Statistics In
Data Science

Introduction

Lecture Notes

Selected Topics

Winter 2022/ 2023

Alois Pichler

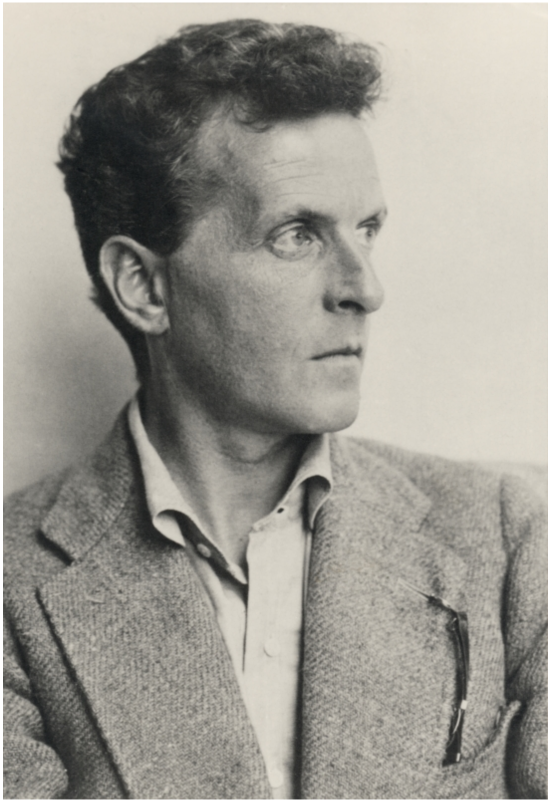
TECHNISCHE UNIVERSITÄT
CHEMNITZ
Faculty of Mathematics

DRAFT
Version as of March 2, 2023

Contents

1 Introduction

> Die Grenzen meiner Sprache
> bedeuten die Grenzen meiner Welt.
>
> Ludwig Wittgenstein, 1889–1951,
> *tractatus logico philosophicus 5.6*



(a) Ludwig Wittgenstein

(b) Julia

Figure 1.1: Alan Edelman: “Good programming language design is applied psychology”

For the online version, see

https://www.tu-chemnitz.de/mathematik/fima/public/mathematischeStatistik.pdf

for an introduction.

Related areas include

(i) data science

(ii) statistical learning

(iii) machine learning

  (a) supervised learning
  (b) unsupervised learning
  (c) reinforcement learning

(iv) statistical pattern recognition

(v) reinforcement learning vs supervised learning

(vi) artificial neural networks, a branch of artificial intelligence

Literature includes Pflug [12], Cressie [6], Bhattacharya et al. [1], Tamhane and Dunlop [16], Kersting and Wakolbinger [8] and Bottou et al. [3] or Bishop [2].

Distributions

2

> Alles was Gegenstand des Denkens ist, ist daher Gegenstand der Mathematik. Die Mathematik ist nicht die Kunst des Rechnens, sondern die Kunst des Nichtrechnens.
>
> David Hilbert, 1862–1943

2.1 BINOMIAL DISTRIBUTION

Definition 2.1. Given the parameters $p \in [0,1]$ and $n \in \mathbb{N}$, the binomial distribution bin(n,p) has the probability mass function $\binom{n}{k} p^k (1-p)^{n-k}$.

Proposition 2.2. *The expectation and variance of a random variable $X \sim$ bin(n,p) are $\mathbb{E}\, X = n \cdot p$ and $\operatorname{var} X = n\,p\,(1-p)$.*

Proof. Indeed, $\mathbb{E}\, X = \sum_{k=0}^n k \cdot P(X=k) = \sum_{k=0}^n k \cdot \binom{n}{k} p^k (1-p)^{n-k} = np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = n \cdot p$, the first assertion.

Further we have that

$$\begin{aligned}\mathbb{E}\, X(X-1) &= \sum_{k=0}^n k(k-1) \cdot P(X=k) = \sum_{k=0}^n k(k-1) \cdot \binom{n}{k} p^k (1-p)^{n-k} \\ &= n(n-1)p^2 \sum_{k=2}^n \binom{n-2}{k-2} p^{k-2} (1-p)^{n-k} = n(n-1)p^2.\end{aligned}$$

It follows that

$$\begin{aligned}\operatorname{var} X &= \left(\mathbb{E}\, X^2\right) - (\mathbb{E}\, X)^2 = \mathbb{E}\, X(X-1) + \mathbb{E}\, X - (\mathbb{E}\, X)^2 \\ &= n(n-1)p^2 + np - (np)^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p),\end{aligned}$$

the remaining assertion. □

Theorem 2.3 (De Moivre–Laplace theorem)**.** *It holds that*

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{1}{2} \frac{(k-\mu_n)^2}{\sigma_n^2}\right),$$

where $\mu_n \coloneqq n\,p$ and $\sigma_n \coloneqq \sqrt{n\,p(1-p)}$.

Proof. We shall employ Stirling's formula, $k! \sim \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$. Then

$$\begin{aligned}
\binom{n}{k}p^k(1-p)^{n-k} &= \frac{n!}{k!\cdot(n-k)!}p^k(1-p)^{n-k}\\
&\sim \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi k}\left(\frac{k}{e}\right)^k\cdot\sqrt{2\pi(n-k)}\left(\frac{n-k}{e}\right)^{n-k}}p^k(1-p)^{n-k}\\
&= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}\frac{n^{n-k}n^k}{k^k(n-k)^{n-k}}p^k(1-p)^{n-k}\\
&= \frac{1}{\sqrt{2\pi\,\frac{k(n-k)}{n}}}\cdot\left(\frac{np}{k}\right)^k\left(\frac{n(1-p)}{n-k}\right)^{n-k}.\\
&= \frac{1}{\sqrt{2\pi\,\frac{k(n-k)}{n}}}\cdot\exp\left(-n\cdot\eta\left(\frac{k}{n}\right)\right),
\end{aligned}$$

where $\eta(t) := t\ln\frac{t}{p}+(1-t)\ln\frac{1-t}{1-p}$. Note, that $\eta'(t)=\log\frac{t}{p}-\log\frac{1-t}{1-p}$ and $\eta''(t)=\frac{1}{t}+\frac{1}{1-t}$, so that $\eta(p)=0$, $\eta'(p)=0$ and $\eta''(p)=\frac{1}{p(1-p)}$; we find the Taylor series expansion $\eta(t)\approx\frac{(t-p)^2}{2p(1-p)}$. Consequently,

$$\begin{aligned}
\binom{n}{k}p^k(1-p)^{n-k} &\sim \frac{1}{\sqrt{2\pi\, n\frac{k}{n}\left(1-\frac{k}{n}\right)}}\cdot\exp\left(-n\cdot\eta\left(\frac{k}{n}\right)\right)\\
&= \frac{1}{\sqrt{2\pi n\,p(1-p)}}\exp\left(-n\frac{(k/n-p)^2}{2p(1-p)}\right)\\
&= \frac{1}{\sqrt{2\pi\cdot np(1-p)}}\exp\left(-\frac{1}{2}\left(\frac{k-np}{\sqrt{np(1-p)}}\right)^2\right)
\end{aligned}$$

and thus the assertion. □

2.2 POISSON DISTRIBUTION

Definition 2.4. The Poisson distribution has probability mass function

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},\qquad k=0,1,2,\dots$$

Proposition 2.5. *It holds that* $\mathbb{E}\,X=\operatorname{var}X=\lambda$.

Proof. Indeed,

$$\mathbb{E}\,X=\sum_{k=0}k\cdot P(X=k)=\sum_{k=0}k\cdot\frac{\lambda^k}{k!}e^{-\lambda}=\lambda\cdot\sum_{k=1}\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda}=\lambda$$

and

$$\begin{aligned}\operatorname{var} X &= \mathbb{E}\, X(X-1) + \mathbb{E}\, X - (\mathbb{E}\, X)^2 \\ &= \sum_{k=0} k(k-1)\cdot\frac{\lambda^k}{k!}e^{-\lambda} + \lambda - \lambda^2 \\ &= \lambda^2 \cdot \sum_{k=2}\frac{\lambda^{k-2}}{(k-2)!}e^{-\lambda} + \lambda - \lambda^2 = \lambda,\end{aligned}$$

the assertion. □

Theorem 2.6 (Poisson limit theorem)**.** *Suppose that* $n\cdot p_n \xrightarrow[n\to\infty]{} \lambda$*, then, for* $k = 0,\ 1, \dots$ *fixed,*

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} \xrightarrow[n\to\infty]{} \frac{\lambda^k}{k!}e^{-\lambda}.$$

Proof. Indeed,

$$\begin{aligned}\binom{n}{k} p_n^k (1-p_n)^{n-k} &\sim \frac{n(n-1)\cdots(n-k+1)}{n^k}\cdot\frac{\lambda^k}{k!}\cdot\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &\sim \frac{\lambda^k}{k!}e^{-\lambda},\end{aligned}$$

as $\left(1-\frac{\lambda}{n}\right)^k \xrightarrow[n\to\infty]{} 1$. Hence the assertion. □

2.3 BENFORD'S LAW

Theorem 2.7 (The significant-digit phenomenon, Newcomb–Benford law)**.** *Let* $X > 0$ *be a random variable and set*

$$h(X) := \text{the first decimal digit in } X.$$

Then, under a mild model assumption, $P\big(h(X) = b\big) = \log_{10}\left(1 + \frac{1}{b}\right)$ *for* $b = 1, \dots, 9$*, cf. Table 2.1.*

b	1	2	3	4	5	6	7	8	9
$P\big(h(X)=b\big)$	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

Table 2.1: Probabilities of Benford's law

Proof. The number X has $n+1$ decimal digits, where $n = \lfloor \log_{10} X \rfloor$. The first decimal digit is $b \in \{1,\ 2, \dots,\ 9\}$, iff

$$\begin{aligned} b\cdot 10^n \le X < (b+1)\cdot 10^n, &\text{ or} \\ \log_{10} b + n \le \log_{10} X < \log_{10}(b+1) + n, &\text{ or} \\ \log_{10} b \le \operatorname{frac}\left(\log_{10} X\right) < \log_{10}(b+1), &\end{aligned}$$

where $\operatorname{frac}(x) := x - \lfloor x \rfloor$ is the fractional part of x. Note that $0 < \log_{10} b < \log_{10}(b+1) \leq 1$. We specify the model assumption so that $\operatorname{frac}\left(\log_{10} X\right) \in [0,1] \sim U$ is uniformly distributed. Then it holds that

$$\{h(X) = b\} = \left\{U \in \left[\log_{10} b,\ \log_{10}(b+1)\right)\right\}$$

with probability $P\big(h(X) = b\big) = \log_{10}(b+1) - \log_{10} b = \log_{10}\left(1 + \frac{1}{b}\right)$, the assertion. □

Corollary 2.8 (Scale invariance)**.** *If X satisfies Benford's law, then λX as well, where $\lambda > 0$.*

Proof. It holds that $\operatorname{frac}\left(\log_{10}(\lambda X)\right) = \operatorname{frac}\left(\log_{10} \lambda + \log_{10} X\right) \sim U$ is uniformly distributed as well and thus the assertion. □

2.4 IMPORTANT DENSITIES IN DATA SCIENCE

Define the functions

(i) $k_1(x) := \frac{1}{e^{\pi x/2} + e^{-\pi x/2}}$,

(ii) $k_2(x) := \frac{2}{\pi\sqrt{12}} \frac{1}{\left(e^{\pi x/\sqrt{12}} + e^{-\pi x/\sqrt{12}}\right)^2}$,

(iii) $k_3(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ and

(iv) $k_4(x) := \frac{\sqrt{2}}{2} \exp(-\sqrt{2}|x|)$ (Laplace distribution).

Lemma 2.9. *All functions (i)–(iii) are densities with unit variance: it holds that*

$$\int_{\infty}^{\infty} k_i(x)\,\mathrm{d}x = 1,\ \int_{\infty}^{\infty} x\, k_i(x)\,\mathrm{d}x = 0 \text{ and } \int_{\infty}^{\infty} x^2\, k_i(x)\,\mathrm{d}x = 1$$

for $k \in \{k_i : i = 1,2,3,4\}$.

Lemma 2.10 (Antiderivatives)**.** *It holds that*

(i) $K_1(x) := \int_{-\infty}^{x} k_1(t)\,\mathrm{d}t = \frac{2}{\pi} \arctan e^{\frac{\pi x}{2}}$,

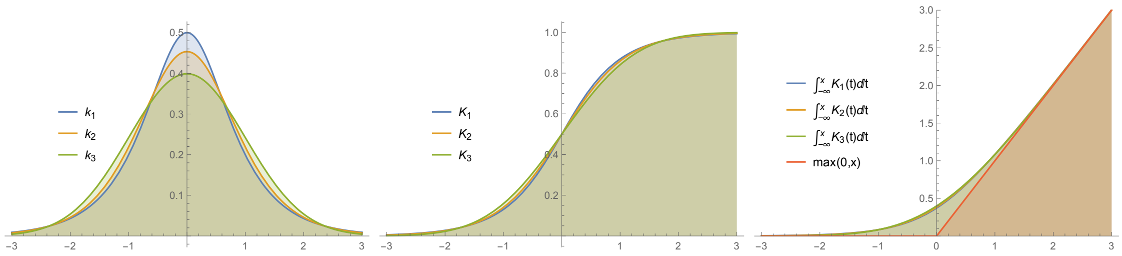
(ii) $K_2(x) := \int_{-\infty}^{x} k_2(t)\,\mathrm{d}t = \frac{1}{1+e^{-\pi x/\sqrt{3}}} = \frac{1}{2}\left(1 + \tanh \frac{\pi x \sqrt{3}}{6}\right)$,

(iii) $K_3(x) := \int_{-\infty}^{x} k_3(t)\,\mathrm{d}t = \Phi(x)$ *and*

(iv) $K_4(x) := \int_{-\infty}^{x} k_4(t)\,\mathrm{d}t = \frac{1}{2} + \frac{\operatorname{sign}(x)}{2}\left(1 - \exp(-\sqrt{2}|x|)\right)$.

Proposition 2.11 (Rectifiers)**.** *It holds that*

(i) $\int_{-\infty}^{x} K(t)\,\mathrm{d}t = \int_{-\infty}^{x} (x-t)\, k(t)\,\mathrm{d}t \geq \max(0, x)$,



(a) pdf (b) cdf (c) integrated cdf

Figure 2.1: Distributions

(ii) $\int_{-\infty}^{x} K_2(t)\,\mathrm{d}t = \frac{\sqrt{3}}{\pi}\log\left(1+e^{\frac{\pi x\sqrt{3}}{3}}\right)$,

(iii) $\int_{-\infty}^{x} K_3(t)\,\mathrm{d}t = x\,\Phi(x)+\varphi(x)$ *and*

(iv) $\int_{-\infty}^{x} K_4(t)\,\mathrm{d}t = \frac{1}{4}\left(\sqrt{2}\exp(-\sqrt{2}|x|)+2(x+|x|)\right)$.

Proof. The equality in (i) follows by integration by parts. For the inequality recall that for X with density k it holds that

$$\begin{aligned} 0 = \mathbb{E}\,X &= -\int_{-\infty}^{0} K(u)\,\mathrm{d}u + \int_{0}^{\infty} 1-K(u)\,\mathrm{d}u \\ &\geq -\int_{-\infty}^{0} K(u)\,\mathrm{d}u + \int_{0}^{x} 1-K(u)\,\mathrm{d}u \\ &= x - \int_{-\infty}^{x} K(u)\,\mathrm{d}u \end{aligned}$$

and thus the assertion. □

Law of Large Numbers

3

All shall be well, and all shall be well,
and all matter of things shall be well.

Julian of Norwich, 1342 – 1416

3.1 WEAK LAW OF LARGE NUMBERS

Proposition 3.1. *Let X, X_i be uncorrelated (not necessarily independent) with $\mathbb{E}\,X = \mathbb{E}\,X_i = \mu$ and $\operatorname{var} X_i \le \sigma^2 < \infty$. Then*

$$P\left(\left|\overline{X}_n - \mu\right| < \varepsilon\right) \xrightarrow[n\to\infty]{} 1$$

for every $\varepsilon > 0$, i.e.,

$$\overline{X}_n \to \mathbb{E}\,X \text{ in probability.}$$

Proof. Note, that $\mathbb{E}\,\overline{X}_n = \mu$ and $\operatorname{var} \overline{X}_n \le \sigma^2/n$. By the Chebyshev inequality, for all $\varepsilon > 0$,

$$P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \le \frac{1}{\varepsilon^2}\,\mathbb{E}\left|\overline{X}_n - \mu\right|^2 \le \frac{\sigma^2}{n\,\varepsilon^2} \xrightarrow[n\to\infty]{} 0,$$

the assertion. □

3.2 HOEFFDING

Lemma 3.2 (Hoeffdings Lemma[1]). *Let $X \in \mathbb{R}$ be a random variable with $\mathbb{E}\,X = 0$ and $X \in [a, b]$ a.s. Then,*

$$\mathbb{E}\,e^{s\,X} \le \exp\left(\frac{s^2(b-a)^2}{8}\right), \qquad s \in \mathbb{R}.$$

Proof. As $x \mapsto e^{sx}$ is convex it follows that

$$e^{s\,x} \le \frac{b-x}{b-a} e^{s\,a} + \frac{x-a}{b-a} e^{s\,b}, \qquad x \in [a, b],$$

[1] Wassily Hoeffding, 1914–1991, Finnish statistician and probabilist

by taking expectations

$$\begin{aligned}\mathbb{E}\, e^{s X} &\le \frac{b}{b-a} e^{s\,a} - \frac{a}{b-a} e^{s\,b}, \\ &= (1-p) e^{s\,a} + p\, e^{s\,b} \\ &= \Big((1-p) + p\, e^{s(b-a)}\Big)\, e^{s\,a} \\ &= e^{\varphi\left(s\cdot(b-a)\right)},\end{aligned} \tag{3.1}$$

where

$$\begin{aligned} p &:= \frac{-a}{b-a} \text{ (recall that } a<0) \text{ and} \\ \varphi(h) &:= \log\left(1-p+p\,e^h\right) - h\cdot p. \end{aligned} \tag{3.2}$$

Observe that

$$\varphi'(h) = \frac{pe^h}{1-p+pe^h} - p$$

so that $\varphi(0) = \varphi'(0) = 0$ and

$$\varphi''(h) = \frac{e^h\cdot(1-p)p}{\left(1+(e^h-1)p\right)^2} = \frac{pe^h}{1-p+pe^h}\left(1-\frac{pe^h}{1-p+pe^h}\right) = \tilde p\,(1-\tilde p) \le \frac{1}{4},$$

with $\tilde p := \frac{pe^h}{1-p+pe^h} \in [0,1]$. By Taylor series expansion it follows that $\varphi(h) \le \frac{h^2}{8}$. Finally choose $h := s\cdot(b-a)$ and observe that $\varphi(h) \le \frac{h^2}{8} = \frac{s^2(b-a)^2}{8}$ thus (3.1), which is the assertion. □

Theorem 3.3 (Hoeffdings inequality)**.** *Let X_i be independent and bounded by $X_i \in [a_i, b_i]$ almost surely. Then, for $S_n := X_1 + \dots + X_n$ and $t>0$,*

$$P\left(S_n - \mathbb{E}\, S_n \ge t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right). \tag{3.3}$$

Proof. With Markov's inequality and $s>0$, $t>0$ we have that

$$\begin{aligned} P\left(S_n - \mathbb{E}\, S_n \ge t\right) &= P\left(e^{s(S_n-\mathbb{E}\, S_n)} \ge e^{s\,t}\right) \\ &\le \frac{1}{e^{s\,t}}\, \mathbb{E}\, e^{s(S_n-\mathbb{E}\, S_n)} \\ &= e^{-s\,t}\prod_{i=1}^n \mathbb{E}\, e^{s(X_i-\mathbb{E}\, X_i)} \\ &\le e^{-s\,t}\prod_{i=1}^n e^{\frac{s^2(b_i-a_i)^2}{8}} \\ &= \exp\left(-s\,t + \frac{s^2}{8}\sum_{i=1}^n (b_i-a_i)^2\right). \end{aligned}$$

Choose $s := \frac{4t}{\sum_{i=1}^n (b_i-a_i)^2}$ (the minimizer with respect to s) to get the assertion, i.e.,

$$P\left(S_n - \mathbb{E}\, S_n \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

□

Corollary 3.4. *Let X_i be independent and bounded by $X_i \in [a, b]$ almost surely with $\mu := \mathbb{E}\, X_i$. Then*

$$P\left(\overline{X}_n - \mu \geq t\right) \leq \exp\left(-n \cdot \frac{2t^2}{(b-a)^2}\right)$$

and

$$P\left(\left|\overline{X}_n - \mu\right| \geq t\right) \leq 2\exp\left(-n \cdot \frac{2t^2}{(b-a)^2}\right) \tag{3.4}$$

Proof. Replace $t \leftarrow t \cdot n$ in (3.3); apply (3.3) to $X_i \leftarrow -X_i$. □

Corollary 3.5. *Let $X_i \sim bin(n, p)$ be independent. Then*

$$P\left(\left|\overline{X}_n - \mu\right| \leq \sqrt{\frac{1}{2n}\log\frac{2}{\eta}}\right) \geq 1 - \eta$$

or, with $H_n := \sum_{i=1}^n X_i$,

$$P\left(H_n - n\,p \geq \varepsilon\, n\right) \leq e^{-2n\varepsilon^2}.$$

Proof. Invert (3.4) (i.e., $\eta = 2e^{-2n\varepsilon^2}$) and choose $t := n\,\varepsilon$ in (3.3). □

3.3 EXPONENTIAL BOUNDS AND LARGE DEVIATION THEORY

This exposition follows Shapiro et al. [14, Section 7.2.9].

