertainty estimates is that the underlying distribution of a given data set is often unknown. Except for the conformal prediction framework, all other methods make some assumptions about the underlying data generating process (even CP makes the exchangeability assumption). One way around this problem is by first transforming the data such that it is sampled according to a well-known distribution.

4.10.5 Conditionality

The above sections considered confidence predictors that were valid in a global sense, i.e. averaged over the whole instance space $X \times Y$. However, as in ordinary probability theory, in many settings it makes more sense to consider conditional statements:

$$P\big(y \in \Gamma^\alpha(x) \mid \kappa(x, y)\big) \geq 1 - \alpha\,, \tag{4.130}$$

where the function $\kappa : X \times Y \to K$ represents the conditioning statement (the set K can be any set). In the conformal prediction literature it is often called a **taxonomy function**.

In practice, the label set K is often finite and discrete, i.e. one considers a finite subdivision of $X \times Y$. The simplest solution to obtain conditional validity in this case is to apply any known algorithm to every conditional class individually. In the case of conformal prediction this gives rise to the notion of **Mondrian conformal predictors**.

In a perfect world, the ultimate goal should be to have exact objectwise validity:

$$P(y \in \Gamma^\alpha(x) \mid x) \geq 1 - \alpha. \tag{4.131}$$

However, in general, one can show that this cannot be attained:

Property 4.10.18 (No-go theorem). Let X be a separable metric space equipped with its canonical Borel σ-algebra and consider a confidence predictor Γ^α at conditional significance level $\alpha \in [0, 1]$. For every probability distribution P on $X \times Y$ and P_X-almost all non-atoms $x \in X$ one has

$$\Pr(\lambda(\Gamma^\alpha(x)) = +\infty) \geq 1 - \alpha \tag{4.132}$$

and

$$\Pr(\mathrm{co}(\Gamma^\alpha(x)) = \mathbb{R}) \geq 1 - 2\alpha\,, \tag{4.133}$$

where λ and co denote the Lebesgue measure and convex hull, respectively.

?? FINISH ??

4.10.6 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 1.6.6 from the p-values produced by a conformal predictor. For every (reasonable) martingale $(X_n)_{n\in\mathbb{N}}$ the Doob-Ville inequality 1.6.8 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 [4].

The general expression of the martingale will be of the form

$$X_i \equiv \prod_{k=0}^{i} f_k(p_k)\,, \tag{4.134}$$

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

$$\int_0^1 f_k(p)\, dp = 1. \tag{4.135}$$

Method 4.10.19 (Power martingale). For every constant $\varepsilon \in [0,1]$ one defines the power martingale as

$$X_i^{\varepsilon} := \prod_{k=0}^{i} \varepsilon p_k^{\varepsilon - 1}. \tag{4.136}$$

One can also construct a **simple mixture** martingale by integrating over ε:

$$X_i := \int_0^1 X_i^{\varepsilon} \, d\varepsilon. \tag{4.137}$$

Because $\varepsilon - 1 \leq 0$, 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 4.10.20 (Plug-in martingale). Let $\hat{\rho}_i(p)$ denote an estimate of the probability density constructed using the p-values $\{p_1, \ldots, p_i\}$. The plug-in martingale is defined by the betting functions

$$f_i := \hat{\rho}_{i-1}. \tag{4.138}$$

Now, if the empirical distribution functions of the p-values converge weakly **??** 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.

A different approach is to use p-values that are constructed directly from disitributional data. For inspiration, consider the following property:

Property 4.10.21. Consider a data sample $(x_n)_{n\in\mathbb{N}}$ and let f, g be two possible probability densities describing this sample. If f is the true density, the likelihood process

$$X_i := \prod_{k=0}^{i} \frac{g(x_i)}{f(x_i)} \tag{4.139}$$

is a martingale.

So likelihood ratios already give rise to test martingales. However, this martingale cannot be used to check for distribution shifts, since it is only a martingale when the true density is used.

Example 4.10.22 (Likelihood nonconformity). For every two probability densities f, g one can define a nonconformity measure as follows:

$$A_{\mathrm{NP}}(\mathcal{B}, z) := \log f(z) - \log g(z)\,, \tag{4.140}$$

i.e. the log-likelihood ratio.[15]

By using this nonconformity measure in combination with the above plug-in approach, one can obtain a likelihood-based changepoint detection algorithm. If the initial distribution is not known, it can be estimated based on the bag $\mathcal{B}$.

