Data Science — Theory

Florian Kalinke
Karlsruhe Institute of Technology (KIT)

Last updated: February 24, 2026

1 Introduction

In data science, we typically wish—broadly speaking—to learn something from data in a principled fashion. *Learning* can have various meanings; for example, we can be interested in answering simple questions like "What is the number of distinct elements in our data set?" to more complicated ones like "Can I predict unseen data?" or "Is there some underlying structure in the data?"

At the heart of answering these and related questions are algorithms, which state *how* to arrive at an answer. In many cases, their application is sufficient for the practitioner. In these lecture notes, we investigate *why* these algorithms work, from a mathematical perspective. Our goal is a deep understanding of a few fundamental methods and also to shed light on cases where these algorithms fail.

Traditionally, learning algorithms are grouped into *supervised* and *unsupervised* approaches. In the former, one observes (input,output)-pairs and the aim is learning a function that maps the inputs to the outputs, subject to generalization to previously unseen data. Formally, the (input,output)-pairs take values in some product space $\mathcal{X} \times \mathcal{Y}$, and the target function is

$$f : \mathcal{X} \to \mathcal{Y}.$$

As an example, consider images of cats and dogs as inputs and the respective labels 'cat' or 'dog' as outputs. For learning, in this case a function f that maps images to class labels, we are given a set of *training data* $((x_i, y_i))_{i=1}^n$ with $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. Alternatively, think of associating with each student 'hours of lecture attendance', 'hours of exam preparation', 'number of questions asked', and 'final score'. Our goal is to predict the 'final score' (the dependent variable or target) from the other attributes (the independent variables or predictors).

The trivial solution to these tasks is memorizing the training data. But this does not generalize (i.e., perform well) to new data, and thus one 'restricts' the functions that can be learned. Note that in the first example, the output is assumed to be discrete while in the second example, we can assume the output to be continuous. We call the discrete case *classification* and the continuous case *regression*. We will see different examples illustrating solutions to the regression

problem throughout the lecture. Indeed, the preceding paragraph only aims to convey some basic intuition and we elaborate the subject in later sections.

For the latter approach, unsupervised learning, the goal is less precise. Here, one wants to find *structure* in the data $(x_i)_{i=1}^n$ with $x_i \in \mathcal{X}$. For example, consider customer data, and one wants to group customers with a similar behavior—a task known as *clustering*. Or, we wish to find a low-dimensional representation for visualization or pre-processing. While we do not discuss clustering, we investigate such dimensionality reduction schemes in these notes.

Before we start in earnest, we analyze an amazing algorithm published quite recently and take the opportunity to review a few properties of probability.

2 Appetizer: Count-distinct problem

In the count-distinct problem, we are given a stream $\mathcal{A} = \langle a_1, \dots, a_m \rangle$ of m elements and want to know the number of distinct elements $n = |\{a_1, \dots, a_m\}|$, processing $\mathcal{A}$ once and with limited memory only. For example, we might wish to know how many different users watched a YouTube video, how many different IP addresses connected to a website, or the number of distinct elements in a large database.

To account for the limited memory, we make a concession: We are willing to accept an (ϵ, δ)-approximation of n. Formally, the goal is to find c such that

$$\Pr\left[(1-\epsilon)n \le c \le (1+\epsilon)n\right] \ge 1 - \delta. \tag{1}$$

Loosely speaking, we aim to find a c that is close (in an ϵ-neighborhood) to n with high probability (larger than $1 - \delta$).

2.1 The CVM algorithm

While problem (1) has been studied for more than four decades and many algorithms exists, Chakraborty et al. [2022] recently proposed the following new solution, now known as CVM algorithm [Knuth, 2023].