Let X_i, be iid, then it holds for $t > 0$ by employing the Chebyshev inequality that

$$P\left(\overline{X}_n \geq a\right) = P\left(e^{t\,\overline{X}_n} \geq e^{t\,a}\right) \leq \frac{1}{e^{t\,a}}\,\mathbb{E}\, e^{t\,\overline{X}_n} = e^{-t\,a}\, M_X\left(\frac{t}{n}\right)^n, \tag{3.5}$$

where $M_X(s) := \mathbb{E}\, e^{s\,X}$ is the *moment generating function* of X.

Suppose that $a > \mu := \mathbb{E}\, X_i$. By taking logarithms in (3.5) we find that

$$\log P\left(\overline{X}_n \geq a\right) \leq -t\,a + n\log M_X\left(\frac{t}{n}\right) = -t\,a + n\,K_X\left(\frac{t}{n}\right),$$

where $K_X(\cdot) := \log M_X(\cdot)$ is the *cumulant generating function* of X. It follows that

$$\frac{1}{n}\log P\left(\overline{X}_n \geq a\right) \leq \inf_{t>0}\left\{-\frac{t}{n}\cdot a + K_X\left(\frac{t}{n}\right)\right\} = -\sup_{t>0}\left\{t\,a - K_X(t)\right\} = -K_X^*(a),$$

where

$$K^*(z) := \sup_{t>0}\left\{t\,z - K(t)\right\} \tag{3.6}$$

is the *convex conjugate* function. In large deviation theory, the function K_X^* is also called the *(large deviations) rate* function. Note that it follows that

$$P\left(\overline{X}_n \geq a\right) \leq e^{-n\cdot K_X^*(a)} \qquad (a > \mu). \tag{3.7}$$

The inequality (3.7) corresponds to the upper bound of Cramér's large deviation theory.

3.4 PROBLEMS

Exercise 3.1. *Show that the optimal t^* in (3.6) satisfies $z = \frac{\mathbb{E}\, X e^{t^* X}}{\mathbb{E}\, e^{t^* X}}$.*

Exercise 3.2. *The moment generating function of a distribution $X \sim$ bin$(1, p)$ is $\mathbb{E}\, e^{t\,X} = 1 - p + p\,e^t$ (compare with (3.2)). Show that the optimal t^* is $t^* = \log \frac{(1-p)z}{p(1-z)}$ and the rate function is*

$$\begin{aligned} K^*(z) &= z \log \frac{(1-p)z}{p(1-z)} - \log\left(1 - p + \frac{(1-p)z}{1-z}\right) \\ &= z \log \frac{z}{p} + (1-z) \log \frac{1-z}{1-p}. \end{aligned}$$

Exercise 3.3. *The moment generating function of a normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ is $M_X(t) = e^{\mu\,t + \frac{1}{2}\sigma^2 t^2}$. Show that the rate function is $K^*(z) := \frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2$. Show as well that this rate is exact in (3.7).*

Exercise 3.4. *Show that the conjugate of $K(t) = \frac{1}{p} t^p$ is $K^*(z) = \frac{1}{q} z^q$, where $\frac{1}{p} + \frac{1}{q} = 1$.*

Sampling techniques, synthetic data

4

> Keinerlei Mystik; Mathematik genügt mir.
>
> Max Frisch, 1911–1991, in Homer Faber

4.1 GENERATION OF RANDOM VARIABLES

4.1.1 The Inverse Transform Method

Definition 4.1 (Uniform distribution)**.** Suppose that $\mathrm{vol}(A) < \infty$. A random variable U is *uniformly distributed* on A (denoted $U \sim \mathcal{U}(A)$, if $P(U \in B) = \frac{\mathrm{vol}(B\cap A)}{\mathrm{vol}(A)}$ for every measurable set B.

Remark 4.2. For a random variable $U \sim \mathcal{U}[0,1]$, it holds that $P(U \le u) = u$ $(u \in [0,1])$.

A random variable X with distribution function F_X often can be obtained by using the inverse transform method. For a univariate, continuous random variable it holds that

$$X \sim F_X^{-1}(U),$$

where U is in $[0,1]$ uniformly distributed. Indeed, we have that

$$F_{F_X^{-1}(U)}(x) = P\big(F_X^{-1}(U) \le x\big) = P\big(U \le F_X(x)\big) = F_X(x), \tag{4.1}$$

and

$$F_{F_X(X)}(u) = P\left(F_X(X) \le u\right) = P\big(X \le F_X^{-1}(u)\big) = F_X\big(F_X^{-1}(u)\big) = u = F_U(u). \tag{4.2}$$

It follows from (4.1) that $F_X^{-1}(U)$ has the same cdf as X, i.e., they cannot be distinguished by their distribution function; as well, $F_X(X)$ and U share the same cdf (cf. (4.2)).

Remark 4.3. Let $U_i([0,1])$ be independent uniforms on the interval $[0,1]$ and $a_i < b_i$ for $i = 1,\dots,d$. Then

$$\begin{pmatrix} a_1 + (b_1 - a_1)U_1 \\ \vdots \\ a_d + (b_d - a_d)U_d \end{pmatrix} \in \mathbb{R}^d \tag{4.3}$$

is uniformly distributed in the rectangle

$$R \coloneqq [a_1, b_1] \times \cdots \times [a_d, b_d]\,. \tag{4.4}$$

Indeed, $P\big(a+(b-a)U\le x\big)=P\left(U\le\frac{x-a}{b-a}\right)=\frac{x-a}{b-a}$ (cf. Remark 4.2), the assertion for $d=1$. For independent U_i, $i=1,\dots,d$,

$$\begin{aligned}P\big(a_i+(b_i-a_i)U_i\le x_i \text{ for } i=1,\dots,d\big)&=\prod_{i=1}^{d}P\big(a_i+(b_i-a_i)U_i\le x_i\big)\\&=\prod_{i=1}^{d}\frac{x_i-a_i}{b_i-a_i}=\frac{\operatorname{vol}([a_1,x_1]\times\dots\times[a_d,x_d])}{\operatorname{vol}([a_1,b_1]\times\dots\times[a_d,b_d])},\end{aligned}$$

the assertion for any rectangle in general dimension d.

Algorithm 1 provides realizations of a random variable $U\sim\mathcal{U}(A)$ for a general set A. Its probability of acceptance is $\frac{\operatorname{vol}(A)}{\operatorname{vol}(R)}$.

Data: A set A with $A\subset R$, where R is a rectangle (cf. (4.4))

Result: Realization of a random variable $U\sim\mathcal{U}(A)$
repeat
  generate a random variable $Y\sim\mathcal{U}(R)$, cf. (4.3)
until $Y\in A$;
return $U:=Y$

Algorithm 1: Realization of a uniform $U\sim\mathcal{U}(A)$ (rejection sampling)

4.1.2 Rejection sampling, acceptance-rejection method — Verwerfungsmethode

Suppose that it is cheap to sample from the multivariate distribution with density $g(\cdot)$ (the proposal distribution) and there is a number $\alpha>1$ such that $f_X(x)\le\alpha\cdot g(x)$ for all $x\in\mathbb{R}^d$. Algorithm 2 describes the method of rejection sampling.

Data: A density function $g(\cdot)$ and $\alpha>1$ so that $f_X(\cdot)\le\alpha\, g(\cdot)$

Result: Realization of a random variable X with density $f_X(\cdot)$
repeat
  generate a random variable Y with density $g(\cdot)$ and
  an independent, uniform $U\in[0,1]$
until $f_X(Y)\ge U\,\alpha\, g(Y)$ *accept* Y;
return $X:=Y$

Algorithm 2: Rejection sampling

Verification of Algorithm 2. Note that

$$P\left(Y \text{ accepted and } Y\in \mathrm{d}x\right)=P\left(U\le\frac{f_X(x)}{\alpha\cdot g(x)} \text{ and } Y\in\mathrm{d}x\right)=\frac{f_X(x)}{\alpha\cdot g(x)}\cdot g(x)\,\mathrm{d}x=\frac{1}{\alpha}f_X(x)\,\mathrm{d}x. \tag{4.5}$$

By integrating all $\mathrm{d}x$ we find the efficiency

$$P\left(Y \text{ accepted}\right) = \int_{\mathbb{R}^d} \frac{1}{\alpha} f_X(x)\,\mathrm{d}x = \frac{1}{\alpha}.$$

It follows that $P(X \in \mathrm{d}x) = P\left(Y \in \mathrm{d}x \mid Y \text{ accepted}\right) = \frac{P(Y \in \mathrm{d}x \text{ and } Y \text{ accepted})}{P(Y \text{ accepted})} = f_X(x)\,\mathrm{d}x$, the assertion. □

4.1.3 Ratio-of-uniforms method

The ratio-of-uniforms method is a variant of rejection sampling to obtain samples from a distribution with given density. The key advantage of the ratio-of-uniforms method is that only *uniform* random variables (and no others) have to be accessible. Basis of the ratio-of-uniforms method is the following:

Theorem 4.4 (cf. Kinderman and Monahan, 1977). *Let $h(\cdot)$ be a function with $\int_{\mathbb{R}^d} h(y)\,\mathrm{d}y < \infty$ and $r > 0$. The volume of*

$$\mathcal{A} := \left\{(v,u) \in \mathbb{R}^d \times \mathbb{R} \colon 0 < u \le \sqrt[rd+1]{h(v/u^r)}\right\} \tag{4.6}$$

is finite. If (V,U) is uniformly distributed in $\mathcal{A}$, then $X := V/U^r = (V_1,\dots,V_d)/U^r \in \mathbb{R}^d$ is a random vector with probability density function $f_X(x) := h(x)/\int_{\mathbb{R}^d} h(y)\,\mathrm{d}y$ (cf. Algorithm 3).

Verification of Theorem 4.4 and Algorithm 3. We shall apply the *change of variables formula*, $\int_{\mathcal{A}} f(y)\mathrm{d}y = \int_{g(\mathcal{A})} f(g^{-1}(x))|(g^{-1})'(x)|\,\mathrm{d}x$. The transformation $g\begin{pmatrix} v_1\\ \vdots\\ v_d\\ u\end{pmatrix} := \begin{pmatrix} v_1/u^r\\ \vdots\\ v_d/u^r\\ u\end{pmatrix}$ with inverse $g^{-1}\begin{pmatrix} x_1\\ \vdots\\ x_d\\ y\end{pmatrix} = \begin{pmatrix} x_1\cdot y^r\\ \vdots\\ x_d\cdot y^r\\ y\end{pmatrix}$ has Jacobian $\det(g^{-1})'\begin{pmatrix} x_1\\ \vdots\\ x_d\\ y\end{pmatrix} = \det\begin{pmatrix} y^r & \ddots & \vdots & x_1\\ 0 & \ddots & 0 & \vdots\\ \vdots & \ddots & y^r & x_d\\ 0 & \dots & 0 & 1\end{pmatrix} = y^{rd}$ and $g(\mathcal{A}) = \left\{(x,y) \in \mathbb{R}^d \times \mathbb{R} \colon 0 < y \le \sqrt[rd+1]{h(x)}\right\}$. The volume of $\mathcal{A}$ is finite, as

$$\begin{aligned}
\operatorname{vol}(\mathcal{A}) &= \int_{\mathcal{A}} 1\,\mathrm{d}u\,\mathrm{d}v_1\dots\mathrm{d}v_d\\
&= \int_{g(\mathcal{A})} y^{rd}\,\mathrm{d}y\,\mathrm{d}x_1\dots\mathrm{d}x_d\\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{y^{rd+1}}{rd+1}\bigg|_{y=0}^{\sqrt[rd+1]{h(x)}} \mathrm{d}x_1\dots\mathrm{d}x_d\\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{h(x)}{rd+1}\mathrm{d}x_1\dots\mathrm{d}x_d < \infty.
\end{aligned} \tag{4.7}$$

The random variable V/U^r are the first d marginals of $g(V,U)$. The marginal density is

$$f_{V/U^r}(x) = \int_0^\infty f_{g(V,U)}(x,y)\,\mathrm{d}y = \int_0^\infty f_{V,U}\left(g^{-1}\begin{pmatrix}x\\y\end{pmatrix}\right)\cdot y^{rd}\,\mathrm{d}y = \int_0^\infty f_{V,U}\begin{pmatrix}xy^r\\y\end{pmatrix}\cdot y^{rd}\,\mathrm{d}y.$$

By design of Algorithm 3, the random vector (V,U) is uniformly distributed in $\mathcal{A}$, so the joint density is

$$f_{V,U}(v,u) = \begin{cases}\frac{1}{\mathrm{vol}(\mathcal{A})} & \text{if } (v,u)\in\mathcal{A},\\ 0 & \text{else,}\end{cases}$$

that is, $f_{V,U}\begin{pmatrix}xy^r\\y\end{pmatrix} = \begin{cases}\frac{1}{\mathrm{vol}(\mathcal{A})} & \text{if } 0\le y\le \sqrt[rd+1]{h(x)},\\ 0 & \text{else.}\end{cases}$ With (4.7), the marginal density is

$$f_{V/U^r}(x) = \int_0^{\sqrt[rd+1]{h(x)}} \frac{y^{rd}}{\mathrm{vol}(\mathcal{A})}\,\mathrm{d}y = \frac{1}{\mathrm{vol}(\mathcal{A})}\,\frac{y^{rd+1}}{rd+1}\bigg|_{y=0}^{\sqrt[rd+1]{h(x)}} = \frac{h(x)}{\int_{\mathbb{R}^d} h(y)\mathrm{d}y}$$

for every $x\in\mathbb{R}^d$. □

Algorithm 3 employs rejection sampling (Algorithm 1) to find uniform points in (4.6) $\subseteq \mathcal{R}$ for a suitable region $\mathcal{R}\subseteq\mathbb{R}^d\times\mathbb{R}$.

Data: A nonnegative function $h(\cdot)$ and a region $\mathcal{R}$ with finite volume containing $\mathcal{A}$, cf. (4.6) (cf. Remark 4.5); a parameter $r>0$

Result: Realization of a random variable X with density $f_X(\cdot) = h(\cdot)/\int_{\mathbb{R}^d} h(y)\mathrm{d}y$

repeat
  generate a random point (V,U) uniformly distribted in $\mathcal{R}$,
  set $Y := V/U^r$; ratio of uniforms
until $U^{rd+1}\le h(Y)$ *reject* Y;
set $X := Y$; accept Y
return X

Algorithm 3: Ratio-of-uniforms method

Remark 4.5. Observe that $u\le \sup_x \sqrt[d+1]{h(x)}$; further, with $x_i := {}^{v_i}\!/_u$, the constraint $u\le \sqrt[d+1]{h(v/u)}$ is equivalent to $v_i\le x_i\cdot\sqrt[d+1]{h(x)}$. For implementations it is thus sufficient (cf. Exercise 4.3) and often convenient to choose the rectangle

$$\mathcal{R} := \underbrace{\cdots\times\Bigg[\underbrace{\inf_{x\in\mathcal{S}} x_i\cdot\sqrt[d+1]{h(x)}}_{=:x_{\ell,i}},\ \underbrace{\sup_{x\in\mathcal{S}} x_i\cdot\sqrt[d+1]{h(x)}}_{=:x_{r,i}}\Bigg]\times\dots}_{\ni V}\quad\times\quad\underbrace{\Big[0,\ \sup_{x\in\mathcal{S}}\sqrt[d+1]{h(x)}\Big]}_{\ni U}\supset\mathcal{A}. \tag{4.8}$$

Remark 4.6. Exercise 4.2 is a remarkable example of how to employ Algorithm 3 to generate variates of a Cauchy distribution.

4.1.4 Composition method

Proposition 4.7. *Suppose that P_j are probability measures and π_j are* mixing coefficients *with $\pi_j \geq 0$ and $\sum_{j=1}^n \pi_j = 1$.*

Let $X_j \sim P_j$ and let $j^* \in \{1, \dots, n\}$ be a random variable with $P(j^* = j) = \pi_j$, then X_{j^*} has measure

$$X_{j^*} \sim \sum_{j=1}^n \pi_j \cdot P_j =: P.$$

Proof. From Bayes' theorem we have that

$$\begin{aligned} P(X_{j^*} \in A) &= \sum_{j=1}^n P(X_{j^*} \in A \mid j^* = j) \cdot P(j^* = j) \\ &= \sum_{j=1}^n P_j(X_j \in A) \cdot P(j^* = j) \\ &= \sum_{j=1}^n \pi_j \cdot P_j(X_j \in A) \end{aligned}$$

and thus the assertion. □

Corollary 4.8. *Suppose that $f_j(\cdot)$ are density functions and π_j are mixing coefficients with $\pi_j \geq 0$ and $\sum_{j=1}^n \pi_j = 1$.*

Let X_j have density $f_j(\cdot)$ and let j^* be a random variable with $P(j^* = j) = \pi_j$, then X_{j^*} has density

$$f_{X_{j^*}}(\cdot) \sim \sum_{j=1}^n \pi_j \cdot f_j(\cdot).$$

4.2 METROPOLIS–HASTINGS

The Metropolis[1]–Hastings[2] algorithm is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of random samples from a probability distribution from which direct sampling is difficult.

Consider a Markov chain where transitions from y to $\mathrm{d}x$ happen with probability $q(\mathrm{d}x|\, y)$. Note, that $\int q(\mathrm{d}x|\, y) = 1$ for every y. Given a measure with density p_m, the subsequent density is $p_{m+1}(x) = \int q(x|\, y)\, p_m(y)\, \mathrm{d}y$.

Definition 4.9. A Markov chain is *stationary* with distribution $p(x)$, if $p(x) = \int q(x|\, y)\, p(y)\, \mathrm{d}y$.

Remark 4.10 (Random walk). A simple example of a Markov chain is the *random walk*, where $q(\cdot|\, y) \sim \mathcal{N}(y, \Sigma_0)$ for some (fixed) covariance Σ_0.

[1] Nicolas Metropolis, 1919–1999, Greek-American physicist
[2] Wilfried Keith Hastings, 1930–2016, statistician

Definition 4.11 (Detailed balance)**.** A Markov chain is said to be *reversible* or *detailed balance*, if there is a probability measure with density p so that $p(x)\,q(y|\,x) = p(y)\,q(x|\,y)$.

Proposition 4.12. *Suppose that a Markov chain is reversible, then it has a stationary distribution.*

Proof. By definition there is a density p so that $p(x)\,q(y|\,x) = p(y)\cdot q(x|\,y)$. It holds that

$$\int q(x|\,y)\,p(y)\,\mathrm{d}y = \int q(y|\,x)\,p(x)\,\mathrm{d}y = p(x)\cdot\int q(y|\,x)\,\mathrm{d}y = p(x),$$

thus p is stationary. □

Remark 4.13*.* Uniqueness of a stationary distribution can be ensured by assuming ergodicity of the Markov chain.

Data: A (unnormalized) density function $\tilde{p}(\cdot)$ and a transition kernel $q(\cdot|\cdot)$
Result: A (possibly correlated) sequence of random variables X_k with density $p(\cdot) = c_{\tilde{p}}\cdot\tilde{p}(\cdot)$

set $k := 0$ and pick an initial value X_0
repeat
  generate a candidate $Y \sim q(\cdot \mid X_k)$,
  compute the Metropolis acceptance ratio

$$A(Y, X_k) := \min\left(1,\ \frac{\tilde{p}(Y)\cdot q(X_k|\,Y)}{\tilde{p}(X_k)\cdot q(Y|\,X_k)}\right), \tag{4.9}$$

  generate an independent uniform $U \in [0,1]$
  if $U \le A(Y, X_k)$ **then**
    set $X_{k+1} = Y$ — accept the candidate
  else
    set $X_{k+1} = X_k$ — reject and copy the old state forward
  end
  set $k = k+1$
until *tired of all this*;

Algorithm 4: Metropolis–Hastings algorithm

The Metropolis–Hastings algorithm (Algorithm 4) generates a sequence of samples from a measure P with associated density $p(x)\,\mathrm{d}x = P(\mathrm{d}x)$, which are (in general) correlated and particularly *not* independent.

Remark 4.14*.* The Metropolis–Hastings algorithm employs the unnormalized density function $\tilde{p}$ instead of the density p. Due to (4.9), the constant $c_{\tilde{p}}^{-1} = \int \tilde{p}(x)\,\mathrm{d}x$ does not have to be known.

Proposition 4.15. *The sequence generated by the Metropolis–Hastings algorithm (Algorithm 4) is detailed balance with stationary distribution $p(\cdot)$.*

Proof. It is apparent that the algorithm defines a Markov process with transition probabilities $q(y|x)\,A(y,x)$. With (4.9) we have that

$$\begin{aligned} p(x)\,q(y|x)\cdot A(y,x) &= \min\left(p(x)\,q(y|x),\ p(y)\,q(x|y)\right)\\ &= \min\left(p(y)\,q(x|y),\ p(x)\,q(y|x)\right)\\ &= p(y)\,q(x|y)\cdot A(x,y). \end{aligned}$$

It follows that $p(\cdot)$ is reversible (detailed balance) and stationary by Proposition 4.12. □

4.3 IMPORTANCE SAMPLING

We have seen in the preceding section that $\frac{1}{n}\sum_{i=1}^n h(X_i) \xrightarrow[n\to\infty]{} \mathbb{E}\,h = \int h\,\mathrm{d}P$ for independent samples X_i chosen from P. I.e., for a density with $f(x)\,\mathrm{d}x = P(\mathrm{d}x)$ we have convergence of the sample means towards its P-expectation, $\frac{1}{n}\sum_{i=1}^n h(X_i) \xrightarrow[n\to\infty]{} \int h\,\mathrm{d}P = \int h(x)\cdot f(x)\,\mathrm{d}x$.