[15] The subscript refers to the Neyman-Pearson lemma 3.7.12 for which this function gives the (logarithm) of the test statistic.

Chapter 5

Fuzzy Sets & Imprecise Probabilities ♣

The main reference for the basics on fuzzy sets is the original paper [81]. For the basics of (ordered) sets, see Section **??** at the beginning of this compendium. The section on imprecise probabilities is mainly based on [9].

This chapter begins with a small organizational remark. Although the content of the current chapter fits better in the parts on general set theory and logic, it does use more advanced concepts from for example topology and category theory. Furthermore, the main application is the characterization of uncertainty in statistics and machine learning. For that reason it was added here.

5.1 Fuzzy sets

Definition 5.1.1 (Fuzzy set). Consider a set X (this set corresponds to the universe of discours in e.g. type theory or category theory). A fuzzy subset of X is a function $A : X \to [0, 1]$. One can interpret the value $A(x)$ at a point $x \in X$ as the grade of membership of x in A. If the function A only takes on values in $\{0, 1\}$, the indicator function of an ordinary subset is obtained.

A fuzzy set is said to be **empty** if its defining function is identically zero.

Remark 5.1.2. One can generalize this definition by replacing $[0, 1]$ by a more general poset (with the necessary properties).

Definition 5.1.3 (Pullback). Consider two sets X, Y and a fuzzy subset A of Y. Given a function $f : X \to Y$ one can define the pullback f^*A as usual:

$$f^*A(x) := A(f(x)). \tag{5.1}$$

The following definition is an immediate generalization of Definition **??**:

Definition 5.1.4 (Fuzzy relation). A fuzzy subset of the product set $X \times X$. This definition can be extended to n-ary relations by considering fuzzy subsets of the n-fold product $X \times \cdots \times X$.

The composition in Definition **??** can be extended through the following formula:

$$S \circ R(x, z) := \sup_{y \in X} \min\big(R(x, y), S(y, z)\big). \tag{5.2}$$

A more exotic construction for fuzzy sets is the following one (note that this only works if the codomain of fuzzy sets is $[0, 1]$):

Definition 5.1.5 (Convex combination). Consider three fuzzy sets A, B, Λ. The convex combination $C \equiv (A, B; \Lambda)$ is defined as follows in analogy to Definition **??**:

$$C(x) := \Lambda(x)A(x) + (1 - \Lambda(x))B(x). \tag{5.3}$$

5.2 Fuzzy measure theory

In this section some of the content of Chapter **??** is generalized to fuzzy set theory. Unless explicitly stated, all concepts will be defined over a general measurable space (X, Σ).

Definition 5.2.1 (Capacity[1]). A set function $\mu : \Sigma \to \mathbb{R}$ satisfying the following conditions:

1. **Grounded**: $\emptyset \in \Sigma \implies \mu(\emptyset) = 0$, and
2. **Monotonicity**: $A \subseteq B \implies \mu(A) \leq \mu(B)$ for all $A, B \in \mathcal{C}$.

If one drops the monotonicity condition, the notion of a **game** is obtained. A capacity is said to be **normalized** (or **regular**) if $\mu(X) = 1$. For infinite sets X, the following two conditions are generally added:

3 **Upward continuity**: If $(A_n)_{n\in\mathbb{N}} \subset \Sigma$ is an increasing sequence, then

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \lim_{n\to\infty} \mu(A_n). \tag{5.4}$$

4 **Downward continuity**: If $(A_n)_{n\in\mathbb{N}} \subset \Sigma$ is a decreasing sequence of compact sets, then

$$\mu\left(\bigcap_{n\in\mathbb{N}} A_n\right) = \lim_{n\to\infty} \mu(A_n). \tag{5.5}$$

Definition 5.2.2 (Alternating capacity). A k-alternating capacity μ satisfies

$$\mu\left(\bigcap_{i=1}^{k} A_i\right) \leq \sum_{I\subset\{1,\ldots,k\}} (-1)^{|I|+1} \mu\left(\bigcup_{i\in I} A_i\right) \tag{5.6}$$

for all measurable sets $A_1, \ldots, A_k$. A 2-alternating capacity is called a **probability measure** if the inequality is saturated for all A_1, A_2. By interchanging the union and intersection symbols (and the inequality sign), the definition of **k-monotone capacities** is obtained. If a capacity is k-alternating (resp. k-monotone) for all $k \geq 2$, it is also said to be **totally alternating** (resp. **totally monotone**).