Algorithm 1 Count-distinct algorithm [Chakraborty et al., 2022]

Input Stream $\mathcal{A} = \langle a_1, \dots, a_m \rangle$, ϵ, δ

1: **Initialize** $p \leftarrow 1$; $\mathcal{X} \leftarrow \emptyset$; thresh $= \lceil \frac{12}{\epsilon^2} \log_2 \frac{8m}{\delta} \rceil$
2: **for** $i = 1, \dots, m$ **do**
3: $\quad \mathcal{X} \leftarrow \mathcal{X} \setminus \{a_i\}$
4: $\quad$ With probability p, $\mathcal{X} \leftarrow \mathcal{X} \cup \{a_i\}$
5: $\quad$ **if** $|\mathcal{X}| =$ thresh **then**
6: $\quad\quad$ Throw away each element of $\mathcal{X}$ with probability $\frac{1}{2}$
7: $\quad\quad$ $p \leftarrow \frac{p}{2}$
8: $\quad\quad$ **If** $|\mathcal{X}| =$ thresh **then Output** $\perp$
9: **Output** $\frac{|\mathcal{X}|}{p}$

Before proving that Algorithm 1 satisfies (1), let us gain some intuition. To simplify following the algorithm, we now assume that each probabilistic step behaves as it would behave in expectation, e.g., if the algorithm "Throws away each element of $\mathcal{X}$ with probability $\frac{1}{2}$", we assume that $|\mathcal{X}|$ is halved. Similarly, if an element is added to $\mathcal{X}$ with probability p (line 4), we assume that every $1/p$-th element is added to $\mathcal{X}$.

We look at two cases.

(i) **All elements distinct.** Suppose $m = 16$, $a_i \neq a_j$ for any $i \neq j$ (implying $n = 16$), and thresh $= 4$. We now heuristically watch Algorithm 1 operate. Indeed, until $i = 4$, the algorithm fills the buffer $\mathcal{X}$, and after line 4, we have $i = 4$ and $\mathcal{X} = \{a_1, a_2, a_3, a_4\}$, when the condition in line 5 evaluates

to true for the first time. Hence, we halve $|\mathcal{X}|$, say, $\mathcal{X} = \{a_1, a_4\}$, and $p = 1/2$.

Continuing right until after line 4 for $i = 5, 6, 7, 8$, by our assumptions, $\mathcal{X} = \{a_1, a_4, a_6, a_8\}$ as every second element is added. Throwing a random half of the elements away and halving p, we have, say, $\mathcal{X} = \{a_4, a_6\}$, and $p = 1/4$.

For $i = 9, \ldots, 16$, the algorithm continues similarly, adding a_{12} and a_{16} to $\mathcal{X}$. When all $m = 16$ elements have been processed, $\mathcal{X}$ and p are halved one last time, i.e., $|\mathcal{X}| = 2$ and $p = 1/8$. The algorithm outputs $\frac{|\mathcal{X}|}{p} = 16$, returning precisely n.

(ii) **Every element three times.** Suppose $m = 16$, the stream is $\mathcal{A} = \langle a_1, a_1, a_1, a_2, \ldots, a_5, a_5, a_5, a_6 \rangle$, $a_i \neq a_j$ for any $i \neq j$ ($n = 6$), and thresh $= 4$.

For the first 9 elements, the algorithm fills the buffer $\mathcal{X}$ so that we have $\mathcal{X} = \{a_1, a_2, a_3\}$. Summarizing its state at the end of each iteration, the algorithm continues as follows, where for $p = 1/2$ we again add only every second element (note the randomness involved; all steps depend on the random halving performed for $i = 10$)

$$
\begin{array}{lll}
i = 10; & \mathcal{X} = \{a_1, a_4\}; & p = 1/2; \\
i = 11; & \mathcal{X} = \{a_1\}; & p = 1/2; \\
i = 12; & \mathcal{X} = \{a_1, a_4\}; & p = 1/2; \\
i = 13; & \mathcal{X} = \{a_1, a_4\}; & p = 1/2; \\
i = 14; & \mathcal{X} = \{a_1, a_4, a_5\}; & p = 1/2; \\
i = 15; & \mathcal{X} = \{a_1, a_4\}; & p = 1/2; \\
i = 16; & \mathcal{X} = \{a_1, a_4, a_6\}; & p = 1/2,
\end{array}
$$

and the output again matches $n = 6$.

The following analysis makes these observations precise.