Suppose that it is difficult to sample from P, but samples from a different measure $Q \gg P$ (the proposal distribution) are cheaply/ easily available. Let Q have density function $g(\cdot)$ and let ξ_i be independent samples from Q. Then

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n h(\xi_i)\frac{f(\xi_i)}{g(\xi_i)} &\xrightarrow[n\to\infty]{} \int h(x)\frac{f(x)}{g(x)}\cdot g(x)\,\mathrm{d}x\\ &= \int h(x)\,f(x)\,\mathrm{d}x\\ &= \int h\,\mathrm{d}P, \end{aligned}$$

i.e., the expectation of h with respect to P can be realized by employing samples from Q and the *likelihood ratio* $R(x) := \frac{g(x)}{f(x)}$.

Note that in contrast to rejection sampling (Algorithm 2 above), importance sampling does *not* discard samples. Instead, the method adjusts the weights (giving thus rise to the name *importance*).

Remark 4.16. For the method to be efficient in practice it is desirable that $R(\cdot) \approx 1$, or even better if $\frac{h(\cdot)}{R(\cdot)} = h(\cdot)\frac{f(\cdot)}{g(\cdot)} \approx$ const. For nonnegative f, the probability density $g(\cdot) := h(\cdot)\cdot f(\cdot)$ is particularly useful.

4.4 PROBLEMS

Exercise 4.1. *Show that the expectation* $\mathbb{E}\,U = \frac{1}{2}(b-a)$ *and variance* $\operatorname{var} U = \frac{1}{12}(b-a)^2$ *of the distribution* $U \sim \mathcal{U}([a,b])$.

Exercise 4.2. *Let* $(U,V) \in \mathcal{R} = \left\{(u,v)\colon u^2+v^2 \le 1\right\}$ *be uniformly distributed. Choose* $h(x) := \frac{1}{1+x^2}$ *and show that* $U/V \sim$ *Cauchy by employing Algorithm 3.*

Exercise 4.3 (Ratio-of-uniforms)**.** *Verify that* (4.6) $\subseteq$ (4.8) $= \mathcal{R}$*, i.e,* $\left\{(u, v) \colon 0 \le u \le \sqrt{h(v/u)}\right\} \subset \left[0,\, \sup_x \sqrt{h(x)}\right] \times \left[-\sup_x \sqrt{x\, h(x)},\, \sup_x \sqrt{x\, h(x)}\right]$.

Exercise 4.4. *Generate variates of a Gamma distribution using the ratio-of-uniforms, Algorithm 3.*

Exercise 4.5. *Discuss and verify the https://www.tu-chemnitz.de/mathematik/fima/public/mathematischeS expectation in (4.5)*

5 Gaussian Distributions

See the Section on *Gaussian distributions* (normal distribution) in the lecture mathematische Statistik.

6 Gaussian processes

6.1 RANDOM FUNCTIONS

Consider a family of functions, often called the *feature maps*, $\varphi_k\colon \mathcal{X} \to \mathbb{R}$, and a sequence $\sigma_k \in \mathbb{R}$, $k = 1, 2, \dots$

Remark 6.1. Note that the realization of the random variable $f\colon \Omega \to \mathbb{R}^{\mathcal{X}}$ is the function $f(\omega)\colon \mathcal{X} \to \mathbb{R}$. We will always have that $\mathcal{X} = \mathbb{R}^d$.

Theorem 6.2 (Random fields)**.** *Let ξ_k be uncorrelated random variables with $\mathbb{E}\,\xi_k = 0$, $\operatorname{var}\xi_k = 1$ and define the* random function *(stochastic process)*

$$\big(f(\omega)\big)(x) \coloneqq \sum_{k=1} \xi_k(\omega)\,\sigma_k\,\varphi_k(x), \qquad x \in \mathcal{X},$$

usually written as random function

$$f(x) = \sum_{k=1} \xi_k\,\sigma\,\varphi_k(x), \qquad x \in \mathcal{X}. \tag{6.1}$$

Then $\mathbb{E}\,f(x) = 0$ and the covariance is

$$k(x, x') \coloneqq \operatorname{cov}\big(f(x),\ f(x')\big) = \sum_{k=1} \sigma_k^2\,\varphi_k(x)\,\varphi_k(x'), \qquad x, x' \in \mathcal{X}.$$

For $\xi_k \sim \mathcal{N}(0,1)$ it holds that

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} k(x_1,x_1) & \dots & k(x_1,x_n) \\ \vdots & & \vdots \\ k(x_n,x_1) & \dots & k(x_n,x_n) \end{pmatrix} \right) = \mathcal{N}(0,\ K), \tag{6.2}$$

where K with $K_{ij} = k(x_i, x_j)$ is the Gram matrix*. The vector f with components $f_i \coloneqq f(x_i)$ follows the multivariate normal distribution*

$$f \sim \mathcal{N}\big(0,\ K\big).$$

Remark 6.3. Suppose that $\xi_k \sim \mathcal{N}(0,1)$ are standard Gaussians, then

$$f(x) \sim \mathcal{N}\left(0,\ \sum_{k=1} \sigma_k^2\,\varphi_k(x)^2\right), \qquad x \in \mathcal{X}.$$

Proof. By linearity, the expectation is

$$\mathbb{E}\, f(x) = \mathbb{E} \sum_{k=1} \xi_k \sigma_k \varphi_k(x) = \sum_{k=1} \sigma_k \varphi_k(x)\ \mathbb{E}\, \xi_k = 0.$$

The covariance thus is

$$\begin{aligned}\operatorname{cov}\big(f(x), f(y)\big) &= \mathbb{E} \sum_{k=1} \xi_k \sigma_k \varphi_k(x) \cdot \sum_{\ell=1} \xi_\ell \sigma_\ell \varphi_\ell(y) \\ &= \sum_{k=1} \sigma_k \varphi_k(x) \cdot \sum_{\ell=1} \sigma_\ell \varphi_\ell(y) \cdot \mathbb{E}\, \xi_k\, \xi_\ell \\ &= \sum_{k=1} \sigma_k^2\, \varphi_k(x)\, \varphi_k(y),\end{aligned}$$

the assertion. □

6.2 GAUSSIAN PROCESSES

Consider a kernel function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a *Gaussian process* f, i.e., a random variable $f\colon \Omega \to \mathbb{R}^{\mathcal{X}}$ (with $\mathcal{X} = \mathbb{R}^d$, e.g.). Recall, that a realization of the random variable $f(\omega)\colon X \to \mathbb{R}$ is a function. For any collection of points $x_1, \dots, x_n \in \mathcal{X}$ it holds that that

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{pmatrix} \right) = \mathcal{N}(0,\, K),$$

where $K_{ij} = k(x_i, x_j)$ is the Gram matrix. The vector f with components $f_i \coloneqq f(x_i)$ follows the multivariate normal distribution

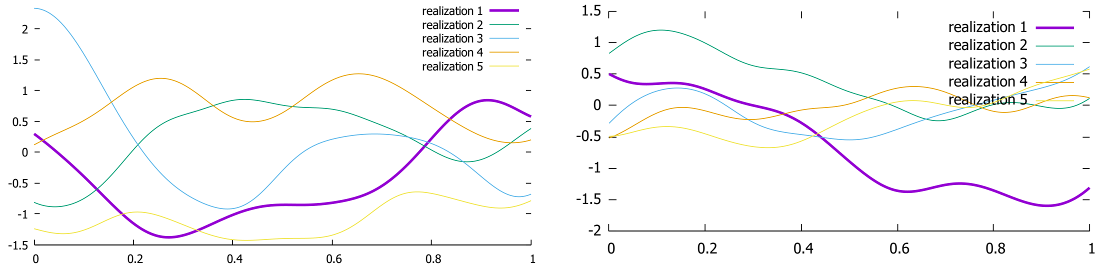
$$f \sim \mathcal{N}(0,\, K).$$

Example 6.4. Consider the exponentially weighted monomials $\varphi_k(x) = \left(\frac{x}{\ell}\right)^k e^{-\frac{1}{2}(x/\ell)^2}$ with $\sigma_k^2 = \frac{1}{k!}$. Then

$$\begin{aligned}k(x, x') &= \sum_{k=0} \frac{1}{k!} \left(\frac{x}{\ell}\right)^k \left(\frac{x'}{\ell}\right)^k e^{-\frac{1}{2}(x/\ell)^2} e^{-\frac{1}{2}(x'/\ell)^2} \\ &= e^{xx'/\ell^2} e^{-\frac{1}{2}(x/\ell)^2} e^{-\frac{1}{2}(x'/\ell)^2} = \exp\left(-\frac{1}{2}\left(\frac{x - x'}{\ell}\right)^2\right).\end{aligned}$$

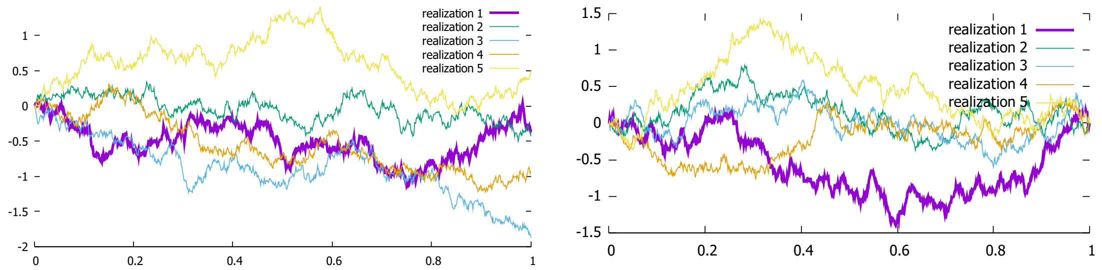
Example 6.5 (Brownian motion)**.** Consider the feature maps $\varphi_k(x) \coloneqq \sqrt{2}\, \sin\Big((k - \frac{1}{2})\pi x\Big)$, and $\sigma_k \coloneqq \frac{1}{(k-\frac{1}{2})\pi}$, then (cf. Figure 6.2a)

$$k(x, y) = \sum_{k=1} \sigma_k^2\, \varphi_k(x)\, \varphi_k(y) = \min(x, y).$$



(a) Exponential kernel, cf. Example 6.4 (b) Sigmoid kernel $k(x,y)=\frac{2}{\exp(\|x-y\|)+\exp(-\|x-y\|)}$

Figure 6.1: Random functions



(a) Brownian motion (b) Brownian bridge

Figure 6.2: Brownian motion and Brownian bridge

Example 6.6 (Brownian bridge)**.** Choose $\varphi_k(x) := \sqrt{2}\sin(k\pi x)$, $\sigma_k := \frac{1}{k\pi}$, then (cf. Figure 6.2b)

$$k(x,y)=\min(x,y)-x\,y=\sum_{k=1}\sigma_k^2\,\varphi_k(x)\,\varphi_k(y)$$

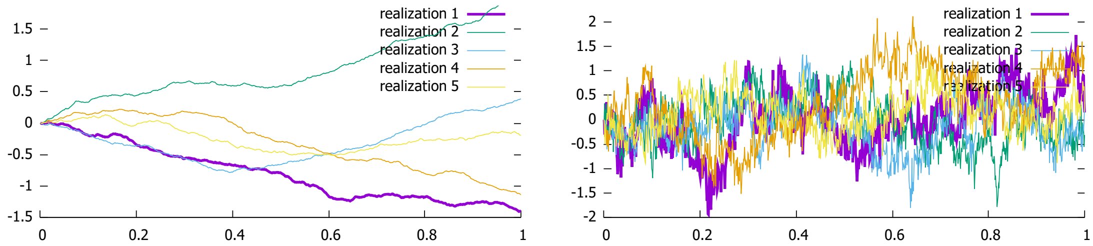
In what follows, we shall assume that there is a symmetric function $k(\cdot,:)$, but the feature functions are not available explicitly. Nonetheless, we can describe the functions.

Example 6.7 (Fractional Brownian motion)**.** The kernel function for the fractional Brownian motion is $2k(x,y)=x^{2H}+y^{2H}-|x-y|^{2H}$, where $H$is the Hurst index; the Wiener process has Hurst index $H=1/2$.

Popular choice for the kernel function include the Matérn $1/2$ kernel[1]

$$k(x,x')=\sigma_f^2\exp\left(-\frac{\|x-x'\|}{\sigma_\ell}\right) \tag{6.3}$$

[1]Bertil Matérn, 1917–2007, Swedish statistician



(a) Hurst index $H = 0.8$; increments are positively correlated

(b) Hurst index $H = 0.2$; increments are negatively correlated

Figure 6.3: Fractional Brownian motion

and the Matérn $3/2$ kernel[2]

$$k(x,x') = \sigma_f^2 \left(1 + \frac{\sqrt{3}\,\|x-x'\|}{\sigma_\ell}\right) \exp\left(-\frac{\sqrt{3}}{\sigma_\ell}\,\|x-x'\|\right). \tag{6.4}$$

Here, the parameter σ_f is called the *signal variance* and σ_ℓ is the *length scale*.

- The Laplace kernel or exponential kernel is

$$k(x,x') = \exp\left(-\frac{\|x-x'\|}{\sigma_\ell}\right);$$

  it is a special case ($\nu = 1/2$) of the following Matérn kernel.

- The general Matérn kernel is

$$k(x,x') = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\,\|x-x'\|}{\sigma_\ell}\right)^\nu \cdot K_\nu\left(\frac{\sqrt{2\nu}}{\sigma_\ell}\,\|x-x'\|\right),$$

  where K_ν is the modified Bessel function of the second kind. A Gaussian process with Matérn covariance is $\lceil \nu \rceil + 1$ times differentiable. For $\nu = k + \frac{1}{2}$ ($k \in \mathbb{N}$), the Matérn kernel simplifies to a polynomial $\times$ exponential function, as in (6.4).

- The squared exponential kernel,

$$k(x,x') = \sigma_f^2 \exp\left(-\frac{1}{2\sigma_\ell^2}\,\|x-x'\|^2\right),$$

  is the Matérn kernel with $\nu \to \infty$. The kernel parameters (σ_f, σ_ℓ, e.g.) and the parameter σ_ε can be estimated by maximizing the log-likelihood function, that is, by maximizing

$$-\frac{1}{2}\log\det\left(K_\vartheta + \sigma_\varepsilon^2 I\right) - \frac{1}{2} y^\top \left(K_\vartheta + \sigma_\varepsilon^2 I\right)^{-1} y$$

[2]Note, that $(1+x)e^{-x} \sim 1 - \frac{x^2}{2} + O(x^3)$

with respect to the parameters of the model ($(\underbrace{\sigma_\varepsilon, \sigma_f, \sigma_\ell}_{\vartheta})$, say).

▹ The inverse multiquadratic kernel (with parameter σ_ℓ) is

$$k(x, x') = \frac{\sigma_f^2}{\sqrt{1 + \frac{1}{2\sigma_\ell^2}\|x - x'\|^2}}.$$

Proposition 6.8. *Suppose that*

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{pmatrix}^{-1}\right).$$

Then the function

$$f(x) \coloneqq \sum_{i=1}^n w_i \cdot k(x, x_i) \tag{6.5}$$

has the distribution (6.2) as well.

Proof. Indeed, $\mathbb{E} f(x) = \sum_{i=1}^n k(x, x_i)\, \mathbb{E}\, w_i = 0$, and

$$\begin{aligned} \operatorname{cov}\big(f(x), f(x_\ell)\big) &= \sum_{i,j=1}^n k(x, x_i)\ \mathbb{E}\, w_i w_j\, k(x_j, x_\ell) \\ &= \sum_{i=1}^n k(x, x_i) \underbrace{\sum_{j=1}^n K_{ij}^{-1}\, k(x_j, x_\ell)}_{\delta_{i\ell}} \\ &= k(x, x_\ell), \end{aligned}$$

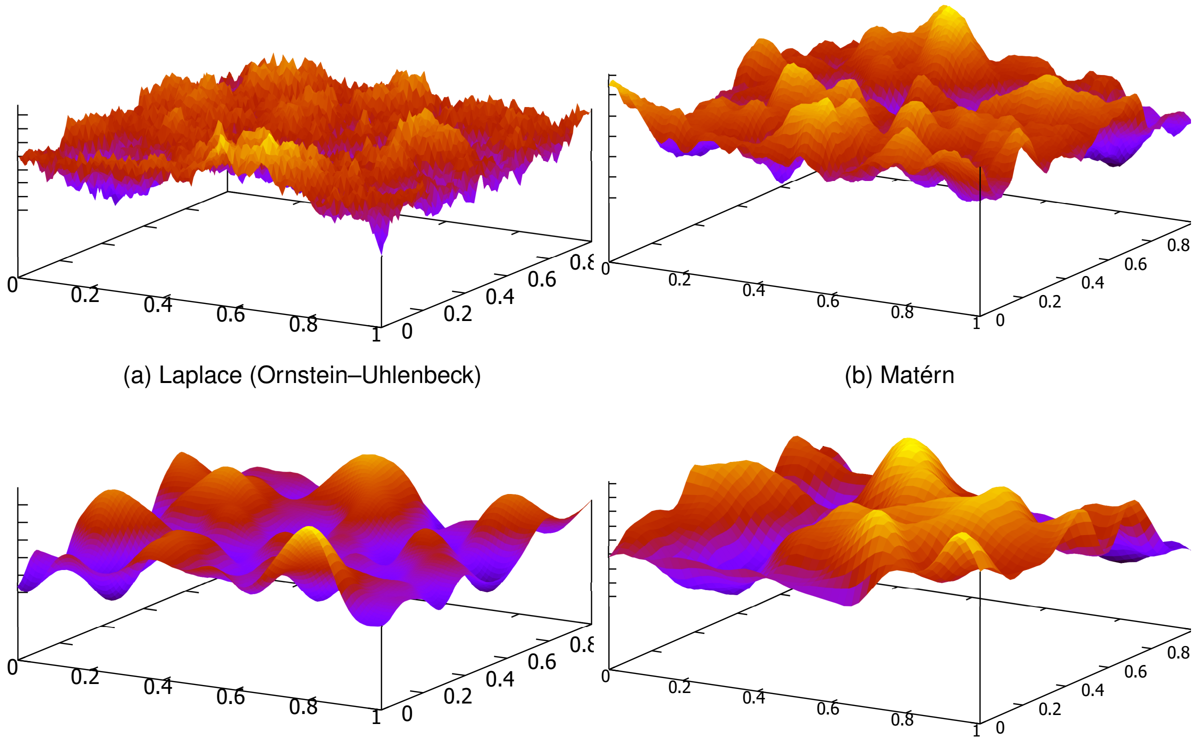
the assertion for $x = x_k$; for convenience, we have set $K \coloneqq \begin{pmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{pmatrix}$. □

The formula (6.5) gives access to the random function f as well.

6.3 GAUSSIAN PROCESS REGRESSION

Suppose the function values at $X = (x_1, \dots, x_n) \in \mathcal{X}^n$ are know ("*training*"), and we were interested in the function values at the new points $\hat{X} \coloneqq (\hat{x}_1, \dots, \hat{x}_m) \in \mathcal{X}^m$. They follow the "signal plus noise" paradigm

$$f_i = f_0(\hat{x}_i) + \varepsilon,$$



(a) Laplace (Ornstein–Uhlenbeck)

(b) Matérn

(c) Gaussian kernel, cf. Figure 6.1a

(d) Sigmoid kernel, cf. Figure 6.1b

Figure 6.4: Realization of two dimensional random function for different, radial kernels

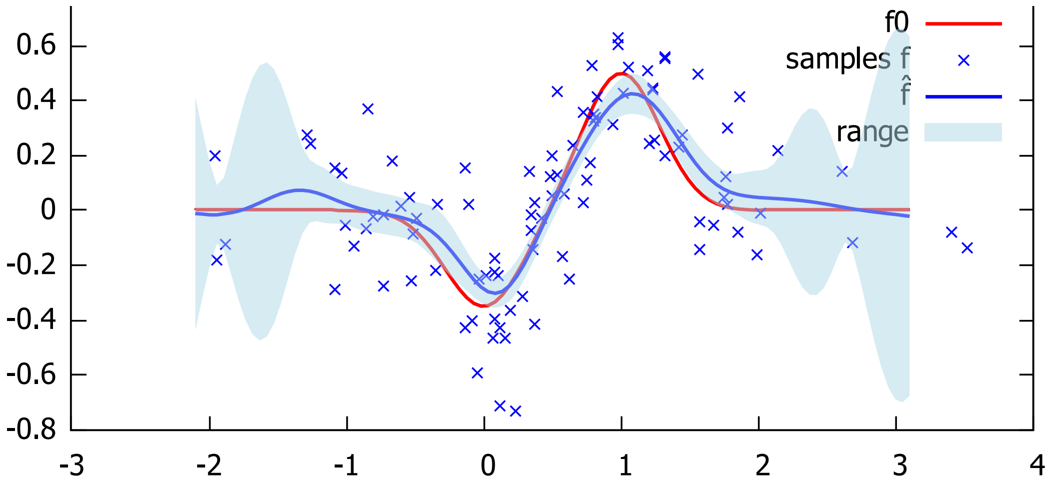


Figure 6.5: Prediction with random functions (6.7)

where $\varepsilon \sim \mathcal{N}(0,\Lambda)$ independent. The joint distribution is

$$\begin{pmatrix} f_0(\hat{X}) \\ f(X) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k(\hat{X},\hat{X}) & k(\hat{X},X) \\ k(X,\hat{X}) & k(X,X)+\Lambda \end{pmatrix} \right),$$

where $f(X) = (f_1,\dots,f_n)$ are the function values observed at $\hat{X}$, $f_0(\hat{X}) = \big(f_0(\hat{x}_0),\dots,f_0(\hat{x}_m)\big)$, $k(\hat{X},X) = \big(k(\hat{x}_i,x_j)\big)_{i,j=1}^{m,n}$, etc.