Property 5.2.3. A capacity $\mu : \Sigma \to \mathbb{R}$ is k-monotone (resp. k-alternating) if and only if its **dual capacity**

$$\overline{\mu}(A) := \mu(X) - \mu(A^c) \tag{5.7}$$

is k-alternating (resp. k-monotone).

[1] Also called a **fuzzy measure**.

Definition 5.2.4 (Choquet integral). Consider a capacity μ on (X, Σ) and a measurable function $f : \mathcal{U} \to \mathbb{R}$, i.e. a function such that $\{y \mid f(y) \geq x\} \in \Sigma$ for all $x \in \mathbb{R}$. The Choquet integral of f is defined as follows:

$$\int_X f \, d\mu := \int_{-\infty}^{0} \big(\mu(\{y \mid f(y) \geq x\}) - \mu(X)\big) \, dx + \int_0^{+\infty} \mu(\{y \mid f(y) \geq x\}) \, dx. \tag{5.8}$$

This integral is not additive, but it is monotonic in f. This expression can be obtained as follows, where functions are decomposed in their positive and negative parts: $f = f^+ - f^-$. The original definition by *Choquet* was for nonnegative functions:

$$\int_X f^+ \, d\mu := \int_0^{+\infty} \mu(\{y \mid f^+(y) \geq x\}) \, dx \,, \tag{5.9}$$

which for simple functions reduces to:

$$\begin{aligned} \int_X f^+ \, d\mu &= \int_X \sum_{i=1}^{n} a_i \mathbb{1}_{\{x \in X \mid f^+(x) = a_i\}} \, d\mu \\ &= \int_X \sum_{i=1}^{n} (a_i - a_{i-1}) \mathbb{1}_{\{x \in X \mid f^+(x) \geq a_i\}} \, d\mu \\ &= \sum_{i=1}^{n} (a_i - a_{i-1}) \mu(\{x \in X \mid f(x) \geq a_i\}). \end{aligned} \tag{5.10}$$

The general Choquet integral then takes the form

$$\int_X f \, d\mu := \int_X f^+ \, d\mu - \int_X f^- \, d\overline{\mu}. \tag{5.11}$$

Property 5.2.5. A capacity is 2-alternating if and only if the associated Choquet integral is subadditive.

Example 5.2.6 (Possibility and necessity). A normalized capacity μ satisfying

$$\mu(A \cup B) = \max\big\{\mu(A), \mu(B)\big\} \tag{5.12}$$

for all $A, B \in \Sigma$. If one replaces the maximum by a minimum, the definition of a **necessity measure** is obtained.

Example 5.2.7 (Belief and plausibility). Belief and plausibility measures are respectively defined as totally monotone and totally alternating normalized capacities. It is not difficult to show that possibility and necessity measures are particular instances of belief and plausibility measures.

Definition 5.2.8 (Consonant capacity). A capacity $\mu : \Sigma \to \mathbb{R}$ such that the **focal sets**, the sets $A \in \Sigma$ such that $\mu(A) > 0$, admit a total order (i.e. they are nested).

Example 5.2.9. Necessity and possibility measures are exactly the consonant belief and plausibility measures.

5.3 Imprecise probabilities

Definition 5.3.1 (Gamble). Consider a set X. The set of gambles over X is the (Banach) space of bounded real-valued functions $\mathcal{B}(X) := \{f : X \to \mathbb{R} \mid \sup_{x \in X} f(x) < +\infty\}$. A subset $\mathcal{D} \subset \mathcal{B}(X)$ 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:** $\mathcal{B}^+(X) \subseteq \mathcal{D}$, where $\mathcal{B}^+(X) := \{f \in \mathcal{B}(X) \mid f > 0\}$, and
4. **Avoiding partial losses:** $\mathcal{B}^-(X) \cap \mathcal{D} = \emptyset$, $\mathcal{B}^-(X) := \{f \in \mathcal{B}(X) \mid f < 0\}$.