2.2 Algorithm analysis

Heuristically having observed that, in expectation, Algorithm 1 performs reasonably well, one core tool for obtaining the multiplicative error bound (1) is Chernoff's inequality.[1]

Theorem 2.2.1 (Chernoff inequality)**.** *Let $X_1, \ldots, X_k$ be independent binary*

[1] We see how one obtains Chernoff's inequality and other so-called concentration inequalities later in the lecture.

random variables ($X_i \in \{0,1\}$), $S = \sum_{i=1}^{k} X_i$, and $\mu = \mathbb{E}S$. Then it holds that

$$\Pr(S \geq (1+\delta)\mu) \leq e^{-\delta^2\mu/(2+\delta)} \qquad \text{for any } \delta \geq 0, \tag{2}$$

$$\Pr(S \leq (1-\delta)\mu) \leq e^{-\delta^2\mu/2} \qquad \text{for any } \delta \in [0,1],$$

$$\Pr(|S-\mu| \geq \delta\mu) \leq 2e^{-\delta^2\mu/3} \qquad \text{for any } \delta \in [0,1]. \tag{3}$$

We now prove the following result, essentially following Chakraborty et al. [2022].

Theorem 2.2.2 (Theorem 2; Chakraborty et al. 2022)**.** *For any data stream $\mathcal{A}$ and any $0 < \epsilon, \delta < 1$, Algorithm 1 satisfies* (1)*.*

Proof. Algorithm 1 has two outcomes; we define the events

$$\begin{aligned} &\text{Fail} : \perp, \\ &\text{Error} : \text{The returned value is not in } [(1-\epsilon)n, (1+\epsilon)n], \end{aligned}$$

and show that[2]

$$\begin{aligned} \Pr(\text{Error}) &= \Pr(\text{Error} \cap \text{Fail}) + \Pr(\text{Error} \cap \text{Fail}^c) \\ &\leq \Pr(\text{Fail}) + \Pr(\text{Error} \cap \text{Fail}^c) \leq \frac{\delta}{8} + \frac{\delta}{2} \leq \delta. \end{aligned}$$

- $\Pr(\text{Fail}) \leq \frac{\delta}{8}$. In the worst case, $|\mathcal{X}| = \text{thresh}$ in every of the m iterations. For $\perp$ to occur, none of the elements of $\mathcal{X}$ are thrown away, which happens with probability $2^{-\text{thresh}}$. Hence,

$$\Pr(\text{Fail}) \leq m2^{-\text{thresh}} \leq m\left(\frac{\delta}{8m}\right)^{12/\epsilon^2} \leq \frac{\delta}{8}.$$

- $\Pr(\text{Error} \cap \text{Fail}^c) \leq \delta/2$. Consider an adapted Algorithm 1', which omits line 8 such that Fail can not happen. Indeed, if Fail does not occur, both algorithms are equivalent. From now on, we consider Algorithm 1'. Corresponding to the event Error of Algorithm 1, define the event

$$\text{Error}_2 = \text{The returned value is not in } [(1-\epsilon)n, (1+\epsilon)n]$$

  for Algorithm 1' and observe that $\Pr(\text{Error} \cap \text{Fail}^c) \leq \Pr(\text{Error}_2)$. The next step is to show that $\Pr(\text{Error}_2) \leq \delta/2$, which will imply the result.

  Denote by $\mathcal{X}_j$ and p_j the values of $\mathcal{X}$ and p at the end of the loop iteration $i = j$, and by $S_j = \{a_1, \ldots, a_j\}$ the distinct first j elements of $\mathcal{A}$. Notice that for each $a_i \in S_j$, $\Pr(a_i \in \mathcal{X}_j) = p_j$. For the analysis, it is convenient to make the value of the corresponding p_j-s explicit (keeping track of how

[2]Recall that for mutually disjoint events $B_1, \ldots, B_k$ and any event A, the law of total probability states that $\Pr(A) = \sum_{i=1}^{k} \Pr(A \cap B_i)$.

often $|\mathcal{X}| = \text{thresh}$ occurred), which we denote by $(p_j, \mathcal{X}_j) = (p_j, \mathcal{Y}_{k,j})$ where $p_j = 2^{-k}$. With this notation

$$\mathbb{E}|\mathcal{Y}_{k,j}| = 2^{-k}|S_j| \leq 2^{-k}n. \tag{4}$$

To gain control of $\Pr(\text{Error}_2)$, we define the event

$$\text{Bad}_2 : p < \text{thresh}/4n \text{ after line 7 of Algorithm 1'},$$

where we note that the event is "Bad" as the output would then overestimate n by 4. We now use the relaxation $\Pr(\text{Error}_2) \leq \Pr(\text{Bad}_2) + \Pr(\text{Error}_2 \cap \text{Bad}_2^c)$ and let $\ell = \lfloor \log_2 \frac{4n}{\text{thresh}} \rfloor$. Note that one can show that $2^{-(\ell+1)} < \text{thresh}/4n \leq 2^{-\ell}$; then, as $p = 2^{-k}$ for some integer k, $p < 2^{-\ell}$ if and only if $p < \text{thresh}/4n$.