It follows from conditional Gaussians (cf. math. statistics, section Normal Distribution or Liptser and Shiryaev [9, Theorem 13.1]) that

$$f_0(\hat{X}) \mid f(X) \sim \mathcal{N}(\hat{\mu},\, \hat{K}),$$

where

$$\hat{\mu} := k(\hat{X},X)\,(k(X,X)+\Lambda)^{-1}\,f(X)$$

is the posterior estimator and

$$\hat{K} := k(\hat{X},\hat{X}) - k(\hat{X},X)\,(k(X,X)+\Lambda)^{-1}\,k(X,\hat{X}).$$

Now consider the special case $\tilde{X} = (x)$. Then the prediction is

$$f_0(x) = k(x,X)\,(k(X,X)+\Lambda)^{-1}\,f(X),$$

the local variance

$$\begin{aligned} &\operatorname{var}\big(f_0(x)\big|\, f(X_1)=f_1,\dots,f(X_n)=f_n\big) \\ &\qquad = k(x,x) - k(x,X)\,(k(X,X)+\Lambda)^{-1}\,k(X,x). \end{aligned} \tag{6.6}$$

does *not* depend on the samples f_i. Note that the variance decreases with additional information, $\operatorname{var}\big(f_0(x)\big|\, f(X)=f\big) \le k(x,x)$.

It is convenient to introduce the auxiliary quantity $w := (k(X,X)+\Lambda)^{-1} f(X)$, i.e.,

$$\lambda\, w_i + \sum_{j=1}^n k(x_i,x_j) w_j = f_i, \qquad i = 1,\dots,n.$$

Then the predicted value is

$$f_0(x) = \sum_{i=1}^n k(x,x_i)\, w_i. \tag{6.7}$$

Figure 6.5 provides an example for predicted function values together with the variance (6.6).

6.4 RECONSTRUCTION OF THE FEATURE FUNCTIONS

Consider the linear operator $Kf(x) := \int_{\mathcal{X}} k(x,y)\, f(y)\, \mathrm{d}y$ with eigenvectors and eigenvalues $K\varphi_k = \lambda_k \varphi_k$. Define the inner product $\langle g \mid f\rangle := \int_{\mathcal{X}} f(x)\, g(x)\, \mathrm{d}x$. Without loss of generality we may assume that $\langle \varphi_k \mid \varphi_k\rangle = 1$. For a symmetric and integrable kernel $k(x,y) = k(y,x)$ the operator K is self-adjoint and we have that there are only countably many eigenvalues, which are mutually orthogonal (i.e., for different eigenvalues). Indeed, $\lambda_\ell \langle \varphi_k \mid \varphi_\ell\rangle = \langle \varphi_k \mid K\varphi_\ell\rangle = \langle K\varphi_k \mid \varphi_\ell\rangle = \lambda_k \langle \varphi_k \mid \varphi_\ell\rangle$, i.e., $\langle \varphi_k \mid \varphi_\ell\rangle = 0$ if $\lambda_k \neq \lambda_\ell$.

Proposition 6.9 (Mercer)**.** *We have that*

$$k(x,x') = \sum_{k=1}^\infty \lambda_k\, \varphi_k(x)\, \varphi_k(x') = \operatorname{cov}\big(f(x),\, f(x')\big),$$

where f is as in (6.1).

Proof. Note that

$$\int_{\mathcal{X}} \sum_{k=1}^\infty \lambda_k \varphi_k(x) \varphi_k(y) \cdot \varphi_\ell(y)\, \mathrm{d}y = \sum_{k=1}^\infty \lambda_k \varphi_k(x) \int_{\mathcal{X}} \varphi_k(y)\, \varphi_\ell(y)\, \mathrm{d}y = \lambda_\ell\, \varphi_\ell(x)$$

for all ℓ. The system $(\varphi_k)_{k\in\mathbb{N}}$ is complete and we thus have that $f(\cdot) = \sum_{\ell=1} f_\ell\, \varphi_\ell(\cdot)$. By linearity thus

$$\int_{\mathcal{X}} \sum_{k=1}^\infty \lambda_k \varphi_k(x) \varphi_k(y) \cdot f(y)\, \mathrm{d}y = \sum_{\ell=1} \lambda_\ell\, f_\ell\, \varphi_\ell(x). \tag{6.8}$$

As well we have that

$$\int_{\mathcal{X}} k(x,y) \cdot f(y)\, \mathrm{d}y = \int_{\mathcal{X}} k(x,y) \sum_{\ell=1} f_\ell \varphi_\ell(y)\, \mathrm{d}y = \sum_{\ell=1} f_\ell\, \lambda_\ell\, \varphi_\ell(x). \tag{6.9}$$

The integrals in (6.8) and (6.9) are equal for all $f(\cdot)$, we thus conclude that the kernels coincide, i.e., $k(x,y) = \sum_{k=1}^\infty \lambda_k\, \varphi(x)\, \varphi_k(y)$. □

Corollary 6.10. *The kernel* $k(\cdot,\cdot)$ *is positively definite iff* $k(x,x')=\varphi(x)^\top\varphi(x')$ *for some function* $\varphi\colon\mathcal{X}\to\mathbb{R}^\mathbb{N}$. *The range of* $\varphi(\cdot)$ *is the* feature space *contained in* $\mathbb{R}^\mathbb{N}$.

Proof. If $k(x,x')=\varphi(x)^\top\varphi(x')$, then k is symmetric ($k(x,x')=k(x',x)$) and

$$\begin{aligned}\langle f \mid Kf\rangle &= \iint_{\mathcal{X}\times\mathcal{X}} f(x)\,k(x,y)\,f(y)\mathrm{d}y\,\mathrm{d}x\\ &= \iint_{\mathcal{X}\times\mathcal{X}} f(x)\,\varphi(x)^\top\varphi(y)\,f(y)\mathrm{d}x\,\mathrm{d}y\\ &= \left(\int_{\mathcal{X}} f(x)\,\varphi(x)\,\mathrm{d}x\right)^\top\left(\int_{\mathcal{X}} f(y)\,\varphi(y)\,\mathrm{d}y\right)\\ &= \left\|\int f_{\mathcal{X}}(x)\,\varphi(x)\,\mathrm{d}x\right\|_{\ell_2}^2 \geq 0.\end{aligned}$$

As for the converse we have from Mercer's theorem that

$$k(x,x')=\sum_{k=1}^\infty \lambda_k\varphi_k(x)\varphi_k(x')=\begin{pmatrix}\sqrt{\lambda_1}\,\varphi_1(x)\\ \sqrt{\lambda_2}\,\varphi_2(x)\\ \vdots\end{pmatrix}^\top\begin{pmatrix}\sqrt{\lambda_1}\,\varphi_1(x')\\ \sqrt{\lambda_2}\,\varphi_2(x')\\ \vdots\end{pmatrix}=\varphi(x)^\top\varphi(x'),\qquad x,x'\in\mathcal{X},$$

as $\lambda_k\geq 0$ for positively definite operators induced by the kernel k. □

6.5 PARAMETERS

6.6 LEARNING

The problem is $\min_x \mathbb{E}_{(u,v)}\,(1-u_i x^\top v_i)_+ + \lambda\,\|x\|^2$.

The problem is $\min_x \mathbb{E}_{(u,v)}(0, v\,x^\top u)_+ + \lambda\,\|x\|^2$.

See Steinwart and Christmann [15]

https://www.cs.princeton.edu/~ehazan/

https://jeremykun.com/2017/06/05/formulating-the-support-vector-machine-optimization-problem/

Definition 6.11 (Loss functions)**.** Loss functions include

- Regression, $y\in\mathbb{R}$, $\ell(y,h):=|y-h|^2$,
- Classification, $y\in\{0,1\}$
  - 0–1-loss, $\ell(y,h):=\frac{1}{2}\,(1-\operatorname{sign}(y\,h))=\mathbb{1}_{(-\infty,0]}(y\,h)$,
  - Hinge loss, $\ell(y,h):=\max(0,1-y\,h)$,
  - Log loss, $\ell(y,h):=\log\big(1+\exp(-yh)\big)$.

7 Probabilistic curve fitting

Nomenclature

t target values

x input values, $x = (x_1, \dots, x_N)^\top$

w parameters, often weights

$p(w)$ prior probability distribution

$p(\mathcal{D} \mid w)$ conditional probability distribution

$p(w \mid \mathcal{D})$ posterior probability distribution

7.1 MAXIMUM LIKELIHOOD ESTIMATION

Definition 7.1. The density of the *multivariate* normal distribution $\mathcal{N}(\mu, \Sigma)$ with mean $\mu \in \mathbb{R}^N$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{N \times N}$ is

$$p(t) = \frac{1}{\sqrt{(2\pi)^N \, \det \Sigma}} \exp\left(-\frac{1}{2}(t-\mu)^\top \Sigma^{-1} (t-\mu)\right). \tag{7.1}$$

Recall, that $\beta := \Sigma^{-1}$ is the *precision matrix* and $P(Y \in \mathrm{d}y) = f(y)\,\mathrm{d}y$, where $f(\cdot)$ is the density function.

In a frequentist's maximum likelihood approach, we are interested in the parameter which maximizes the probability of the particular observations x and t, i.e.,

$$w_{\mathsf{ML}} \in \arg\max_w p(x \mid w). \tag{7.2}$$

Example 7.2. Consider independent normals

$$p(x_1, \dots, x_N \mid \mu) := \prod_{n=1}^N \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(x_n-\mu)^2\right) = \sqrt{\frac{\beta}{2\pi}}^N \exp\left(-\frac{\beta}{2}\sum_{i=1}^N (x_n-\mu)^2\right)$$

as in (7.1). The maximum of the corresponding *sum-of-squares error function*

$$\mu_{\mathsf{ML}} \in \arg\max_\mu p(x \mid \mu) = \arg\min_\mu \sum_{n=1}^N (x_n-\mu)^2$$

is attained at $\mu_{\mathsf{ML}} = \frac{1}{N}\sum_{n=1}^N x_n$.

Example 7.3. Consider independent normals

$$p(x_1,\dots,x_N \mid \mu,\beta) := \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(x_n-\mu)^2\right) = \sqrt{\frac{\beta}{2\pi}}^{N} \exp\left(-\frac{\beta}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right)$$

as in (7.1). The maximizers of the problem $(\mu_{\mathsf{ML}},\beta_{\mathsf{ML}}) \in \arg\max_{(\mu,\beta)} p(x \mid \mu,\beta)$ minimize

$$-\log p(x_1,\dots,x_N \mid \mu,\beta) = \frac{\beta}{2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\log\beta;$$

they are $\mu_{\mathsf{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$ and

$$\frac{1}{\beta_{\mathsf{ML}}} = \sigma^2_{\mathsf{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathsf{ML}})^2. \tag{7.3}$$

7.2 MAXIMUM LIKELIHOOD CURVE FITTING

Suppose we want to predict $y(x)$ depending on x. Suppose further a sample of observations (x_n,t_n) is available, where $t := (t_1,\dots,t_N)$ are the *target values* and $x := (x_1,\dots,x_N)$. By picking the parameter w we want to select the function $y(x,w)$, which fits best to the sample observed.

Example 7.4. We assume the distribution

$$p(t_1,\dots,t_N \mid x_1,\dots,x_N,w,\beta) := \prod_{n=1}^{N} \mathcal{N}\big(t_n \mid y(x_n,w),\beta\big).$$

Maximizing the likelihood $\max_w \mathcal{N}\big(t \mid y(x,w)\big)$ corresponds to minimizing the log-likelihood

$$w_{\mathsf{ML}} \in \arg\min_{w} \frac{\beta}{2}\sum_{n=1}^{N}\big(t_n-y(x_n,w)\big)^2 - \frac{N}{2}\log\beta. \tag{7.4}$$

As above we have that $\frac{1}{\beta_{\mathsf{ML}}} = \sigma^2_{\mathsf{ML}} = \frac{1}{N}\sum_{n=1}^{N}(t_n-y(x_n,w_{\mathsf{ML}}))^2$.

Example 7.5. Suppose that $y(x,w) = w^\top g(x) = w_1\,g_1(x)+\dots+w_M\,g_M(x)$, then the problem (7.4) reads

$$w_{\mathsf{ML}} \in \underset{(w_1,\dots,w_M)}{\arg\min} \frac{\beta}{2}\sum_{n=1}^{N}\left(t_n-\sum_{m=1}^{M} w_m\cdot g_m(x_n)\right)^2 - \frac{N}{2}\log\beta, \tag{7.5}$$

which we address further below.

7.3 SIMPLE BAYES

Definition 7.6. The conditional probability is $P(A \mid C)$ satisfies the *product rule*

$$P(A \cap C) = P(A \mid C) \cdot P(C). \tag{7.6}$$

Proposition 7.7 (Law of total probability[1])**.** *Suppose that $(C_k)_{k=1}^K$ is a partition of the sample space (i.e., $\bigcup_{k=1}^K C_k = \Omega$ and $C_j \cap C_k = \emptyset$ whenever $j \neq k$), then the* sum rule

$$P(A) = \sum_k P(A \cap C_k)$$

and

$$P(A) = \sum_{k=1}^K P(A \mid C_k) \cdot P(C_k) \tag{7.7}$$

hold true.

Theorem 7.8 (Bayes' Theorem)**.** *It holds that*

$$P(C \mid A) := \frac{P(A \mid C) \cdot P(C)}{P(A)}. \tag{7.8}$$

Corollary. *For a partition C_k, $k = 1, \dots, K$, it holds that*

(i) $P(C_k \mid A) = \frac{P(A|C_k) \cdot P(C_k)}{P(A)} = \frac{P(A|C_k) \cdot P(C_k)}{\sum_j P(A|C_j) \cdot P(C_j)}$ *and particularly*

$$P(C \mid A) = \frac{P(A \mid C)\, P(C)}{P(A \mid C)\, P(C) + P(A \mid C^c)\, P(C^c)}; \tag{7.9}$$

(ii) $P(C \mid A) = \sum_k P(C \mid A \cap C_k) \cdot P(C_k \mid A)$,

(iii) $P(B \mid A) = \sum_k P(B \mid A \cap C_k) \cdot P(C_k)$ *if B is independent with every C_k,*

(iv) $P(A_1 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdot \ldots P(A_n \mid A_1 \cap \dots \cap A_{n-1})$.

Epistemological interpretation of (7.9): For proposition C and evidence or background A:

(i) $P(C)$ is the *prior* probability, is the initial degree of belief in C;

(ii) $P(C^c) = 1 - P(C)$ is the corresponding probability of the initial degree of belief against C;

(iii) $P(A \mid C)$ is the conditional probability or likelihood, is the degree of belief in A, given that the proposition C is true;

[1] Gesetz der totalen Wahrscheinlichkeit

(iv) $P(A \mid C^{c})$ is the conditional probability or likelihood, is the degree of belief in A, given that the proposition C is false;

(v) $P(C \mid A)$ is the *posterior probability*, is the probability for C after taking into account A for and against C.

In data science, we typically use the Bayes rule for densities. We can rewrite (7.6) as

$$p(w \mid \mathcal{D}) = \frac{p(w, \mathcal{D})}{p(\mathcal{D})}.$$

By Bayes' theorem (cf. (7.8)) we have that

$$p(w \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid w)}{p(\mathcal{D})} \cdot p(w), \tag{7.10}$$

where, by (7.7),

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid w)\, p(w)\, \mathrm{d}w.$$

The denominator $p(\mathcal{D})$ in (7.10) does not depend on w. It follows that

$$\arg\max_w p(w \mid \mathcal{D}) = \arg\max_w p(\mathcal{D} \mid w) \cdot p(w).$$

For this reason, Bayes' theorem (7.10) is often stated as

$$\underbrace{p(w \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{p(\mathcal{D} \mid w)}_{\text{likelihood}} \times \underbrace{p(w)}_{\text{prior}}. \tag{7.11}$$

7.4 BAYESIAN CURVE FITTING

The Bayesian framework assumes a distribution for the prior w, for example

$$p(w) = \mathcal{N}\left(w \mid 0,\, \alpha^{-1}\mathbb{1}\right) = \left(\frac{\alpha}{2\pi}\right)^{M} \exp\left(-\frac{\alpha}{2} w^\top w\right); \tag{7.12}$$

here, $w \in \mathbb{R}^M$ and $\alpha \in \mathbb{R}$ is a *hyperparameter*. By Bayes' theorem (7.11) we infer that

$$\begin{aligned} p(w \mid t, x) &\propto p(t, x \mid w) \times p(w) \\ &= \sqrt{\frac{\beta}{2\pi}}^{N} \exp\left(-\frac{\beta}{2} \sum_{n=1}^{N} \left(t_n - y(x_y, w)\right)^2\right) \times \sqrt{\frac{\alpha}{2\pi}}^{M} \exp\left(-\frac{\alpha}{2} w^\top w\right). \end{aligned} \tag{7.13}$$

Maximizing with respect to w

$$w \in \arg\max_w p(w \mid t, x) = \arg\min_w \sum_{n=1}^{N} \left(t_n - y(x_n, w)\right)^2 + \frac{\alpha}{\beta} w^\top w.$$

This is a regularization with parameter $\lambda := \frac{\alpha}{\beta}$.

We can also include the precision β as a parameter, then the problem is

$$p(w \mid t, x, \beta) \propto p(t, x, \beta \mid w) \times p(w) = (7.13),$$

which corresponds to maximizing

$$(w, \beta) \in \underset{(w,\beta)}{\arg\max}\, p(w \mid t, x) = \underset{(w,\beta)}{\arg\min}\, \frac{\beta}{2}\sum_{n=1}^{N}\big(t_n - y(x_n, w)\big)^2 - \frac{N}{2}\log\beta + \frac{\alpha}{2} w^\top w. \tag{7.14}$$

We conclude from (7.3) that $\frac{1}{\beta_{\text{ML}}} = \frac{1}{N}\sum_{i=1}^{N}\left(t_n - y(x_n, w_{\text{ML}})\right)^2$, where w_{ML} is optimal in (7.14).

Assume that $y(x, w) = w^\top y(x) = \sum_{m=1}^{M} w_m\, y_m(x)$ so that the problem is to minimize

$$\beta \sum_{n=1}^{N}\left(t_n - \sum_{m=1}^{M} w_m\, y_m(x_n)\right)^2 + \alpha \sum_{m=1}^{M} w_m^2$$

with respect to w. Differentiating with respect to w_k gives the first order condition,

$$-2\beta \sum_{n=1}^{N}\left(t_n - \sum_{m=1}^{M} w_m\, y_m(x_n)\right) \cdot y_k(x_n) + 2\alpha\, w_k = 0.$$

This is the k^{th} row in the the normal equations $-\beta Y^\top t + \beta Y^\top Y\, w = -\alpha\, \mathbb{1} w$, where $Y := \big(y_m(x_n)\big)_{n,m} \in \mathbb{R}^{N\times M}$, $t := (t_n)_{n=1}^{N}$ and $w := (w_m)_{m=1}^{M}$. It follows that

$$w = \beta\left(\alpha\, \mathbb{1} + \beta\, Y^\top Y\right)^{-1} Y^\top t = \beta\, S\, Y^\top t,$$

where $S^{-1} := \alpha\, \mathbb{1} + \beta\, Y^\top Y$. Note that the posterior mean is

$$m(x) = y(x)^\top w = \beta\; y(x)^\top S\, Y^\top t$$

and variance

$$s(x)^2 = \beta^{-1} + y(x)^\top S\, y(x),$$

resulting in the predictive distribution

$$p(t \mid x, w, \beta) = \mathcal{N}\left(t \mid m(x), s(x)^2\right).$$

Methods for Classification 8

Suppose that X_i have mean μ_i and variance Σ_i. Then the linear *feature* $w^\top X$ has expectation $w^\top\mu_i$ and variance $w^\top\Sigma_i w$. Note that μ_i and Σ_i can be estimated by $\hat\mu_i = \frac{1}{|C_i|}\sum_{j\in C_i} x_j$ and $\hat\Sigma_i = \frac{1}{|C_i|}\sum_{i\in C_i}(x_j-\hat\mu_i)(x_j-\hat\mu_i)^\top$. The matrix $\hat\Sigma$ is often estimated $\hat\Sigma := \frac{1}{|n|}\sum_{j=1}^n (x_j-\hat\mu)(x_j-\hat\mu)^\top$, where $\hat\mu = \frac{1}{n}\sum_{j=1}^n x_j$.