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

Property 5.3.2 (Order structure). The collection of all coherent sets of desirable gambles over a space X can be given a poset structure by inclusion, with $\mathcal{B}^+(X)$ as its least element. If $\mathcal{D} \subset \mathcal{D}^+$, then $\mathcal{D}$ is said to be less **committal** than $\mathcal{D}'$.

Definition 5.3.3 (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 $\lambda P_1 + (1-\lambda)P_2, \lambda \in [0,1]$.

List of Symbols

The following symbols are used throughout the summary:

Abbreviations

FIP	finite intersection property
GA	geometric algebra
PVM	projection-valued measure
TVS	topological vector space

Operations

e	identity element of a group
∂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
$X \times Y$	cartesian product of the sets 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

$\mathcal{B}_0(V, W)$	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
$C(X, Y)$	set of continuous functions between two topological spaces X and Y
D^n	standard n-disk
$\mathbf{Open}(X)$	category of open subsets of a topological space X
$\mathrm{Pin}(V)$	pin group of the Clifford algebra $C\ell(V, Q)$
$\mathbf{Sh}(\mathbf{C}, J)$	category of J-sheaves on a site $(\mathbf{C}, J)$
S^n	standard n-sphere
$S^n(V)$	space of symmetric rank n tensors over a vector space V
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
$\emptyset$	empty set

$[a, b]$	closed interval
$]a, b[$	open interval
$\Lambda^n(V)$	space of antisymmetric rank n tensors over a vector space V
$\rho(A)$	resolvent set of a bounded linear operator A

Bibliography

[1] Jonathan R. Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. Technical report, 1994.

[2] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(Oct):1705--1749, 2005.

[3] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371-–421, 2008.

[4] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

[5] Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. Bregman Voronoi diagrams. *Discrete & Computational Geometry*, 44(2):281--307, 2010.

[6] Shun-ichi Amari. *Information Geometry and Its Applications*. Springer Publishing Company, Incorporated, 2016.

[7] Niclas Sandgren and Petre Stoica. On moving average parameter estimation. Technical Report 2006-022, Department of Information Technology, Uppsala University, 2006.

[8] 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.

[9] Thomas Augustin, Frank PA Coolen, Gert De Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

[10] Tetsuji Miwa, Michio Jimbo, Michio Jimbo, and E Date. *Solitons: Differential Equations, Symmetries and Infinite-dimensional Algebras*, volume 135. Cambridge University Press, 2000.

[11] Vladimir I. Arnol'd. *Mathematical Methods of Classical Mechanics*, volume 60. Springer Science & Business Media, 2013.

[12] Edwin J. Beggs and Shahn Majid. *Quantum Riemannian Geometry*. Springer, 2020.

[13] Marc Henneaux and Claudio Teitelboim. *Quantization of Gauge Systems*. Princeton university press, 1992.

[14] Mark Hovey. *Model Categories*. Number 63. American Mathematical Soc., 2007.

[15] Gregory M. Kelly. *Basic Concepts of Enriched Category Theory*, volume 64. CUP Archive, 1982.

[16] Mukund Rangamani and Tadashi Takayanagi. *Holographic Entanglement Entropy.* Springer, 2017.

[17] Saunders Mac Lane. *Categories for the Working Mathematician*, volume 5. Springer Science & Business Media, 2013.

[18] Peter T. Johnstone. *Topos Theory.* Dover Publications, 2014.

[19] Charles W. Misner, Kip S. Thorne, and John A. Wheeler. *Gravitation.* Princeton University Press, 2017.

[20] Carlo Rovelli and Francesca Vidotto. *Covariant Loop Quantum Gravity: An Elementary Introduction to Quantum Gravity and Spinfoam Theory.* Cambridge University Press, 2014.

[21] Richard W. Sharpe. *Differential Geometry: Cartan's Generalization of Klein's Erlangen Program*, volume 166. Springer Science & Business Media, 2000.

[22] John C. Baez, Irving E. Segal, and Zhengfang Zhou. *Introduction to Algebraic and Constructive Quantum Field Theory.* Princeton University Press, 2014.