- $\Pr(\text{Bad}_2) \leq \frac{\delta}{4}$. Denote by $\text{Bad}_{2,j}$ the event that "the j-th iteration of the **for** loop in Algorithm 1' is the first iteration where $p < 2^{-l}$" = "$p_{j-1} = 2^{-l}$ and $p_j = 2^{-(l+1)}$." To bound $\Pr(\text{Bad}_{2,j})$, note that in case the event happens $|\mathcal{X}_j| = |\mathcal{Y}_{\ell,j}| = \text{thresh}$, to obtain

  $$\Pr(\text{Bad}_{2,j}) \leq \Pr(|\mathcal{Y}_{\ell,j}| \geq \text{thresh}) \leq e^{-\frac{9\text{thresh}}{20}} \leq \frac{\delta}{4m},$$

  where we applied the Chernoff bound (2) after subtracting $\mathbb{E}|\mathcal{Y}_{\ell,j}|$ from both sides, using that $\mathbb{E}|\mathcal{Y}_{\ell,j}| \leq \text{thresh}/4$ (by (4) and the definition of ℓ) on the r.h.s., and setting $\delta = \frac{3\text{thresh}}{4\mathbb{E}|\mathcal{Y}_{\ell,j}|}$ in the inequality. Hence, $\Pr(\text{Bad}_2) \leq \sum_{j=1}^{m} \Pr(\text{Bad}_{2,j}) \leq \frac{\delta}{4}$.
- $\Pr(\text{Error}_2 \cap \text{Bad}_2{}^c) \leq \frac{\delta}{4}$. We define $\text{Error}_{2,q}$ as "$p_m = 2^{-q}$ and $\frac{|\mathcal{X}_m|}{2^{-q}} \notin [(1-\epsilon)n, (1+\epsilon)n]$." Then $\Pr(\text{Error}_{2,q}) = \Pr(|\mathcal{Y}_{q,m}| \notin [(1-\epsilon)\frac{n}{2^q}, (1+\epsilon)\frac{n}{2^q}])$. It follows that

  $$\begin{aligned}\Pr(\text{Error}_2 \cap \text{Bad}_2^c) &\leq \sum_{q=0}^{\ell} \Pr(\text{Error}_{2,q}) \leq \sum_{q=0}^{\ell} 2e^{-\frac{\epsilon^2 n}{3\cdot 2^q}} \\ &\leq 2(\ell+1)e^{-\frac{\epsilon^2 n}{3\cdot 2^\ell}} \leq \frac{\delta}{4},\end{aligned}$$

  where, for the first inequality, we used the definition of Bad_2. The second inequality follows by (3). For the penultimate term, we upper bound the sum by $\ell + 1$ times its maximum, noting that the terms are increasing in q. □

Exercise 2.2.1. *What is the memory requirement of Algorithm 1?*

Exercise 2.2.2. *Show that $p < 2^{-\ell}$ if and only if $p < \text{thresh}/4n$.*

Exercise 2.2.3. *Verify the applications of the Chernoff inequalities.*

2.3 Notes

This section illustrates a few of the probabilistic tools we employ repeatedly throughout these notes: union bounds, conditioning, the law of total probability, and concentration inequalities, to name but a few. We investigate these tools in more detail over the following few sections, including proving the Chernoff inequalities.

As already noted, the count-distinct problem is well-studied and we refer to Chakraborty et al. [2022, Section 3] for related work. For a slightly different approach to the analysis of the CVM algorithm, see Knuth [2023]

3 Probability review

In this section, we recall (measure-theoretic) probability theory. The goal is to obtain a working foundation quickly—this is not a lecture in measure-theory and many important results are stated without proof. Please refer to the notes section for references to proofs and for more details on this material.

3.1 Measure and probability spaces

We start by defining our notions of main interest, connecting them to probability immediately afterwards.