8.1 (Linear) Discriminant Analysis

Consider the probability densities $p\,(x\mid y=0)$ or $p\,(x\mid y=1)$. The decision can be based on the likelihood ratio by $\frac{p(x|y=1)}{p(x|y=0)} \lessgtr 1$. For normal distributed random variables $\mathcal{N}(\mu_0,\Sigma_0)$ and $\mathcal{N}(\mu_1,\Sigma_1)$ the criterion reduces to

$$(x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0)+\log\det\Sigma_0-(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)-\log\det\Sigma_1 > T, \tag{8.1}$$

where T is some threshold. Note, that (8.1) describes an ellipsoid. Assuming that $\Sigma=\Sigma_0=\Sigma_1$ the criterion further reduces to

$$w^\top x > c$$

with $w=\Sigma^{-1}(\mu_1-\mu_0)$ and $c=\frac{1}{2}\left(T-\mu_0^\top\Sigma^{-1}\mu_0+\mu_1^\top\Sigma^{-1}\mu_1\right)$.

8.2 Fisher's Linear Discriminant

Fisher[1] defined the *separation* S between these two to be the ratio of the variance between the classes to the variance within the classes,

$$S=\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}}=\frac{(w^\top\mu_1-w^\top\mu_0)^2}{w^T\Sigma_1 w+w^T\Sigma_0 w}=\frac{(w^\top(\mu_1-\mu_0))^2}{w^T(\Sigma_0+\Sigma_1)w}=\frac{w^\top S_b w}{w^\top\Sigma w}, \tag{8.2}$$

where $S_b=(\mu_1-\mu_0)(\mu_1-\mu_0)^\top$. This measure is, in some sense, a measure of the signal-to-noise ratio for the class labelling.

The maximum separation occurs when S is large. Note, that S is invariant with respect to re-scaling of w. The first order conditions for the Lagrangian

$$L(w,\lambda):=\left(w^\top\Delta\mu\right)^2-\lambda\left(w^\top\Sigma w-1\right)$$

[1] Ronald Fisher, 1890–1962, British statistician

includes

$$0 = \frac{\partial}{\partial w} L = 2\left(w^\top \Delta\mu\right) \Delta\mu^\top - \lambda\left((\Sigma w)^\top + w^\top \Sigma\right)$$
$$= 2\left(w^\top \Delta\mu\right) \Delta\mu^\top - 2\lambda w^\top \Sigma$$

from which follows that

$$w \propto (\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0). \tag{8.3}$$

This is Fisher's linear discriminant, the same solution as for linear discriminant analysis (LDA, Section 8.1 above), but does not require the assumptions made there.

Remark 8.1. Differentiating S directly gives $\frac{\partial S}{\partial w} \propto \Delta\mu^\top - w^\top\Sigma$, which again characterizes Fisher's linear discriminant (8.3).

Remark 8.2. Note that the optimal vector w in (8.2) maximizes the Rayleight quotient $S = \frac{w^\top S_b w}{w^\top \Sigma w} = \frac{\tilde{w}^\top \Sigma^{-1/2} S_b \Sigma^{-1/2} \tilde{w}}{\tilde{w}^\top \tilde{w}}$, where $\tilde{w} := \Sigma^{1/2} w$ so that $\tilde{w}$ is an eigenvector and satisfies $\Sigma^{-1/2} S_b \Sigma^{-1/2} \tilde{w} = S\,\tilde{w}$, or equivalently, $\Sigma^{-1} S_b\, w = S\, w$. Hence, w is an eigenvector of $\Sigma^{-1} S_b$ for the Eigenvalue S.

Remark 8.3 (Shrinkage). Occasionally, one considers the matrix $(1-\lambda)\Sigma + \lambda\mathbb{1}$ for some *shrinkage intensity* or *regularisation parameter* λ.

8.3 PERCEPTION ALGORITHM

Consider Rosenblatt's[2] Perceptron, i.e., the nonlinear classifier $y(x) = \operatorname{sign}(w^\top \phi(x))$. Define the target values $t = 1$ ($t = -1$, resp.) if $x \in C_1$ ($x \in C_2$, resp.). Note, that $t_i \cdot w^\top \phi(x_i) > 0$ for correctly classified data. The perception criterion is $E_P(w) = -\sum_{i \in \mathcal{M}} t_i \cdot w^\top \phi(x_i)$, where $\mathcal{M}$ collects misclassified patterns. The perception algorithm is $w^{\tau+1} = w^\tau + \eta\, t_n\, \phi(x_n)$, where $n \in \mathcal{M}$ is misclassified.

8.4 MULTIPLE CLASSES

Classifiers for multiple classes $C_1, \dots, C_K$ can be obtained by $y_k(x) := w_k^\top x + w_{k0}$ and the classification

$$x \in C_k \iff k \in \operatorname*{arg\,max}_{k'=1,\dots,K} w_{k'}^\top x + w_{k'0}.$$

These classes are necessarily convex.

8.5 PROBABILISTIC METHODS

Recall from Bayes' theorem that

$$p(C_k \mid x) = \frac{p(x \mid C_k) \cdot p(C_k)}{\sum_{k=1}^K p(x \mid C_k) \cdot p(C_k)} = \frac{\exp(a_k)}{\sum_{j=1}^K \exp(a_j)},$$

[2]Frank Rosenblatt, 1928–1971, American psychologist notable in the field of artificial intelligence

where

$$a_k(x) := \log\big(p(x \mid C_k) \cdot p(C_k)\big).$$

In particular we have that

$$\begin{aligned} p(C_1 \mid x) &= \frac{p(x \mid C_1) \cdot p(C_1)}{p(x \mid C_1) \cdot p(C_1) + p(x \mid C_2) \cdot p(C_2)} \\ &= \frac{1}{1 + \frac{p(x|C_2) \cdot p(C_2)}{p(x|C_1) \cdot p(C_1)}} \\ &= \frac{1}{1 + \exp(-a)} = S(a), \end{aligned}$$

where $a(x) := \log \frac{p(x|C_1) \cdot p(C_1)}{p(x|C_2) \cdot p(C_2)} = a_1(x) - a_2(x)$ and $S(x) = \frac{1}{1+\exp(-x)}$ is the *logistic sigmoid* function.

Remark 8.4. Suppose that $p(\cdot \mid C_k)$ is the density of a normal distribution $\mathcal{N}(\mu_k, \Sigma)$. Then $a(x) = w^\top x + w_0$, where $w = \Sigma^{-1}(\mu_2 - \mu_1)$ and $w_0 = -\frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^\top \Sigma^{-1}\mu_2 + \log \frac{p(C_1)}{p(C_2)}$. It follows that $p(C_1 \mid x) = S(w^\top x + w_0)$.

For general classes, $a_k(x) := w_k^\top x + w_{k0}$, where $w_k = \Sigma^{-1}\mu_k$ and $w_{k0} = -\frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k + \log p(C_k)$.

8.6 SUPPORT VECTORS

Lemma 8.5. *The linear equation $w^\top x = b$ defines a hyperplane. The point on the hyperplane closest (in Euclidean norm) to the origin is $w\frac{b}{\|w\|^2}$. The distance to the hyperplane is $\frac{b}{\|w\|}$.*

Proof. Apparently, $p := w\frac{b}{\|w\|^2}$ is on the hyperplane, as $w^\top p = b$.

Note that $p \propto w$, the normal vector. For any other vector x on the plane it holds that $x - p \perp p$ (indeed, $p^\top(x-p) = \frac{b}{\|w\|^2}(w^\top x - w^\top p) = \frac{b}{\|w\|^2}\left(b - w^\top w \frac{b}{\|w\|^2}\right) = 0$) and thus $w^\top(p + (x-p)) = b$ for which the norm is $\|x\|^2 = \|p\|^2 + \|x-p\|^2 \geq \|p\|^2$. □

Corollary 8.6. *The distance of the hyperplanes $w^\top x - b = \pm 1$ is*

$$\frac{2}{\|w\|}. \tag{8.4}$$

Proof. The hyperplanes are parallel, so the points closest to the origin are closest to each other. Their distance is $\frac{b+1}{\|w\|} - \frac{b-1}{\|w\|} = \frac{2}{\|w\|}$. □

8.7 LINEARLY SEPARABLE DATA – HARD MARGIN

Let $D := \{(x_i, y_i) \colon i = 1, \dots, m\}$ be a set of data with $y_i \in \{-1, 1\}$. We are looking for a linear rule consisting of w and b separating the data in the distinct sets $I_+ := \{i \colon y_i > 0\}$ and $I_- := \{i \colon y_i < 0\}$. A correct linear classifier satisfies $\operatorname{sign}(w^\top x_i + b) = y_i$ or, equivalently, $y_i(w^\top x_i + b) \geq 0$ for all $i \leq m$.

Definition 8.7. The geometric margin of a hyperplane w with respect to a dataset D is the shortest distance from a training points x_i to the hyperplane defined by w. The *best hyperplane* has the largest possible margin.

Problem 8.8 (Support vectors)**.** By rescaling the plane parameters w and b, the classifications defined by the hyperplane are $w^\top x_i - b \geq 1$ for $i \in I_+$ and $w^\top x_i - b \leq -1$ for $i \in I_-$. The hyperplane midway between the classification points (x_i, y_i) with largest distance (margin, cf. (8.4)) is given by

$$\begin{aligned} &\underset{\text{in } w,\, b}{\text{minimize}} && \frac{1}{2}\|w\|^2 \\ &\text{subject to} && y_i \left(w^\top x_i - b\right) \geq 1 \text{ for all } i = 1, \dots, m. \end{aligned} \tag{8.5}$$

The classifier is given by $x \mapsto \operatorname{sign}(w^\top x - b)$, where b and w are the support vectors solving the preceding optimization problem. Note that the problem (8.5) is convex.

8.8 NOT LINEARLY SEPARABLE DATA – SOFT MARGIN

Definition 8.9 (Hinge[3] loss)**.** For an intended output $t = \pm 1$ and a classifier score y, the *hinge loss* (or *ramp function*) is

$$\ell(y; t) := \max(0,\ 1 - y \cdot t) = (1 - y \cdot t)_+ .$$

Note, that $\ell(w^\top x_i - b; t) = 0$, if $t = y_i$ and the constraints (8.5) are satisfied. We thus wish to solve

$$\underset{\text{in } w,\, b}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^n \max\left(0,\ 1 - y_i\left(w^\top x_i - b\right)\right) + \frac{\lambda}{2}\|w\|^2, \tag{8.6}$$

where the parameter λ[4] determines the trade-off between increasing the margin size and ensuring that the x_i lie on the correct side of the margin. Thus, for sufficiently small values of λ, the second term in the loss function will become negligible, hence, it will behave similar to the hard-margin SVM, if the input data are linearly classifiable, but will still learn if a classification rule is viable or not.

Remark 8.10. Note, that $\ell(\cdot)$ is a convex function. Further, the objective (8.6) is convex and the problem does not involve constraints.

8.8.1 Dualization

We may rewrite the problem (8.6) as

$$\underset{\text{in } w,\, b,\, s}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1} s_i + \frac{\lambda}{2}\|w\|^2 \tag{8.7}$$

$$\text{subject to } y_i\left(w^\top x_i - b\right) \geq 1 - s_i \text{ and} \qquad (\alpha_i \geq 0) \tag{8.8}$$

$$s_i \geq 0 \text{ for all } i = 1, \dots, n, \qquad (\beta_i \geq 0) \tag{8.9}$$

[3] Drehgelenk, Scharnier in German

[4] $\frac{1}{\lambda}$ is also known as the *soft margin parameter*.

where the slack variable s_i quantifies the amount to which the constraint (8.8) is violated.

The Lagrangian is

$$L(w,b,s;\alpha_i,\beta_i) \coloneqq \frac{1}{n}\sum_{i=1}^n s_i+\frac{\lambda}{2}\|w\|^2+\frac{\lambda}{n}\sum_{i=1}^n \alpha_i\cdot\Big(1-s_i-y_i\left(w^\top x_i-b\right)\Big)-\frac{\lambda}{n}\sum_{i=1}^n \beta_i\cdot s_i, \tag{8.10}$$

which we minimize with respect to the primal variables w, b and s for fixed Lagrange multipliers $\alpha_i \ge 0$ and $\beta_i \ge 0$ corresponding to the inequality constraints in (8.7). The first order conditions are

$$\frac{\partial L}{\partial w_j} = \lambda\, w_j - \frac{\lambda}{n}\sum_{i=1}^n \alpha_i\, y_i\, x_{i,j} = 0, \qquad j = 1,\dots,m, \tag{8.11}$$

$$\frac{\partial L}{\partial s_j} = \frac{1}{n}\left(1-\lambda\alpha_j-\lambda\beta_j\right) = 0, \qquad j = 1,\dots,m \text{ and} \tag{8.12}$$

$$\frac{\partial L}{\partial b} = \frac{\lambda}{n}\sum_{i=1}^n \alpha_i\, y_i = 0. \tag{8.13}$$

From (8.11) it follows that the support vector is

$$w = \frac{1}{n}\sum_{i=1}^n \alpha_i\, y_i\, x_i. \tag{8.14}$$

It follows from (8.12) that

$$\beta_i = \frac{1}{\lambda} - \alpha_i. \tag{8.15}$$

The Lagrange multipliers α_i and β_i correspond to inequality constraints in (8.7), so they are nonnegative, i.e., $0 \le \alpha_i \le \frac{1}{\lambda}$. The Lagrangian (8.10) thus simplifies to

$$\begin{aligned} L(w,b,s;\alpha_i,\beta_i) &= \frac{1}{n}\sum_{i=1} s_i + \frac{\lambda}{2}\|w\|^2 \\ &\qquad + \frac{\lambda}{n}\sum_i \alpha_i - \frac{\lambda}{n}\sum_i \alpha_i s_i - \lambda\, w^\top \underbrace{\frac{1}{n}\sum_{i=1}^n \alpha_i y_i x_i}_{w \text{ by } (8.14)} + \underbrace{\frac{\lambda}{n}\sum_{i=1}^n \alpha_i y_i\, b}_{=0 \text{ by } (8.13)} \\ &\qquad - \frac{\lambda}{n}\sum_{i=1}^n \underbrace{\left(\frac{1}{\lambda}-\alpha_i\right)}_{=\beta_i \text{ by } (8.15)} \cdot s_i \\ &= -\frac{\lambda}{2}\|w\|^2 + \frac{\lambda}{n}\sum_i \alpha_i \end{aligned}$$

by convex duality. The convex dual to the preceding problem (8.7)–(8.9) is

$$\begin{aligned} \underset{\text{in } \alpha}{\text{maximize}} \quad & \frac{1}{n}\sum_{i=1}^n \alpha_i - \frac{1}{2}\|w\|^2 = \frac{1}{n}\sum_{i=1}^n \alpha_i - \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\, y_i y_j\, x_i^\top x_j \\ \text{subject to} \quad & \frac{1}{n}\sum_{i=1}^n y_i\,\alpha_i = 0 \text{ and} \qquad \text{(cf. (8.13))} \\ & 0 \le \alpha_i \le \frac{1}{\lambda}. \end{aligned} \tag{8.16}$$

Remark 8.11. Note, that (x_i, y_i) is correctly classified, if $s_i = 0$. By complementary slackness we have that $\alpha_i < \frac{1}{\lambda} \underset{(8.15)}{\iff} \beta_i > 0 \underset{(8.9)}{\implies} s_i = 0$.

The offset b can be recovered by finding an x_i on the margin's boundary (i.e., $\alpha_i < \frac{1}{\lambda}$) and solving

$$y_i\left(w^\top x_i - b\right) = 1 \iff b = w^\top x_i - y_i$$

(as $y_i^2 = 1$). The classification then is $x \mapsto \operatorname{sign}\left(\sum_{i=1}^n \alpha_i y_i x_i^\top x - b\right)$.

8.8.2 The kernel trick I

The dual problem can be generalized by involving a kernel function $k(x, y)$ and solving

$$\begin{aligned} \underset{\text{in } \alpha}{\text{maximize}} \quad & \frac{1}{n}\sum_{i=1}^n \alpha_i - \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\, y_i y_j\, k(x_i, x_j) \\ \text{subject to} \quad & \frac{1}{n}\sum_{i=1}^n y_i\,\alpha_i = 0 \text{ and} \\ & 0 \le \alpha_i \le \frac{1}{\lambda} \end{aligned} \tag{8.17}$$

instead. The hyperplane $\frac{1}{n}\sum_{i=1}^n \alpha_i\, y_i\, k(x_i, x) = \text{const}$ then specifies the classification rule.

8.8.3 The kernel trick II

Consider the (unconstrained) optimization problem

$$\underset{\text{in } f(\cdot)}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i); f_i\big) + \frac{\lambda}{2}\|f\|_k^2. \tag{8.18}$$

The Lagrangian of the equivalent reformulation

$$\begin{aligned} \underset{\text{in } f(\cdot),\, u\in\mathbb{R}^n}{\text{minimize}} \quad & \frac{1}{n}\sum_{i=1}^n \ell\big(u_i; f_i\big) + \frac{\lambda}{2}\|f\|_k^2 \\ \text{subject to } & u_i = \big\langle k(\cdot, x_i), f(\cdot)\big\rangle \text{ for } i = 1, \dots, n \end{aligned}$$

with dual parameters (shadow costs) $\alpha = (\alpha_i)_{i=1}^n$ is

$$L(f,u;\alpha) := \frac{1}{n}\sum_{i=1}^n \ell\big(u_i; f_i\big) + \frac{\lambda}{2}\|f\|_k^2 + \frac{\lambda}{n}\sum_{i=1}^n \alpha_i\Big(u_i - \big\langle k(\cdot,x_i), f(\cdot)\big\rangle\Big)$$
$$= \frac{1}{n}\sum_{i=1}^n \big(\ell\big(u_i; f_i\big) + u_i\cdot\lambda\alpha_i\big) + \frac{\lambda}{2}\Big\|f(\cdot) - \frac{1}{n}\sum_{i=1}^n \alpha_i k(\cdot,x_i)\Big\|_k^2 - \frac{\lambda}{2n^2}\sum_{i,j=1}^n \alpha_i k(x_i,x_j)\alpha_j$$

with dual function

$$d(\alpha) := \inf_{f,u} L(f,u;\alpha).$$

This objective is minimal for $f(\cdot) = \frac{1}{n}\sum_{i=1}^n \alpha_i\, k(\cdot,x_i)$ and thus

$$d(\alpha) = -\frac{1}{n}\sum_{i=1}^n \ell^*(-\lambda\alpha_i; f_i) - \frac{\lambda}{2n^2}\sum_{i,j=1}^n \alpha_i k(x_i,x_j)\alpha_j,$$

where $= \inf_{u\in\mathbb{R}} \ell(u;y) - u\cdot\alpha = -\sup_{u\in\mathbb{R}} u\cdot\alpha - \ell(u;y) = -\ell^*(\alpha;y)$ is the convex conjugate function, cf. (3.6). The optimization problem (8.18) thus is

$$\begin{array}{c}\text{maximize}\\ \text{in } \alpha\in\mathbb{R}^n\end{array} \quad -\frac{1}{n}\sum_{i=1}^n \ell^*(-\lambda\alpha_i; f_i) - \frac{\lambda}{2n^2}\sum_{i,j=1}^n \alpha_i k(x_i,x_j)\alpha_j. \tag{8.19}$$

8.8.4 The kernel trick III

A particular situation arises for $k(x,y) = \varphi(x)^\top\varphi(y)$, where $\varphi\colon \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ maps the data into the *feature space* with $d_2 > d_1$. The solution of (8.17) is $w = \frac{1}{n}\sum_{i=1}^n \alpha_i\, y_i\, \varphi(x_i)^\top$ and the classification reads

$$w^\top\varphi(x) = \frac{1}{n}\sum_{i=1}^n \alpha_i\, y_i\, \varphi(x_i)^\top\varphi(x) = \frac{1}{n}\sum_{i=1}^n \alpha_i\, y_i\, k(x_i,x),$$

which is known as the *kernel trick*, or *kernel substitution*.

The classification problem can be stated as

$$\begin{array}{c}\text{minimize}\\ \text{in } w\end{array} \quad J(w) := \frac{1}{2}\sum_{i=1}^n \big(w^\top\varphi(x_i) - y_i\big)^2 + \frac{\lambda}{2}w^\top w. \tag{8.20}$$

Differentiating with respect to w gives the first order conditions

$$\nabla_w J = \sum_{i=1}^n \big(w^\top\varphi(x_i) - y_i\big)\,\varphi(x_i) + \lambda\, w = 0,$$

or

$$w = \sum_{i=1}^n \underbrace{\frac{1}{\lambda}\big(y_i - w^\top\varphi(x_i)\big)}_{=:a_i}\,\varphi(x_i) = \varphi^\top a,$$

where $\varphi = \big(\varphi(x_1), \dots, \varphi(x_n)\big)^\top$ is the design matrix.

Substituting $w = \varphi^\top a$ in (8.20) gives the problem

$$\underset{\text{in } a}{\text{minimize}} \quad \tilde{J}(a) := \frac{1}{2}\sum_{i=1}^{n}\big(a^\top \varphi\, \varphi(x_i) - y_i\big)^2 + \frac{\lambda}{2} a^\top \varphi\varphi^\top a \tag{8.21}$$

$$= \frac{1}{2} a^\top \varphi\varphi^\top \varphi\varphi^\top a - a^\top \varphi\varphi^\top y + \frac{1}{2} y^\top y + \frac{\lambda}{2} a^\top \varphi\varphi^\top a$$

$$= \frac{1}{2} a^\top K K a - a^\top K y + \frac{1}{2} y^\top y + \frac{\lambda}{2} a^\top K a, \tag{8.22}$$

where $K = \varphi\varphi^\top$ is the Gram[5] matrix with entries $K_{ij} = \varphi(x_i)^\top \varphi(x_j) =: k(x_i, x_j)$. The solution of the problem (8.22) is $a = (K + \lambda \cdot \mathbb{1})^{-1} y$. The final prediction is

$$y(x) = w^\top \varphi(x) = \varphi(x)^\top w = \varphi(x)^\top \varphi^\top a = k(x)^\top (K + \lambda \mathbb{1})^{-1} y,$$

where $k_i(x) = \varphi(x)^\top \varphi(x_i) = k(x_i, x)$.