[23] Raoul Bott and Loring W. Tu. *Differential Forms in Algebraic Topology.* Graduate Texts in Mathematics. Springer New York, 1995.

[24] The Univalent Foundations Program. *Homotopy Type Theory: Univalent Foundations of Mathematics.* `https://homotopytypetheory.org/book`, Institute for Advanced Study, 2013.

[25] Bruce Blackadar. *Operator Algebras: Theory of C∗-Algebras and von Neumann Algebras.* Springer, 2013.

[26] Marek Capinski and Peter E. Kopp. *Measure, Integral and Probability.* Springer Science & Business Media, 2013.

[27] Georgiev Svetlin. *Theory of Distributions.* Springer, 2015.

[28] Gerd Rudolph and Matthias Schmidt. *Differential Geometry and Mathematical Physics: Part II. Fibre Bundles, Topology and Gauge Fields.* Springer, 2017.

[29] Martin Schottenloher. *A Mathematical Introduction to Conformal Field Theory*, volume 759. 2008.

[30] Dusa McDuff and Deitmar Salamon. *Introduction to Symplectic Topology.* Oxford Graduate Texts in Mathematics. Oxford University Press, 2017.

[31] John C. Baez and Peter May. *Towards Higher Categories*, volume 152 of *IMA Volumes in Mathematics and its Applications.* Springer, 2009.

[32] 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.

[33] Geoffrey Compère. *Advanced Lectures on General Relativity*, volume 952. Springer, 2019.

[34] Pavel Etingof, Shlomo Gelaki, Dmitri Nikshych, and Victor Ostrik. *Tensor Categories*, volume 205. American Mathematical Soc., 2016.

[35] 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.

[36] Charles A. Weibel. *An Introduction to Homological Algebra*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1994.

[37] Peter J. Hilton and Urs Stammbach. *A Course in Homological Algebra*. Springer.

[38] Jean-Luc Brylinski. *Loop Spaces, Characteristic Classes and Geometric Quantization*. Birkhauser.

[39] Antoine Van Proeyen and Daniel Freedman. *Supergravity*. Cambridge University Press.

[40] William S. Massey. *A Basic Course in Algebraic Topology*. Springer.

[41] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to Quantum Field Theory*. Westview Press.

[42] Nadir Jeevanjee. *An Introduction to Tensors and Group Theory for Physicists*. Birkhauser.

[43] Yvonne Choquet-Bruhat, Cecile DeWitt-Morette, and Margaret Dillard-Bleick. *Analysis, Manifolds and Physics, Part 1: Basics*. North-Holland.

[44] Yvonne Choquet-Bruhat and Cecile DeWitt-Morette. *Analysis, Manifolds and Physics, Part 2*. North-Holland.

[45] Herbet Goldstein, John L. Safko, and Charles P. Poole. *Classical Mechanics*. Pearson.

[46] Franco Cardin. *Elementary Symplectic Topology and Mechanics*. Springer.

[47] Walter Greiner and Joachim Reinhardt. *Field Quantization*. Springer.

[48] Walter Greiner. *Quantum Mechanics*. Springer.

[49] B. H. Bransden and Charles J. Joachain. *Quantum Mechanics*. Prentice Hall.

[50] Heydar Radjavi and Peter Rosenthal. *Invariant Subspaces*. Dover Publications.

[51] Max Karoubi. *K-Theory: An Introduction*. Springer.

[52] Damien Calaque and Thomas Strobl. *Mathematical Aspects of Quantum Field Theories*. Springer, 2015.

[53] Ivan Kolar, Peter W. Michor, and Jan Slovak. *Normal Operations in Differential Geometry*. Springer.

[54] Stephen B. Sontz. *Principal Bundles: The Classical Case*. Springer.

[55] Stephen B. Sontz. *Principal Bundles: The Quantum Case*. Springer.

[56] William Fulton and Joe Harris. *Representation Theory: A First Course*. Springer.

[57] Peter Petersen. *Riemannian Geometry*. Springer.

[58] Charles Nash and Siddharta Sen. *Topology and Geometry for Physicists*. Dover Publications.

[59] Ian M. Anderson. *The Variational Bicomplex*.