8.9 PROBLEMS

Exercise 8.1. *Show that the conjugate of the hinge loss is* $\ell^*(z;t) = \begin{cases} \frac{z}{t} & \text{if } \frac{z}{t} \in [-1,0], \\ +\infty & \text{else} \end{cases}$.

[5] Jørgen Pedersen Gram, 1850–1916, Danish actuary and mathematician

9 Neural Networks

9.1 FORWARD PROPAGATION

Definition 9.1 (Prediction functions for Classification)**.** Prediction functions for classification include

- Support vector machines, $h\big(x,(w,b)\big)=w^\top x+b$,
- Deep neural networks, $h\big(x,(W_1,\dots W_J,b_1,\dots,b_J\big):=(S_J\circ\dots\circ S_1)\,(x)$, where $S_j(x):=h(W_jx+b_j)$ for some nonlinear activation function h and $S_J=s$ is the sigmoid function, $s(x)=\frac{1}{1+e^{-x}}$.

$a_j:=W_jx+b_j$ at the layer j is called an activation. the activation $a_j:=\sum_i w_{ji}^{(1)}x_i+w_{j0}^{(1)}$, where the parameters $w_{j0}^{(1)}$ are called *biases*. For an activation function $h(\cdot)$ set $z_j:=h(a_j)$. A typical activation function is $h(x)=\max(0,x)$. for Forward propagation is the evaluation of the neural network, i.e.,

$$\Phi\colon x\mapsto s\left(T_L\,h\left(\sum_j T_{L-1}\dots T_2 h\,(T_1x)\right)\right),$$

where

$$\begin{aligned}T_\ell\colon \mathbb{R}^{n_{\ell-1}}&\to\mathbb{R}^{n_\ell}\\ x&\mapsto A_\ell\,x+b_\ell\end{aligned}$$

and $h(x_1,\dots,x_n):=\big(h(x_1),\dots,h(x_n)\big)$.

Mathematical foundations of neural networks include

- the universal approximation theorem and
- the Kolmogorov–Arnold representation theorem.

10 Stochastic Approximation

In what follows we assume that $f\colon \mathbb{R}^n \to \mathbb{R}$ is sufficiently smooth. We follow Pflug [12]. See also Nemirovski et al. [11].

10.1 GRADIENT METHOD

Proposition 10.1. *Suppose that the gradient of $f\colon \mathbb{R}^d \to \mathbb{R}$ is Lipschitz, i.e.,*

$$\|\nabla f(y) - \nabla f(x)\| \le L \|y - x\|, \tag{10.1}$$

then

$$f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|^2. \tag{10.2}$$

Proof. Consider the mapping $t \mapsto f(x + t\,h)$ for some fixed direction $h \in \mathbb{R}^d$. With Cauchy–Schwarz it holds that

$$\begin{aligned} f(x+h) - f(x) &= \int_0^1 f'(x + t\,h)^\top h \,\mathrm{d}t \\ &= f'(x)^\top h + \int_0^1 \left(f'(x + t\,h) - f'(x)\right)^\top h \,\mathrm{d}t \\ &\le f'(x)^\top h + \int_0^1 \|f'(x + t\,h) - f'(x)\| \, \|h\| \,\mathrm{d}t \end{aligned}$$

and with Lipschitz continuity (10.1) thus further

$$\begin{aligned} f(x+h) - f(x) &\le f'(x)^\top h + \int_0^1 L \|t\,h\| \, \|h\| \,\mathrm{d}t \\ &= f'(x)^\top h + L \|h\|^2 \int_0^1 t \,\mathrm{d}t \\ &= f'(x)^\top h + \frac{L}{2} \|h\|^2. \end{aligned} \tag{10.3}$$

The assertion follows with $h = y - x$. □

Remark 10.2. The condition in the preceding proposition is true, if $f \in C^2$ with uniformly bounded Hessian, $\left\|\nabla^2 f(x)\right\| \le L < \infty$.

Lemma 10.3 (Steepest descent)**.** *The gradient $f'(x) = \nabla f(x)$ is the direction of steepest ascent.*

Proof. By Taylor's series expansion it holds that $f(x+t\,h) = f(x) + t\cdot f'(x)^\top h + o(t)$. Among all $h\in\mathbb{R}^n$ with $\|h\| = \|f'(x)\|$ the descent $\frac{1}{t}\big(f(x+t\,h) - f(x)\big) + o(1) = f'(x)^\top h$ is largest for the direction $h = -f'(x)$. □

Definition 10.4. The steepest descent algorithm is

$$x_{k+1} := x_k - \alpha_k \cdot \nabla f(x_k), \tag{10.4}$$

where $\alpha_k > 0$ is an appropriate step size (learning rate).

Example 10.5. Let $f(x) = \frac{c}{2}x^2$, then $x_{k+1} = x_k - \alpha_k \cdot c x_k = x_k(1 - c\alpha_k)$. For the sequence to converge (to the minimum, which is 0) we need $|1 - c\alpha_k| < 1$, i.e., $\alpha_k \in \left(0, \frac{2}{c}\right)$. Note, that $\alpha_k = \alpha$ does not lead to convergence, if $\alpha \geq \frac{2}{c}$ (usually, we don't know c). Hence we need $\alpha_k \to 0$, as $k\to\infty$. Note, that

$$x_k = x_0 \cdot \prod_{\ell=0}^{k-1} (1 - c\,\alpha_\ell)\,.$$

It holds that $\prod_{\ell=0}^{k-1}(1 - c\alpha_\ell) < \infty$, iff $c\sum_{\ell=0}\alpha_\ell < \infty$. For $\alpha_k \to 0$ we necessarily need that $\sum_{k=0}\alpha_k = \infty$.

Lemma 10.6 (Steepest descent)**.** *Suppose that f is bounded from below and $x \mapsto f'(x)$ is Lipschitz with constant L. Suppose further that $\alpha_k > 0$, $\alpha_k \to 0$ as $k\to\infty$ and $\sum_{k=1}^\infty \alpha_k = \infty$ in the sequence (10.4). Then the sequence $f(x_k)$ converges and $\|f'(x_k)\| \xrightarrow[k\to\infty]{} 0$.*

Proof. With (10.2) and the step $h := -\alpha_k \cdot f'(x_k)$ in (10.3) we have

$$f(x_{k+1}) - f(x_k) \leq -\alpha_k \|f'(x_k)\|^2 + \frac{\alpha_k^2 L}{2}\|f'(x_k)\|^2 = -\left(\alpha_k - \frac{\alpha_k^2 L}{2}\right)\|f'(x_k)\|^2\,. \tag{10.5}$$

As $\alpha_k - \alpha_k^2\frac{L}{2} > 0$ for $k > N$ large enough it follows that $f(x_k)$ is strictly decreasing for $k > N$.

Recall that $f(x_{\ell+1})$ is bounded from below, thus

$$-\infty < C - f(x_N) \leq f(x_{\ell+1}) - f(x_N) \leq -\sum_{k=N}^{\ell}\left(\alpha_k - \alpha_k^2\frac{L}{2}\right)\|f'(x_k)\|^2$$

and the sequence $f(x_k)$ converges. Further, the series

$$\sum_{k=N}^{\ell}\left(\alpha_k - \alpha_k^2\frac{L}{2}\right)\cdot\|f'(x_k)\|^2 < \infty$$

converges. Since $\sum_{k=N}^{\ell}\left(\alpha_k - \alpha_k^2\frac{L}{2}\right) \xrightarrow[\ell\to\infty]{} \infty$ it follows that $\liminf_{k\to\infty}\|f'(x_k)\|^2 = 0$.

Suppose that $\limsup_{k\to\infty} \|f'(x_k)\| > 2\varepsilon > 0$. Let $m_i < n_i < m_{i+1}$ be chosen so that

$$\begin{aligned} &\|f'(x_k)\| > \varepsilon \text{ for } k \in [m_i, n_i) \text{ and} \\ &\|f'(x_k)\| \le \varepsilon \text{ for } k \in [n_i, m_{i+1}). \end{aligned} \tag{10.6}$$

Let k_0 be large enough so that $\sum_{k=k_0} \alpha_k \|f'(x_k)\|^2 < \varepsilon^2/L$. Then, for k large enough so that $m_i > k_0$ and j, $\ell \in [m_i, n_i)$, it holds that

$$\left\|f'(x_{\ell+1}) - f'(x_j)\right\| = \left\|\sum_{k=j}^{\ell} f'(x_{k+1}) - f'(x_k)\right\| \le L \sum_{k=j}^{\ell} \alpha_k \|f'(x_k)\| < \frac{L}{\varepsilon} \sum_{k=j}^{\ell} \alpha_k \|f'(x_k)\|^2 < \frac{L}{\varepsilon}\frac{\varepsilon^2}{L} = \varepsilon$$

by Lipschitz continuity of f' and (10.4) and because $1 < \frac{\|f'(x_k)\|}{\varepsilon}$ by (10.6). It follows that $\|f'(x_k)\| \le \left\|f'(x_{n_i})\right\| + \left\|f'(x_{n_i}) - f'(x_k)\right\| \le \varepsilon + \varepsilon$ for $k \in [m_i, n_i)$. But $\|f'(x_k)\| \le \varepsilon$ for $j \in [n_i, m_{i+1})$ and thus $\limsup \left\|f'(x_j)\right\| < 2\varepsilon$. This contradicts the assumption and thus $\|f'(x_k)\| \xrightarrow[k\to\infty]{} 0$. □

10.2 STOCHASTIC APPROXIMATION

Stochastic gradient descent, also known as *sequential gradient descent* or *stochastic approximation* dates back to Robbins and Monro [13]. The presentation here follows Bottou, Curtis, and Nocedal [3]. We consider the stochastic and particular optimization problem (EM–algorithm)

$$f(x) := \min_{x\in\mathcal{X}} \mathbb{E}\, f(x,\xi) = \min_{x\in\mathcal{X}} \int_{\mathbb{R}^d} f(x,\xi)\, P(\mathrm{d}\xi).$$

input : x_0 and a sequence $\alpha_k > 0$, $k = 0, 1, 2, \dots$ with (10.11)
output : a random sequence x_k

for $k = 0, 1, 2, \dots$ **do**
    generate a new sample ξ_k
    compute the stochastic (gradient) vector $g(x_k, \xi_k)$ and
    set $x_{k+1} := x_k - \alpha_k \cdot g(x_k, \xi_k)$
end

Algorithm 5: Stochastic gradient descent

Example 10.7 (Cf. Kalman filters)**.** Consider the problem $\min_x \mathbb{E}_\xi\, f(x,\xi)$ with $f(x,\xi) := \frac{1}{2}(x-\xi)^2$. Note, that $g(x,\xi) := \nabla_x f(x,\xi) = x - \xi$. Choose x_0 arbitrary and $\alpha_k := \frac{1}{k+1}$, set

$$x_{k+1} := x_k - \alpha_k \cdot g(x_k, \xi_k) = x_k - \alpha_k \cdot (x_k - \xi_k)\,.$$

Then $x_k = \frac{1}{k}\sum_{j=0}^{k-1} \xi_j = \overline{\xi}_k \to \mathbb{E}\,\xi$ by the law of large numbers.

Proof. The statement is apparently correct for $k = 0$ and $k = 1$. Indeed, note that $x_1 = x_0 - 1 \cdot (x_0 - \xi_0) = \xi_0$ and $x_2 = x_1 - \frac{1}{2}(x_1 - \xi_1) = \xi_0 - \frac{1}{2}(\xi_0 - \xi_1) = \frac{1}{2}(\xi_0 + \xi_1)$. By induction,

$$x_{k+1} = \frac{1}{k}\sum_{j=0}^{k-1}\xi_j - \frac{1}{k+1}\left(\frac{1}{k}\sum_{j=0}^{k-1}\xi_j - \xi_k\right) = \frac{1}{k}\left(1 - \frac{1}{k+1}\right)\sum_{j=0}^{k-1}\xi_j + \frac{1}{k+1}\xi_k,$$

from which the assertion is immediate. □

Remark 10.8. For Kalman filters see Williams [18] or Brockwell and Davis [4], Liptser and Shiryaev [10].

The gradient $d := g(x_k, \xi_k)$ depends on ξ_k and thus $x_{k+1} = x_{k+1}(\xi_k)$ is random. We shall indicate randomness with respect to ξ_k given x_k explicitly by writing $\mathbb{E}_{\xi_k}$, etc.

Corollary 10.9 (Corollary to Lemma 10.6)**.** *Suppose that (10.1) holds true in Algorithm 5, then*

$$\mathbb{E}_{\xi_k} f(x_{k+1}, \xi_k) \le f(x_k, \xi_k) - \alpha_k \nabla f(x_k)^\top \mathbb{E}_{\xi_k} g(x_k, \xi_k) + \frac{L\,\alpha_k^2}{2}\mathbb{E}_{\xi_k} \|g(x_k, \xi_k)\|^2. \quad (10.7)$$

Proof. The assertion follows from (10.5) by taking expectations for the stochastic gradient $d := g(x_k, \xi_k)$. □

Corollary 10.10. *Suppose that $g(x, \xi)$ is an unbiased estimator for $\nabla f(x, \xi)$ (for example, $g(x, \cdot) := \nabla_x F(x, \cdot)$), then*

$$\mathbb{E}_{\xi_k} f(x_{k+1}) \le f(x_k) - \left(\alpha_k - \frac{L\,\alpha_k^2}{2}\right)\|\nabla f(x_k)\|^2.$$

Remark 10.11. Recall that $\operatorname{var} g = \mathbb{E}\, g\, g^\top - (\mathbb{E}\, g)(\mathbb{E}\, g)^\top \in \mathbb{R}^{d\times d}$ and

$$\operatorname{trace}\operatorname{var} g(x_k, \xi_k) = \sum_{i=1}^{d} \operatorname{var} g_i(x_k, \xi_k) = \mathbb{E}_{\xi_k}\|g(x_k, \xi_k)\|^2 - \left\|\mathbb{E}_{\xi_k} g(x_k, \xi_k)\right\|^2.$$

Theorem 10.12. *Suppose that*

(i) $\nabla f(x_k)^\top \mathbb{E}_{\xi_k} g(x_k, \xi_k) \ge \mu \|\nabla f(x_k)\|^2$ *for some* $\mu > 0$,

(ii) $\left\|\mathbb{E}_{\xi_k} g(x_k, \xi_k)\right\| \le \mu_G \|\nabla f(x_k)\|$ *for some* $\mu_G \ge \mu$ *and*

(iii) $\mathbb{V}\left(g(x_k, \xi_k)\right) := \mathbb{E}_{\xi_k}\|g(x_k, \xi_k)\|^2 - \left\|\mathbb{E}_{\xi_k} g(x_k, \xi_k)\right\|^2 \le M + M_V \|\nabla f(x_k)\|^2$.

Then it holds that

$$\begin{aligned}\mathbb{E}_{\xi_k} f(x_{k+1}) - f(x_k) &\le -\mu\alpha_k \|\nabla f(x_k)\|^2 + \frac{L\,\alpha_k^2}{2}\mathbb{E}_{\xi_k}\|g(x_k, \xi_k)\|^2 && (10.8)\\ &\le -\left(\mu - \frac{\alpha_k\, L\, M_G}{2}\right)\alpha_k \|\nabla f(x_k)\|^2 + \frac{L\,\alpha_k^2\, M}{2}, && (10.9)\end{aligned}$$

where $M_G := M_V + \mu_G^2 \ge \mu^2 > 0$.

Proof. From (10.7) we conclude with (i) that

$$\begin{aligned}\mathbb{E}_{\xi_k} f(x_{k+1}) - f(x_k) &\le -\alpha_k \nabla f(x_k)^\top \mathbb{E}_{\xi_k} g(x_k, \xi_k) + \frac{L\,\alpha_k^2}{2} \mathbb{E}_{\xi_k} \|g(x_k, \xi_k)\|^2 \\ &\le -\alpha_k \mu \|\nabla f(x_k)\| + \frac{L\,\alpha_k^2}{2} \mathbb{E}_{\xi_k} \|g(x_k, \xi_k)\|^2, \end{aligned} \tag{10.10}$$

which is (10.8).

From (iii) and (ii) we deduce

$$\begin{aligned}\mathbb{E}_{\xi_k} \|g(x_k, \xi_k)\|^2 &\le M + M_V \|\nabla f(x_k)\|^2 + \left\|\mathbb{E}_{\xi_k} g(x_k, \xi_k)\right\|^2 \\ &\le M + M_V \|\nabla f(x_k)\|^2 + \mu_G^2 \|\nabla f(x_k)\|^2 \\ &= M + M_G \|\nabla f(x_k)\|^2 .\end{aligned}$$

Eq. (10.9) follows now with (10.10). □

In what follows we will use the total expectation $\mathbb{E} f(x_k) = \mathbb{E}_{\xi_1} \dots \mathbb{E}_{\xi_k} f(x_k)$.

Theorem 10.13. *Suppose that* $\alpha_k > 0$ *so that*

$$\sum_k \alpha_k = \infty \text{ and } \sum_k \alpha_k^2 < \infty. \tag{10.11}$$

Then

$$\liminf_{k\to\infty} \mathbb{E} \|\nabla f(x_k)\|^2 = 0. \tag{10.12}$$

Proof. Taking *total* expectation in (10.9) we get, for k large enough (note, that $\frac{\alpha_k L M_G}{2} \xrightarrow[k\to\infty]{} 0$),

$$\begin{aligned}\mathbb{E} f(x_{k+1}) - \mathbb{E} f(x_k) &\le -\left(\mu - \frac{\alpha_k L M_G}{2}\right) \alpha_k \mathbb{E} \|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2 M}{2} \\ &\le -\frac{\mu\,\alpha_k}{2} \mathbb{E} \|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2 M}{2}.\end{aligned}$$

Without loss of generality we assume that the latter inequality holds for *all* $k \in \{1, 2, \dots, K\}$. Summing both inequalities gives

$$f_{\inf} - \mathbb{E} f(x_1) \le -\mathbb{E} f(x_{k+1}) - \mathbb{E} f(x_1) \le -\frac{\mu}{2} \sum_{k=1}^K \alpha_k \mathbb{E} \|\nabla f(x_k)\|^2 + \frac{L\,M}{2} \sum_{k=1}^K \alpha_k^2,$$

or

$$\sum_{k=1}^K \alpha_k \mathbb{E} \|\nabla f(x_k)\|^2 \le \frac{2}{\mu} \left(\mathbb{E} f(x_1) - f_{\inf}\right) + \frac{L\,M}{\mu} \sum_{k=1}^K \alpha_k^2.$$

It follows that

$$\sum_{k=1}^K \alpha_k \mathbb{E} \|\nabla f(x_k)\|^2 < \infty. \tag{10.13}$$

As well it follows that

$$\frac{1}{A_K}\sum_{k=1}^{K}\alpha_k\,\mathbb{E}\,\|\nabla f(x_k)\|^2 \xrightarrow[K\to\infty]{} 0, \tag{10.14}$$

where $A_K := \sum_{k=1}^{K}\alpha_k$.

Now suppose that (10.12) would not hold true, but this were a contradiction to (10.13). Hence the result. □

Corollary 10.14. *Choose the index $k(K) \in \{0,\ 1, \dots,\ K\}$ with probability $\frac{\alpha_k}{A_K}$. It holds that*

$$\left\|\nabla f(x_{k(K)})\right\| \xrightarrow[k\to\infty]{} 0 \tag{10.15}$$

in probability.

Proof. From Markov's inequality we have that

$$P\left(\|\nabla f(x_k)\| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\,\mathbb{E}\,\|\nabla f(x_k)\|^2 \xrightarrow[k\to\infty]{} 0$$

by (10.14). □

Corollary 10.15. *If $f \in C^2$ and $x \mapsto \|\nabla f(x_k)\|$ has Lipschitz derivatives, then*

$$\lim_{k\to\infty}\mathbb{E}\,\|\nabla f(x_k)\|^2 = 0.$$

By employing Doob's martingale convergence theorems it is possible to establish almost sure convergence in (10.15).

11 Entropy and information

11.1 ENTROPY

Let P ($P(\mathrm{d}x) = p(x)\,\mathrm{d}x$ or $P = \sum_i p_i\,\delta_{x_i}$, resp.) and Q ($Q(\mathrm{d}x) = q(x)\,\mathrm{d}x$, $Q = \sum_i q_i\,\delta_{x_i}$, resp.) be probability measures.

Definition 11.1 (Cross entropy, differential entropy)**.** The *entropy* is

$$H(P) \coloneqq -\sum_i p_i \log p_i \qquad \left(H(P) \coloneqq -\int p(x)\log p(x)\,\mathrm{d}x,\ \text{resp.}\right), \tag{11.1}$$

the *cross entropy* is

$$H(P,Q) \coloneqq -\sum_i p_i \log q_i \qquad \left(H(P,Q) \coloneqq -\int p(x)\log q(x)\,\mathrm{d}x,\ \text{resp.}\right).$$

Note, that $H(P) = H(P,P)$.

The quantity $I(i) \coloneqq -\log q_i$ ($I(x) \coloneqq -\log q(x)$) is also called *self-information* or *information content.*[1]

Remark 11.2*.* The entropy H (cf. (11.1)) does *not* involve the locations x_i. Further, as $p_i > 0$, the entropy (and the cross entropy) is nonnegative.

Example 11.3. Consider the distribution $P(\{x_1\}) = p$ and $P(\{x_2\}) = 1-p$, then $H = -p\log p - (1-p)\log(1-p)$.

Corollary 11.4 (Log sum inequality)**.** *Let $a_i, b_i > 0$ and $a \coloneqq \sum_i a_i$ ($b \coloneqq \sum_i b_i$, resp.). It holds that*

$$\sum_i a_i \log\frac{a_i}{b_i} \ge a\log\frac{a}{b}. \tag{11.2}$$

Equality holds iff $\frac{a_i}{b_i}$ = const for all i.

Proof. The function $\varphi(x) \coloneqq x\cdot\log x$ is convex in $\mathbb{R}_{\ge 0}$ (indeed, $\varphi''(x) = \frac{1}{x} > 0$ for $x > 0$). With Jensen's inequality,[2]

$$\sum_i a_i\log\frac{a_i}{b_i} = b\cdot\sum_i \frac{b_i}{b}\varphi\left(\frac{a_i}{b_i}\right) \ge b\cdot\varphi\left(\sum_i \frac{b_i}{b}\frac{a_i}{b_i}\right) = b\,\varphi\left(\frac{a}{b}\right) = a\log\frac{a}{b}$$

and hence the assertion. □

[1] Informationsgehalt, dt.

[2] Jensens inequality states that $\varphi(\mathbb{E}\,X) \le \mathbb{E}\,\varphi(X)$, provided that φ is convex.

Remark 11.5. The entropy of the uniform distribution $U(\{x_1,\dots,x_n\})$ with $P(\{x_i\})=\frac{1}{n}$ is $H(P)=-\sum_i \frac{1}{n}\log\frac{1}{n}=\log n$.

Proposition 11.6. *For a discrete random variable with n possible realizations it holds that $0\le H(P)\le \log n$.*

Proof. Note first that $p\log p\le 0$ for $p\in(0,1)$ and thus $H=-\sum_i p_i\log p_i\ge 0$.

With $a_i:=p_i$ and $b_i:=1$ (i.e., $a=1$ and $b=n$) the log sum inequality (11.2) states that

$$\sum_i p_i\log p_i=\sum_i p_i\log\frac{p_i}{1}\ge 1\cdot\log\frac{1}{n}=-\log n$$

and thus $H(P)=-\sum p_i\log p_i\le\log n$. □

Remark 11.7. The entropy may be negative for continuous distributions. Indeed, for the uniform distribution $U[a,b]$ with density $p(x)=\frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$ it holds that $H=-\int_a^b\log\frac{1}{b-a}\frac{\mathrm{d}x}{b-a}=\log(b-a)$.

Theorem 11.8. *The uniform distribution has largest entropy among all distributions with fixed support.*

Proof. For discrete distributions the statement follows from Proposition 11.6 and Remark 11.5.

As for continuous distributions (with support $[a,b]$) we have with Jensen's inequality

$$\begin{aligned}\int_a^b p(x)\log p(x)\,\mathrm{d}x &= (b-a)\frac{1}{b-a}\int_a^b \varphi\big(p(x)\big)\,\mathrm{d}x\\ &\ge (b-a)\,\varphi\left(\frac{1}{b-a}\int_a^b p(x)\,\mathrm{d}x\right)\\ &=(b-a)\,\varphi\left(\frac{1}{b-a}\right)\\ &=(b-a)\,\frac{1}{b-a}\log\frac{1}{b-a}\\ &=-\log(b-a),\end{aligned}$$

from which the assertion is immediate with Remark 11.7. □

Theorem 11.9. *The probability measure with maximum entropy given moment constraints $\mathbb{E}\,r_i(X)=\alpha_i$, $i=1,\dots,n$, has density $p(x)=\frac{e^{-\lambda_1 r_1(x)-\dots-\lambda_n r_n(x)}}{e^{\lambda_0-1}}$ for $\lambda_0,\lambda_1,\dots,\lambda_n$ appropriate.*

Proof. The Lagrangian function is

$$L(x;\lambda_1,\dots,\lambda_n)=-\int p(x)\log p(x)\mathrm{d}x+\lambda_0\left(1-\int p(x)\mathrm{d}x\right)+\sum_{i=1}^n\lambda_i\left(\alpha_i-\int p(x)r_i(x)\mathrm{d}x\right).$$

Differentiating with respect to $p(x)$ (without going into detail; recall, that we are interested in the optimal p) reveals the first order conditions

$$0 = \frac{\partial L}{\partial p(x)} = -\log p(x) - 1 - \lambda_0 - \sum_{i=1}^n \lambda_i\, r_i(x)$$

and hence the result. □

Corollary 11.10 (Normal distribution)**.** *The normal distribution $\mathcal{N}(\mu, \sigma^2)$ attains maximal entropy given the variance σ^2; the maximal entropy is $\frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2} \approx 1.42 + \log\sigma$.*

Proof. Choose $r_1(x) = x$ and $r_2(x) = x^2$. From the preceding theorem we have that

$$p(x) = e^{1-\lambda_0-\lambda_1 x-\lambda_2 x^2} = e^{-\lambda_2\left(x+\lambda_1/2\lambda_2\right)^2+\lambda_1^2/4\lambda_2^2+1-\lambda_0}$$

is optimal, the optimal density p thus is the density of a normal distribution. To meet the moment constraints, the parameters λ_0, λ_1 and λ_2 have to be adjusted accordingly. The only normal distribution meeting all constraints is $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$. The maximal entropy is

$$-\int \underbrace{\left(\log\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu)^2\right)}_{\log p(x)} p(x)\,\mathrm{d}x = \frac{\log\left(2\pi\sigma^2\right)}{2} + \frac{1}{2}$$

and thus the assertion. □

Corollary 11.11. *The Laplace distribution with density $p(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$ maximizes the entropy given the constraint $\mathbb{E}\,|x-\mu| = b$.*

Remark 11.12 (Relation between continuous and discrete entropy). For continuous densities $p(x)$ and $q(x)$ set $x_i := i\cdot\Delta$, $p_i := \int_{x_i}^{x_{i+1}} p(x)\,\mathrm{d}x$ and $q_i := \int_{x_i}^{x_{i+1}} q(x)\,\mathrm{d}x$ for all $i \in \mathbb{Z}$. For the approximating measures $P_\Delta := \sum_{i\in\mathbb{Z}} p_i\,\delta_{x_i}$ and $Q_\Delta := \sum_{i\in\mathbb{Z}} q_i\,\delta_{x_i}$ it holds that

$$\begin{aligned}
H(P_\Delta, Q_\Delta) &= -\sum_i p_i \log q_i \\
&\approx -\sum_i \Delta\cdot p(x_i)\log\left(\Delta\cdot q(x_i)\right) \\
&= -\sum_i \Delta\cdot p(x_i)\log q(x_i) - \sum_i \Delta\cdot p(x_i)\log\Delta \\
&\approx -\int p(x)\log q(x)\mathrm{d}x - \log\Delta \\
&= H(P,Q) - \log\Delta
\end{aligned}$$

for $\Delta > 0$ small.

Proposition 11.13. *Let π have marginals P and Q, then*

$$\max\big(H(P), H(Q)\big) \le H(\pi) \le H(P \otimes Q) = H(P) + H(Q),$$

where $P \otimes Q$ is the product measure.[3]

Proof. Set $a_{ij} := \pi_{ij}$, $b_{ij} := p_i\, q_j$ and observe that $a = \sum_{ij} \pi_{ij} = 1$ and $b = \sum_{ij} p_i\, q_j = 1$. The log sum inequality (11.2) (with double index) gives $\sum_{ij} \pi_{ij} \log \frac{\pi_{ij}}{p_i\, q_j} \ge 1 \log \frac{1}{1} = 0$. That is,

$$\sum_{ij} \pi_{ij} \log \pi_{ij} \ge \sum_{ij} \pi_{ij} \log p_i + \sum_{ij} \pi_{ij} \log q_j = \sum_i p_i \log p_i + \sum_j q_j \log q_j,$$

or $H(\pi) \le H(P) + H(Q)$, the second inequality. Equality holds for $a_{ij} = b_{ij}$, i.e., the product measure.

Further recall that $p_i = \sum_j \pi_{ij}$ and that

$$\begin{aligned} H(\pi) &= -\sum_i p_i \log p_i - \sum_{ij} \pi_{ij} \log \pi_{ij} + \sum_{ij} \pi_{ij} \log p_i \\ &= -\sum_i p_i \log p_i - \sum_{ij} \pi_{ij} \underbrace{\log \frac{\pi_{ij}}{p_i}}_{\le 0} \\ &\ge -\sum_i p_i \log p_i \\ &= H(P), \end{aligned}$$

from which the remaining assertion follows. □

Every bivariate measure π can be disintegrated as $\pi(A \times B) = \sum_{i \in A} P(B \mid i)\, P(i)$ (or $\pi(A \times B) = \int_A P(B \mid x)\, P(\mathrm{d}x)$), where P is the marginal measure.

Proposition 11.14. *Let π have marginal P and σ have marginal Q. It holds that*

$$H(\pi, \sigma) = H(P, Q) + \sum_i P_i \cdot H\big(P(\cdot|i), Q(\cdot|i)\big).$$

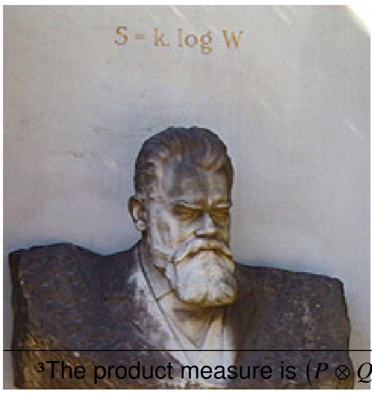


Figure 11.1: Ludwig Boltzmann, 1844–1906

[3] The product measure is $(P \otimes Q)(A \times B) := P(A) \cdot Q(B)$.

Proof: Indeed,

$$
\begin{aligned}
H(P,Q)+&\sum_i P_i\cdot H\big(P(\cdot|x_i),Q(\cdot|x_i)\big)\\
&=-\sum_i P_i\log Q_i-\sum_i P_i\sum_j\frac{\pi_{ij}}{P_i}\log\frac{\sigma_{ij}}{Q_i}\\
&=-\sum_i P_i\log Q_i-\sum_i\sum_j\pi_{ij}\log\sigma_{ij}+\sum_i\sum_j\pi_{ij}\log Q_i\\
&=-\sum_i P_i\log Q_i-\sum_{i,j}\pi_{ij}\log\sigma_{ij}+\sum_i P_i\log Q_i\\
&=-\sum_{i,j}\pi_{ij}\log\sigma_{ij}\\
&=H(\pi,\sigma),
\end{aligned}
$$

11.2 RELATIVE ENTROPY

Definition 11.15 (Kullback[4]–Leibler[5] divergence, relative entropy)**.** For probability measures P and Q we define

$$D\big(P\|Q\big)\coloneqq H(P,Q)-H(P);$$

for $P\not\ll Q$ we set $D\big(P\|Q\big)\coloneqq\infty$.

Divergence $D\big(P\parallel Q\big)$ is often called Kullback–Leibler divergence and also denoted as $D\big(P\parallel Q\big)=D_{KL}\big(P\parallel Q\big)=KL\big(P\parallel Q\big)$.

In the context of machine learning, $D(P\|Q)$ is often called the *information gain* achieved if Q is used instead of P. By analogy with information theory, it is also called the *relative entropy* of P with respect to of Q.

Example 11.16. Let Q denote the counting measure, $Q(\{x_i\})=\frac{1}{n}$ for all $i=1,\dots,n$. Then $D\big(P\|Q\big)=\sum_i p_i\log\frac{p_i}{1/n}=\sum_i p_i\log p_i+\sum_i p_i\log n=\sum_i p_i\log p_i+\log n$ and $D\big(Q\|P\big)=\sum_i\frac{1}{n}\log\frac{1/n}{p_i}=-\log n-\frac{1}{n}\sum_i\log p_i$.

Remark 11.17*.* The Kullback–Leibler divergence is asymmetric in general: $D\big(P\|Q\big)\neq D\big(Q\|P\big)$.

Theorem 11.18. *Let P and Q be probability measures on the same space with $\mathrm{d}P=Z\,\mathrm{d}Q$. The* divergence *between P and Q is*

$$D\big(P\parallel Q\big)\coloneqq\mathbb{E}_Q\left(Z\log Z\right)=\int Z\log Z\,\mathrm{d}Q=\int\log Z\,\mathrm{d}P=\mathbb{E}_P\log Z.$$

[4]Solomon Kullback, 1907–1994, American mathematician
[5]Richard Leibler, 1914–2003, American mathematician

Proof. For discrete measures let $P = \sum_i p_i\,\delta_{x_i}$ and $Q = \sum_i q_i\,\delta_{x_i}$. Note, that $Z(x_i) = \frac{\mathrm{d}P}{\mathrm{d}Q}(x_i) = \frac{p_i}{q_i}$ and thus

$$D(P\|Q) = \sum_i p_i\,\log p_i - \sum_i p_i\,\log q_i = \sum_i p_i \log\frac{p_i}{q_i} = \mathbb{E}_P \log Z.$$

For continuous measures $Q(\mathrm{d}x) = q(x)\,\mathrm{d}x$ and $P(\mathrm{d}x) = p(x)\,\mathrm{d}x = \frac{p(x)}{q(x)}q(x)\,\mathrm{d}x = \frac{p(x)}{q(x)}Q(\mathrm{d}x)$ we find the likelihood ratio $Z(x) = \frac{p(x)}{q(x)}$ so that

$$D(P\|Q) = \int p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x = \int\left(\frac{p(x)}{q(x)}\log\frac{p(x)}{q(x)}\right)q(x)\,\mathrm{d}x = \mathbb{E}_Q\, Z\log Z \tag{11.3}$$

and thus the assertion. □

Definition 11.19. More generally, for f convex with $f(1) = 0$, the f*-divergence* between P and Q is

$$D_f(P\|Q) \coloneqq \mathbb{E}_Q\, f(Z).$$

Remark 11.20. The Kullback–Leibler divergence is the f-divergence for $f(x) \coloneqq x\cdot\log x$.

Proposition 11.21 (Gibb's inequality)**.** *It holds that $D_f(P\|Q) \ge 0$. Equality holds iff $P = Q$.*

Proof. Note first that Z is a density with respect to Q. Indeed, $Z \ge 0$ and $\mathbb{E}_Q Z = \int\frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q = \int \mathrm{d}P = 1$. The function f is convex (in particular, $f\colon x\mapsto x\cdot\log x$ is convex). From Jensen's inequality it follows that

$$D(P\|Q) = \mathbb{E}_Q\, f(Z) \ge f(\mathbb{E}_Q\, Z) = f(1) = 0,$$

the assertion. □

Corollary 11.22. *It holds that $H(P,Q) \ge H(P)$ and thus $D(P\|Q) \ge 0$.*

Theorem 11.23 (Product measures)**.** *Let P_1, P_2, Q_1 and Q_2 be measures, then it holds that*

$$D(P_1\otimes P_2 \parallel Q_1\otimes Q_2) = D(P_1 \parallel Q_1) + D(P_2\parallel Q_2).$$

Proof. The Radon–Nikodym derivative is

$$\begin{aligned}(P_1\otimes P_2)(\mathrm{d}x,\mathrm{d}y) &= P_1(\mathrm{d}x)\cdot P_2(\mathrm{d}y)\\ &= Z_1(x)Q_1(\mathrm{d}x)\cdot Z_2(y)Q_1(\mathrm{d}x)\\ &= Z_1(x)Z_2(y)(Q_1\otimes Q_2)(\mathrm{d}x,\mathrm{d}y).\end{aligned}$$

It follows that

$$\begin{aligned}
D\big(P_1 \otimes P_2 \| Q_1 \otimes Q_2\big) &= \iint Z_1(x) Z_2(y) \log\big(Z_1(x) Z_2(y)\big) Q_1(\mathrm{d}x) Q_2(\mathrm{d}y) \\
&= \iint Z_1(x) Z_2(y) \log\big(Z_1(x)\big) Q_1(\mathrm{d}x) Q_2(\mathrm{d}y) \\
&\qquad + \iint Z_1(x) Z_2(y) \log\big(Z_2(y)\big) Q_1(\mathrm{d}x) Q_2(\mathrm{d}y) \\
&= \int Z_1(y) \log\big(Z_1(x)\big) Q_1(\mathrm{d}x) \cdot \int Z_2(y) Q_2(\mathrm{d}y) \\
&\qquad + \int Z_1(y) Q_1(\mathrm{d}x) \cdot \int Z_2(y) \log\big(Z_2(y)\big) Q_2(\mathrm{d}y) \\
&= D\big(P_1 \| Q_1\big) + D\big(P_2 \| Q_2\big),
\end{aligned}$$

the assertion. □

Theorem 11.24 (Convexity)**.** *For $\lambda \in [0,1]$ it holds that*

$$D\big((1-\lambda)P_0 + \lambda P_1 \;\|\; (1-\lambda)Q_0 + \lambda Q_1\big) \le (1-\lambda)\, D\big(P_0 \;\|\; Q_0\big) + \lambda\, D\big(P_1 \;\|\; Q_1\big).$$

Proof. The Radon–Nikodym derivative is $\frac{\mathrm{d}(1-\lambda)P_0+\lambda P_1}{\mathrm{d}(1-\lambda)Q_0+\lambda Q_1} = \frac{(1-\lambda)p_0+\lambda p_1}{(1-\lambda)q_0+\lambda q_1}$. By the log sum inequality (Corollary 11.4) we find that

$$\begin{aligned}
\big((1-\lambda)p_0 + \lambda p_1\big) \log \frac{(1-\lambda)p_0 + \lambda p_1}{(1-\lambda)q_0 + \lambda q_1} &\le \\
&\le (1-\lambda) p_1 \log \frac{(1-\lambda)p_1}{(1-\lambda)q_1} + \lambda p_0 \log \frac{\lambda p_0}{\lambda q_0}.
\end{aligned}$$

Integration gives the desired inequality. □

Theorem 11.25. *Let π be a bivariate measure with marginals P and Q. It holds that*

$$D\big(\pi \| P \otimes Q\big) = H(P) + H(Q) - H(\pi). \tag{11.4}$$

Proof. Indeed,

$$D\big(\pi \| P \otimes Q\big) = \sum_{i,j} \pi_{ij} \log \frac{\pi_{ij}}{p_i\, q_j} = \sum_{i,j} \pi_{ij} \log \pi_{i,j} - \sum_{i,j} \pi_{ij} \log p_i - \sum_{i,j} \pi_{ij} \log q_j.$$

As the marginals of π coincide with P and Q it follows that

$$\begin{aligned}
D\big(\pi \| P \otimes Q\big) &= \sum_{i,j} \pi_{ij} \log \pi_{ij} - \sum_i p_i \log p_i - \sum_j q_j \log q_j \\
&= H(P) + H(Q) - H(\pi),
\end{aligned}$$

the assertion. □

Theorem 11.26 (Data processing theorem)**.** *Let T be a measurable. Then it holds that*

$$D\big(P^T \parallel Q^T\big) \le D\big(P \parallel Q\big).$$

Kullback comments on the preceding theorem,

> "statistical processing will not increase the information (discrimination information) contained in the data".

Proof. Denote by p and q (p^T, q^T, resp.) the densities of P and Q (the push-forward P^T, Q^T, resp.). From the definition and by changing the variables we have that

$$D\big(P^T\|Q^T\big) = \mathbb{E}_{P^T} \log\frac{P^T}{Q^T} = \int \log\frac{p^T(y)}{q^T(y)}\,P^T(\mathrm{d}y) = \int \log\frac{p^T\big(T(x)\big)}{q^T\big(T(x)\big)}\,P(\mathrm{d}x),$$

and thus

$$\begin{aligned} D\big(P\|Q\big) - D\big(P^T\|Q^T\big) &= \int \log\frac{p(x)}{q(x)} - \log\frac{p^T\big(T(x)\big)}{q^T\big(T(x)\big)}\,P(\mathrm{d}x) \\ &= \int p(x)\log\frac{p(x)\cdot q^T\big(T(x)\big)}{q(x)\cdot p^T\big(T(x)\big)}\,\mathrm{d}x. \end{aligned}$$

Now set $s(x) := \frac{p(x)\cdot q^T(T(x))}{q(x)\cdot p^T(T(x))}$ so that

$$\begin{aligned} D\big(P\|Q\big) - D\big(P^T\|Q^T\big) &= \int \frac{q(x)\cdot p^T\big(T(x)\big)}{q^T\big(T(x)\big)} s(x)\log s(x)\,\mathrm{d}x \\ &= \int s(x)\log s(x)\,\mu(\mathrm{d}x), \end{aligned} \tag{11.5}$$

where $\mu(\mathrm{d}x) = \frac{q(x)\cdot p^T(T(x))}{q^T(T(x))}\,\mathrm{d}x$.