[60] Joel Robbin and Dietmar Salamon. The maslov index for paths. *Topology*, 32(4):827--844, 1993.

[61] Nima Moshayedi. 4-manifold topology, donaldson-witten theory, floer homology and higher gauge theory methods in the BV-BFV formalism. 2021.

[62] Edward Witten. Global anomalies in string theory. In *Symposium on Anomalies, Geometry, Topology*, 6 1985.

[63] F.A. Berezin and M.S. Marinov.

[64] Paul A. M. Dirac. Generalized Hamiltonian dynamics. *Canadian Journal of Mathematics*, 2:129–148, 1950.

[65] Angelo Vistoli. Notes on Grothendieck topologies, fibered categories and descent theory. *arXiv:math/0412512*, 2004.

[66] Emily Riehl and Dominic Verity. The theory and practice of Reedy categories. *Theory and Applications of Categories*, 29, 2013.

[67] Emily Riehl. Homotopical categories: From model categories to $(\infty, 1)$-categories. 2019. arXiv:1904.00886.

[68] Floris Takens. A global version of the inverse problem of the calculus of variations. *Journal of Differential Geometry*, 14(4):543--562, 1979.

[69] John Baez and Alexander Hoffnung. Convenient categories of smooth spaces. *Transactions of the American Mathematical Society*, 363(11):5789--5825, 2011.

[70] John C. Baez and Alissa S. Crans. Higher-dimensional algebra vi: Lie 2-algebras. 2003. arXiv:math/0307263.

[71] John C. Baez and Aaron D. Lauda. Higher-dimensional algebra v: 2-groups. 2003. arXiv:math/0307200.

[72] Edward Witten. Supersymmetry and Morse theory. *J. Diff. Geom*, 17(4):661--692, 1982.

[73] Urs Schreiber. *From Loop Space Mechanics to Nonabelian Strings*. PhD thesis, 2005.

[74] John C. Baez and Urs Schreiber. Higher gauge theory. 2005. arXiv:math/0511710.

[75] Jade Master. Why is homology so powerful? 2020. arXiv:2001.00314.

[76] 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.

[77] Richard Palais. The symmetries of solitons. *Bulletin of the American Mathematical Society*, 34(4):339--403, 1997.

[78] Michael F. Atiyah. Topological quantum field theory. *Publications Mathématiques de l'IHÉS*, 68:175--186, 1988.

[79] 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.

[80] Benoît Tuybens. Entanglement entropy of gauge theories. 2017.

[81] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338--353, 1965.

[82] John C. Baez, Alexander E. Hoffnung, and Christopher Rogers. Categorified symplectic geometry and the classical string. *Communications in Mathematical Physics*, 293:701--725, 2010.

[83] Charles Rezk. A model for the homotopy theory of homotopy theory. *Transactions of the American Mathematical Society*, 353(3):973--1007, 2001.

[84] Peter May. A note on the splitting principle. *Topology and Its Applications*, 153(4):605--609, 2005.

[85] Irina Markina. Group of diffeomorphisms of the unit circle as a principal $U(1)$-bundle.

[86] Sjoerd E. Crans. Localizations of transfors. 1998.

[87] Tom Leinster. Basic bicategories. 1998. arXiv:math/9810017.

[88] Alexander E. Hoffnung. Spans in 2-categories: A monoidal tricategory. 2011. arXiv:1112.0560.

[89] Eugenia Cheng and Nick Gurski. The periodic table of n-categories for low dimensions ii: Degenerate tricategories. 2007. arXiv:0706.2307.

[90] Mehmet B. Şahinoğlu, Dominic J. Williamson, Nick Bultinck, Michael Mariën, Jutho Haegeman, Norbert Schuch, and Frank Verstraete. Characterizing topological order with matrix product operators. 2014. arXiv:1409.2150.

[91] 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.

[92] Guifré Vidal. Efficient classical simulation of slightly entangled quantum computations. *Phys. Rev. Lett.*, 91, 2003.

[93] 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.

[94] Domenico Fiorenza. An introduction to the Batalin-Vilkovisky formalism. 2004. arXiv:math/0402057v2.