With $f(x) = x\cdot\log x$ we have the Taylor series expansion

$$s(x)\log s(x) = f\big(s(x)\big) = \underbrace{f(1)}_{=0} + \underbrace{f'(1)}_{=1}\big(s(x)-1\big) + \frac{1}{2}f''\big(h(x)\big)\big(s(x)-1\big)^2, \tag{11.6}$$

where $h(x) \in \big(1, s(x)\big)$; as $s(x) > 0$ we also have $h(x) > 0$. Now note that

$$\int s(x)\,\mathrm{d}\mu(x) = \int \frac{p(x)\cdot q^T\big(T(x)\big)}{q(x)\cdot p^T\big(T(x)\big)}\cdot\frac{q(x)\cdot p^T\big(T(x)\big)}{q^T\big(T(x)\big)}\,\mathrm{d}x = \int p(x)\,\mathrm{d}x = 1$$

and $f''(x) = \frac{1}{x} > 0$ for $x > 0$ and thus the assertion follows with (11.5) and (11.6). ◻

Theorem 11.27 (Pinsker's inequality[6])**.** *It holds that*

$$\|P - Q\|_\infty \le \sqrt{\frac{1}{2} D(P \parallel Q)},$$

where

$$\|P - Q\|_\infty := \sup\{|P(A) - Q(A)| : A \text{ measurable}\}$$

is the total variation distance.

Proof. Cf. Tsybakov [17]. □

11.3 GIBBS MEASURES

Theorem 11.28. *The minimum of the entropy $\mathbb{E}\, Z \log Z$ subject to the moment constraint $\mathbb{E}\, YZ = E$ and $\mathbb{E}\, Z = 1$ is attained at $Z^* = \frac{\mathbb{E}\, Y e^{\lambda Y}}{\mathbb{E}\, e^{\lambda Y}}$, where λ is chosen so that $\mathbb{E}\, Z = E$.*

Proof. The Lagrangian is

$$L(\lambda, \gamma, Z) = \mathbb{E}\, Z \log Z + \lambda\, (\mathbb{E}\, YZ - E) + \gamma\, (\mathbb{E}\, Z - 1)\,.$$

The derivatives with respect to the parameters are

$$\begin{aligned} \frac{\partial}{\partial \lambda} L(Z; \lambda, \gamma) &= \mathbb{E}\, YZ - E = 0, \\ \frac{\partial}{\partial \gamma} L(Z; \lambda, \gamma) &= \mathbb{E}\, Z - 1 = 0 \text{ and} \\ \frac{\partial}{\partial Z} L(Z; \lambda, \gamma)(H) &= \mathbb{E}\, (\log Z + 1 + \lambda Y + \gamma \mathbb{1})\, H = 0 \end{aligned}$$

for all directions H, and thus $Z = \exp(-1 - \gamma - \lambda Y)$. It follows from $\mathbb{E}\, Z = 1$ that $Z = \frac{e^{-\lambda Y}}{\mathbb{E}\, e^{-\lambda Y}}$, where λ is chosen so that $\frac{\mathbb{E}\, Y e^{-\lambda Y}}{\mathbb{E}\, e^{-\lambda Y}} = E$. □

Corollary 11.29 (Maximum entropy, discrete version)**.** *The maximum among all probabilities $p_i \ge 0$ so that $\sum_i p_i\, y_i = E$ with respect to $H(P) = -\sum_i p_i \log p_i$ is attained at $p_i = \frac{e^{-\lambda y_i}}{\sum_j e^{-\lambda y_j}}$ for some appropriate $\lambda \in \mathbb{R}$.*

Definition 11.30 (Gibbs measure, Boltzmann distribution)**.** The Gibbs measure has the density $Z\, \mathrm{d}P = \frac{e^{-\lambda Y}}{Z(\lambda)}\, \mathrm{d}P$, where $Z(\lambda) := \mathbb{E}\, e^{-\lambda Y}$ is the *partition function*. For the Boltzmann distribution the parameter is the inverse temperature, $\lambda = \frac{1}{kT}$.

Here, Y can be interpreted as energy with average energy E; states with low energy are more likely, as states with high energy cool down to lower energy.

[6] Mark Semenovich Pinsker, 1925–2003, Russian mathematician

Definition 11.31 (Gibbs softmax, aka. LogSumExp)**.** The Gibbs softmax is

$$\max_\beta(x_1,\dots,x_n) \coloneqq \frac{1}{\beta}\log\sum_{i=1}^n e^{\beta x_i} \tag{11.7}$$

and the softmin is

$$\min_\beta(x_1,\dots,x_n) \coloneqq -\frac{1}{\beta}\log\sum_{i=1}^n e^{-\beta x_i}.$$

11.4 REFERENCES

A comprehensive source for information theory is the book Cover and Thomas [5]. Some parts here follow Kersting and Wakolbinger [8, Chapter VI].

11.5 PROBLEMS

Exercise 11.1. *Verify that the Kullback–Leibler divergence is not symmetric, cf. Remark 11.17.*

Exercise 11.2. *Compare the Gibbs softmax (softmin, resp.) with*

$$\max_\beta(x_1,\dots,x_n) \coloneqq \frac{\sum_{i=1}^n x_i\, e^{\beta x_i}}{\sum_{i=1}^n e^{\beta x_i}}$$

and

$$\min_\beta(x_1,\dots,x_n) \coloneqq \frac{\sum_{i=1}^n x_i\, e^{-\beta x_i}}{\sum_{i=1}^n e^{-\beta x_i}}.$$

12 Cluster analysis

Definition 12.1 (Wasserstein distance)**.** Let P and Q be probability measures. The Wasserstein distance is

$$d(P,Q) := \inf \left(\iint d(x,y)^r \, \pi(\mathrm{dx},\mathrm{dy}) \right)^{1/r}, \tag{12.1}$$

where the infimum is among all bivariate probability measures π with marginals P and Q, i.e.,

$$\begin{aligned} \pi(A \times Y) &= P(A) \text{ and} \\ \pi(X \times B) &= Q(B). \end{aligned}$$

The discrete version of the Wasserstein distance reads

$$\begin{aligned} \text{minimize } & \sum_{i,j} \pi_{ij} d_{ij}^r \\ \text{subject to } & \sum_j \pi_{ij} = p_i, \\ & \sum_i \pi_{ij} = q_j, \\ & \pi_{ij} \geq 0. \end{aligned}$$

12.1 FAST COMPUTATION

Definition 12.2 (Sinkhorn distance)**.** The Sinkhorn distance $d_\alpha(P,Q)$ is (12.1) above, except that π satisfies the additional constraint $KL\,(\pi \mid P \otimes Q) \leq \alpha$.

Remark 12.3. Recall from (11.4) that

$$\begin{aligned} D_{KL}\,(\pi \mid P \otimes Q) &= \sum_{i,j} \pi_{ij} \log \frac{\pi_{ij}}{p_i\, q_j} \\ &= \sum_{i,j} \pi_{ij} \left(\log \pi_{ij} - \log p_i - \log q_j \right) \\ &= \sum_{i,j} \pi_{ij} \log \pi_{ij} - \sum_i p_i \log p_i - \sum_j q_j \log q_j \\ &= H(P) + H(Q) - H(\pi). \end{aligned}$$

Definition 12.4 (Regularized Sinkhorn distance)**.** The regularized Sinkhorn distance is given by

$$\begin{aligned} \text{minimize} \quad & \sum_{i,j} \pi_{ij} d_{ij}^r + \frac{1}{\lambda} \sum_{i,j} \pi_{ij} \log \pi_{ij} \\ \text{subject to} \quad & \sum_j \pi_{ij} = p_i, \\ & \sum_i \pi_{ij} = q_j, \\ & \pi_{ij} \geq 0, \end{aligned} \tag{12.2}$$

where $\lambda > 0$ is a regularization parameter.

Proposition 12.5. *There are vectors β and γ so that the optimal π in the Sinkhorn distance ((12.2) or Definition 12.2) satisfies*

$$\pi = \operatorname{diag}(\beta) \cdot K \cdot \operatorname{diag}(\gamma), \qquad K_{ij} := e^{-\lambda d_{ij}}.$$

They can be found by Sinkhorn's fixed point iteration by re-scaling the rows and columns successively. To this end set $(r_{n+1}, c_{n+1}) := (r_n./Kc_n, c_n./r_nK)$, or $r_{n+2} = r_n./Kc_n./r_nK$.

Proof. Define the Lagrangian

$$\begin{aligned} L(\pi; \lambda, \beta, \gamma) := \sum_{i,j} \pi_{ij}\, d_{ij} + \frac{1}{\lambda} \left(H(P) + H(Q) - \alpha + \sum_{i,j} \pi_{ij} \log \pi_{ij} \right) \\ + \beta^\top (\pi \cdot \mathbb{1} - p) + \left(\mathbb{1}^\top \cdot \pi - q\right)^\top \gamma \end{aligned}$$

so that $\frac{\partial L}{\partial \pi_{ij}} = \frac{1}{\lambda}\left(\log \pi_{ij} + 1\right) + d_{ij} + \beta_i + \gamma_j = 0$, i.e.,

$$\pi_{ij} = e^{-\lambda \beta_i - 1/2} \cdot e^{-\lambda \cdot d_{ij}} \cdot e^{-\lambda \gamma_j - 1/2}. \tag{12.3}$$

λ is the Lagrange parameter associated with the constraint $KL\,(\pi \mid P \otimes Q) \leq \alpha$.

The Lagrangian for the regularized problem is

$$L(\pi; \lambda, \beta, \gamma) := \sum_{i,j} \pi_{ij}\, d_{ij} + \frac{1}{\lambda} \left(\sum_{i,j} \pi_{ij} \log \pi_{ij} \right) + \beta^\top (\pi \cdot \mathbb{1} - p) + \left(\mathbb{1}^\top \cdot \pi - q\right)^\top \gamma$$

so that again $\frac{\partial L}{\partial \pi_{ij}} = \frac{1}{\lambda}\left(\log \pi_{ij} + 1\right) + d_{ij} + \beta_i + \gamma_j = 0$.

It follows from (12.3) that $\pi = \operatorname{diag}(\tilde\beta) \cdot K \cdot \operatorname{diag}(\tilde\gamma)$ for some vectors $\tilde\beta$ and $\tilde\gamma$, where $K_{ij} := e^{-\lambda d_{ij}}$ and β, γ are Lagrange parameters. □

12.2 REFERENCES

include Sinkhorn-Knopp algorithm and Gabriel Peyré, https://www.youtube.com/watch?v=4FtamHah29M.

Lorenz curve and Gini coefficient

13

> Jedenfalls bin ich überzeugt, daß *der* nicht würfelt.
>
> Albert Einstein, *Brief an Max Born*, 1926

13.1 LORENTZ CURVE

For nonnegative random variables the following are often considered in economics.

Definition 13.1. The Lorenz[1] curve is

$$L(p) := \frac{\int_0^p F_X^{-1}(u)\,\mathrm{d}u}{\int_0^1 F_X^{-1}(u)\,\mathrm{d}u}, \qquad p \in [0,1].$$

Remark 13.2. The Lorenz curve is convex and, provided that $X \geq 0$, $0 \leq L(p) \leq 1$. Further, $L(p) = 0$ if X is not integrable (i.e., $\mathbb{E}\, X = \infty$) and $p < 1$.

Definition 13.3. The Gini[2] coefficient is

$$G := 1 - 2 \cdot \int_0^1 L(p)\,\mathrm{d}p.$$

Remark 13.4. The Gini coefficient with $G \in [0,1]$ is a summary statistics of the Lorenz curve and a measure of inequality in a population. It is a measure of statistical dispersion (spread). $G = 0$ (or small) identifies an 'all are equal' (similar) distribution, while $G = 1$ (or large) identifies large deviations within the population.

Remark 13.5. Einkommensverteilung in Deutschland

Proposition 13.6. *Alternatively expressions for the Gini coefficient include (cf. Fig-*

[1] Max Otto Lorenz, 1876–1959, American economist
[2] Corrado Gini, 1884–1965, Italian statistician

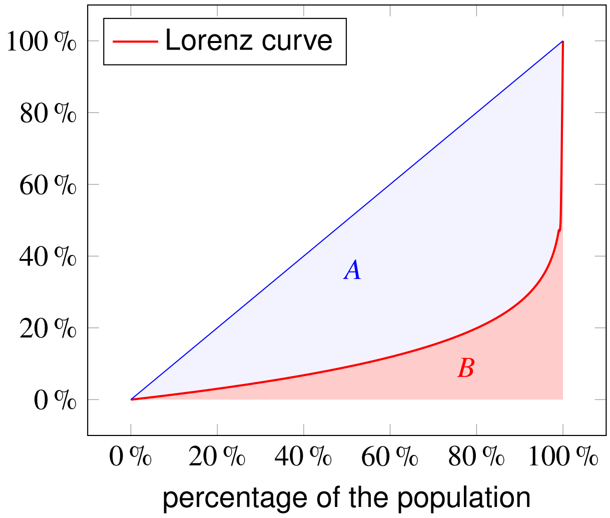


Figure 13.1: Lorenz curve of a Pareto distribution (Gini coefficient $G \approx 0.75$) exhibiting Pareto's 80/ 20 rule

ure 13.1)

$$
\begin{aligned}
G &= \frac{A}{A+B} = 2A = 1-2B \\
&= \frac{1}{\mu}\int_0^\infty F_X(x)\big(1-F_X(x)\big)\,\mathrm{d}x \qquad (13.1)\\
&= \frac{1}{\mu}\int_0^1 u(1-u)\,\mathrm{d}F_X^{-1}(u) \\
&= \frac{1}{2\mu}\int_0^\infty\int_0^\infty f(x)f(y)\,|x-y|\,\mathrm{d}x\,\mathrm{d}y \qquad (13.2)\\
&= \frac{1}{2\mu}\int_0^1\int_0^1 \left|F_X^{-1}(u)-F_X^{-1}(v)\right|\,\mathrm{d}u\,\mathrm{d}v \qquad (13.3)\\
&= \frac{1}{2\mu}\,\mathbb{E}\,|X-X'|, \qquad (13.4)
\end{aligned}
$$

where f_X is the density, $\mu = \mathbb{E}\,X$ the mean and X' an independent copy of X.

Remark 13.7. Recall, that $\operatorname{var} X = \frac{1}{2}\int_{-\infty}^\infty\int_{-\infty}^\infty f(x)f(y)(x-y)^2\,\mathrm{d}x\,\mathrm{d}y = \mathbb{E}\left(X-X'\right)^2$ and compare with (13.2) and (13.4).

Proof. Indeed,

$$\begin{aligned}\mu\cdot\int_0^1 L(p)\,\mathrm{d}p &= \int_0^1\int_0^p F^{-1}(u)\,\mathrm{d}u\mathrm{d}p = \int_0^1 F^{-1}(u)\cdot\int_u^1 1\,\mathrm{d}p\,\mathrm{d}u\\ &= \int_0^1 (1-u)F^{-1}(u)\,\mathrm{d}u \qquad (13.5)\\ &= \int_0^\infty \big(1-F(x)\big)f(x)\cdot x\,\mathrm{d}x = -\frac{(1-F(x))^2}{2}x\Big|_{x=0}^\infty + \int_0^\infty \frac{(1-F(x))^2}{2}\,\mathrm{d}x\\ &= \int_0^\infty \frac{(1-F(x))^2}{2}\,\mathrm{d}x.\end{aligned}$$

It follows further that $\mu G = \mu - 2\mu\int_0^1 L(p)\,\mathrm{d}p = \int_0^\infty 1-F(x)\,\mathrm{d}x - \int_0^\infty \big(1-F(x)\big)^2\,\mathrm{d}x = \int_0^\infty F(x)\big(1-F(x)\big)\,\mathrm{d}x$, which is (13.1).

Note next that

$$\begin{aligned}\int_0^1 \left|F^{-1}(u)-x\right|\,\mathrm{d}u &= \int_0^{F(x)} x-F^{-1}(u)\,\mathrm{d}u + \int_{F(x)}^1 F^{-1}(u)-x\,\mathrm{d}u\\ &= F(x)x-(1-F(x))x - \int_0^{F(x)} F^{-1}(u)\,\mathrm{d}u + \int_{F(x)}^1 F^{-1}(u)\,\mathrm{d}u\\ &= 2F(x)x - x - \int_0^{F(x)} F^{-1}(u)\,\mathrm{d}u + \mu - \int_0^{F(x)} F^{-1}(u)\,\mathrm{d}u\\ &= x - 2(1-F(x))x + \mu - 2\int_0^{F(x)} F^{-1}(u)\,\mathrm{d}u.\end{aligned}$$

Now substitute $x \leftarrow F^{-1}(v)$ so that

$$\int_0^1 \left|F^{-1}(u)-F^{-1}(v)\right|\,\mathrm{d}u = F^{-1}(v) - 2(1-v)F^{-1}(v) + \mu - 2\int_0^v F^{-1}(u)\,\mathrm{d}u$$

and thus further

$$\begin{aligned}&\int_0^1\int_0^1 \left|F^{-1}(u)-F^{-1}(v)\right|\,\mathrm{d}u\mathrm{d}v\\ &\quad= \int_0^1 F^{-1}(v)\,\mathrm{d}v - 2\int_0^1 (1-v)F^{-1}(v)\,\mathrm{d}v + \mu - 2\mu\int_0^1 L(p)\,\mathrm{d}p\\ &\quad\underset{(13.5)}{=} \mu - 2\mu\int_0^1 L(v)\,\mathrm{d}v + \mu - 2\mu\int_0^1 L(p)\,\mathrm{d}p = 2\mu G,\end{aligned}$$

and thus the assertion (13.3) follows. The others are obvious. □

Fact 13.8 (Statistics for Gini's coefficient)**.** *It follows from (13.2) and (13.5) and the fact that $F_n^{-1}(i/n) = X_{(i)}$ that a (biased) estimator for Gini's coefficient is*

$$G \underset{(13.3)}{\approx} \frac{\frac{1}{n^2}\sum_{i,j=1}^n \left|X_i-X_j\right|}{2\cdot\frac{1}{n}\sum_{i=1}^n X_i} \underset{(13.5)}{\approx} \frac{n+1}{n} - 2\frac{\frac{1}{n}\sum_{i=1}^n\left(1-\frac{i-1}{n}\right)X_{(i)}}{\frac{1}{n}\sum_{i=1}^n X_i}.$$

Distribution	pdf	Gini coefficient
Dirac delta distribution	$\delta(\cdot - x_0)$	0
Uniform distribution	$\mathbb{1}_{[a,b]}$	$\frac{b-a}{3(b+a)}$
Exponential distribution	$\lambda e^{-\lambda x},\ x \geq 0$	$\frac{1}{2}$
Pareto distribution	$\frac{\alpha\, x_{min}^{\alpha}}{x^{\alpha+1}},\ x \geq x_{min}$	$\begin{cases} \frac{1}{2\alpha-1} & \alpha \geq 1 \\ 1 & 0 < \alpha < 1 \end{cases}$
Weibull	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$	$1 - 2^{-k}$

Table 13.1: Gini coefficient of selected distributions

13.2 PROBLEMS

Exercise 13.1. *Verify that the Lorenz curve is* $L(p) = 1 - (1-p)^{1-\frac{1}{\alpha}}$ *for the Pareto distribution and* $p + (1-p)\log(1-p)$ *for the exponential distribution.*

Exercise 13.2. *Verify the Gini coefficients in Table 13.1.*

14 Stochastic global optimization

Zhigljavsky and Žilinskas [19]

15 Dynamic optimization

The Fleten et al. [7]

Bibliography

[1] R. N. Bhattacharya, L. Lin, and V. Patrangenaru. *A course in mathematical statistics and large sample theory*. Springer, 2016. doi:10.1007/978-1-4939-4032-5. 6

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York Inc., 2006. ISBN 0387310738. URL https://www.springer.com/de/book/9780387310732. 6

[3] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi:10.1137/16m1080173. 6, 55

[4] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer New York, 2nd edition, 1987. doi:10.1007/978-1-4419-0320-4. URL https://books.google.de/books?id=ZW_ThhYQiXIC. 56

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley John + Sons, 2006. ISBN 0471241954. 68

[6] N. A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, Inc., 1993. doi:10.1002/9781119115151. 6

[7] S.-E. Fleten, E. Haugom, A. Pichler, and C. J. Ullrich. Structural estimation of switching costs for peaking power plants. *European Journal on Operational Research*, 285(1):23–33, 2020. doi:10.1016/j.ejor.2019.03.031. 77

[8] G. Kersting and A. Wakolbinger. *Elementare Stochastik*. Springer Basel, 2010. doi:10.1007/978-3-0346-0414-7. 6, 68

[9] R. S. Liptser and A. N. Shiryaev. *Statistics of Random Processes I*. Springer, 2nd edition, 2001. doi:10.1007/978-3-662-13043-8. 33

[10] R. S. Liptser and A. N. Shiryaev. *Statistics of Random Processes II*. 2nd edition, 2001. doi:10.1007/978-3-662-10028-8. 56

[11] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi:10.1137/070704277. 53

[12] G. Ch. Pflug. *Optimization of Stochastic Models*, volume 373 of *The Kluwer International Series in Engineering and Computer Science*. Kluwer Academic Publishers, 1996. doi:10.1007/978-1-4613-1449-3. 6, 53

[13] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi:10.1214/aoms/1177729586. 55

[14] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming*. MOS-SIAM Series on Optimization. SIAM, third edition, 2021. doi:10.1137/1.9781611976595. 15

[15] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer New York, 2008. doi:10.1007/978-0-387-77242-4. 35

[16] A. C. Tamhane and D. D. Dunlop. *Statistics and Data Analysis: From Elementary to Intermediate*. PRENTICE HALL, 1999. ISBN 0137444265. 6

[17] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2008. doi:10.1007/b13794. 67

[18] D. Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, 1991. doi:10.1017/CBO9780511813658. URL http://books.google.com/books?id=RnOJeRpk0SEC. 56

[19] A. Zhigljavsky and A. Žilinskas. *Stochastic Global Optimization*. Springer US, 2008. doi:10.1007/978-0-387-74740-8. 75

Index