[95] Stefan Cordes, Gregory Moore, and Sanjaye Ramgoolam. Lectures on 2d Yang-Mills theory, equivariant cohomology and topological field theories. arXiv:hep-th/9411210v2.

[96] Donald C. Ferguson. A theorem of Looman-Menchoff. `http://digitool.library.mcgill.ca/thesisfile111406.pdf`.

[97] Holger Lyre. Berry phase and quantum structure. arXiv:1408.6867.

[98] Florin Belgun. Gauge theory. `http://www.math.uni-hamburg.de/home/belgun/Gauge4.pdf`.

[99] Vladimir Itskov, Peter J. Olver, and Francis Valiquette. Lie completion of pseudogroups. *Transformation Groups*, 16:161--173, 2011.

[100] Richard Borcherds. Lie groups. `https://math.berkeley.edu/~reb/courses/261/`.

[101] Andrei Losev. From Berezin integral to Batalin-Vilkovisky formalism: A mathematical physicist's point of view. 2007.

[102] Edward Witten. Coadjoint orbits of the Virasoro group. *Comm. Math. Phys.*, 114(1):1--53, 1988.

[103] Sidney R. Coleman and Jeffrey E. Mandula. All possible symmetries of the S-matrix. *Phys. Rev.*, 159, 1967.

[104] Emily Riehl. Monoidal algebraic model structures. *Journal of Pure and Applied Algebra*, 217(6):1069--1104, 2013.

[105] Valter Moretti. Mathematical foundations of quantum mechanics: An advanced short course. *International Journal of Geometric Methods in Modern Physics*, 13, 2016.

[106] Antonio Michele Miti. Homotopy comomentum maps in multisymplectic geometry, 2021.

[107] John E. Roberts. Spontaneously broken gauge symmetries and superselection rules. 1974.

[108] Richard Sanders. Commutative spectral triples & the spectral reconstruction theorem.

[109] Jean Gallier. Clifford algebras, Clifford groups, and a generalization of the quaternions, 2008. arXiv:0805.0311.

[110] Bozhidar Z. Iliev. Normal frames for general connections on differentiable fibre bundles. arXiv:math/0405004.

[111] Piotr Stachura. Short and biased introduction to groupoids. arXiv:1311.3866.

[112] Fosco Loregian. Coend calculus. arXiv:1501.02503.

[113] Frederic Schuller. Lectures on the geometric anatomy of theoretical physics. `https://www.youtube.com/channel/UC6SaWe7xeOp31Vo8cQG1oXw`.

[114] Nima Amini. Infinite-dimensional Lie algebras. `https://people.kth.se/~namini/PartIIIEssay.pdf`.

[115] Peter Selinger. Lecture notes on lambda calculus.

[116] Nigel Hitchin. Lectures on special Lagrangian submanifolds. `https://arxiv.org/abs/math/9907034v1`, 1999.

[117] Olivia Caramello. Lectures on topos theory at the university of Insubria. `https://www.oliviacaramello.com/Teaching/Teaching.htm`.

[118] Derek Sorensen. An introduction to characteristic classes. `http://derekhsorensen.com/docs/sorensen-characteristic-classes.pdf`, 2017.

[119] Arun Debray. Characteristic classes. `https://web.ma.utexas.edu/users/a.debray/lecture_notes/u17_characteristic_classes.pdf`.

[120] Pascal Lambrechts.

[121] Chris Tiee. Contravariance, covariance, densities, and all that: An informal discussion on tensor calculus. `https://ccom.ucsd.edu/~ctiee/notes/tensors.pdf`, 2006.

[122] Emily Riehl. Homotopy (limits and) colimits. `http://www.math.jhu.edu/~eriehl/hocolimits.pdf`.

[123] Andreas Gathmann. Algebraic geometry. `https://www.mathematik.uni-kl.de/~gathmann/class/alggeom-2019/alggeom-2019.pdf`.

[124] Will J. Merry. Algebraic topology. `https://www.merry.io/algebraic-topology`.

[125] Stacks project. `https://stacks.math.columbia.edu/`.

[126] The nlab. `https://ncatlab.org/nlab`.

[127] Wikipedia. `https://www.wikipedia.org/`.

[128] Joost Nuiten. Cohomological quantization of local prequantum boundary field theory. Master's thesis, 2013.

Index