Definition 3.1.1 (Measurable set, measurable space, and sigma-algebra)**.** *Let Ω be any set. $\mathcal{S}$ is a sigma-algebra on Ω if*

(i) $\Omega \in \mathcal{S}$,

(ii) if $A \in \mathcal{S}$ *then* $A^c = \Omega \setminus A \in \mathcal{S}$ *(closed under complementation), and*

(iii) if $(A_i) = A_1, A_2, \ldots \in \mathcal{S}$ *then* $A_1 \cup A_2 \cup \cdots \in \mathcal{S}$ *(closed under countable unions).*

An $A \in \mathcal{S}$ is a measurable set. $(\Omega, \mathcal{S})$ is a measurable space.

In probability, Ω is the set of all possible outcomes and called the sample space. The sigma-algebra $\mathcal{S}$ is the class of events (i.e., the measurable sets) that we consider. Think of throwing a six-sided die, which has outcomes $\Omega = \{1, 2, 3, 4, 5, 6\}$. Assume that we are interested in the event that the die shows an even number, i.e., $A = \{2, 4, 6\}$. We could take $\mathcal{S} = \{\Omega, \{\}, A, A^c\}$, which is a sigma-algebra. An alternative is $\mathcal{S} = 2^\Omega$, which allows to measure all possible events.

In most cases, Ω is infinite and $\mathcal{S}$ can not be written down easily.

Definition 3.1.2 (Measure, measure space)**.** *A set function $\mu : \mathcal{S} \to [0, \infty]$ is a measure if*

(i) $\mu(\emptyset) = 0$,

(ii) $\mu(A) \geq 0$ *for any* $A \in \mathcal{S}$, *and*

(iii) $\mu\Big(\bigcup_{i=1}^\infty A_i\Big) = \sum_{i=1}^\infty \mu(A_i)$ *for any mutually disjoint countable collection* (A_i) $(A_i \in \mathcal{S})$.

The triple $(\Omega, \mathcal{S}, \mu)$ is called measure space.

In other words, a measure is a non-negative and countably additive set function acting on elements of a sigma algebra and vanishing on the empty set.

For technical reasons, we include the definition of sigma-finiteness.

Definition 3.1.3 (Finite measure, sigma-finite measure)**.** *Let $(\Omega, \mathcal{S})$ be a measurable space. A measure μ on $(\Omega, \mathcal{S})$ is finite if $\mu(\Omega) < \infty$. If there exists a sequence $(A_i) \in \mathcal{S}$ such that*

$$\Omega = \bigcup_{i=1}^{\infty} A_i \text{ and } \mu(A_i) < \infty \text{ for every } i \in \mathbb{N},$$

we call μ sigma-finite.

The following result is the fundamental property of measure; notice its close connection to the monotone convergence of the integral (introduced later below).

Theorem 3.1.1 (Monotone convergence)**.** *Let $(\Omega, \mathcal{S}, \mu)$ be a measure space and (A_i) a sequence in $\mathcal{S}$. Suppose that $A_i \subseteq A_{i+1}$ for all $i \in \mathbb{N}$ and that $\bigcup_{i=1}^{\infty} A_i = A$. Then $\mu(A_n) \uparrow \mu(A)$.*

Proof. Let $n \in \mathbb{N}$, $G_1 = A_1$, and $G_i = A_i \setminus A_{i-1}$ for $i \geq 2$. As the G_i-s are disjoint, we have

$$\mu(A_n) = \mu(G_1 \cup G_2 \cup \cdots G_n) = \sum_{i=1}^{n} \mu(G_i) \uparrow \sum_{i=1}^{\infty} \mu(G_i) = \mu(A)$$

as $n \to \infty$. □

We next define probability.

Definition 3.1.4 (Probability measure, probability space)**.** *Let $(\Omega, \mathcal{S}, \mu)$ be a measure space. If $\mu(\Omega) = 1$, we call μ a probability measure and $(\Omega, \mathcal{S}, \mu)$ probability space.*

By convention, one often denotes probability measures by P or Q; the probability spaces then are $(\Omega, \mathcal{S}, P)$ and $(\Omega, \mathcal{S}, Q)$, respectively. For concreteness, recall the die example with $\Omega = \{1, 2, 3, 4, 5, 6\}$, consider the probability space $(\Omega, 2^{\Omega}, P)$, and assume a fair die. The probability of an even throw is $P(\{2, 4, 6\}) = P(\{2\} \cup \{4\} \cup \{6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = 3/6 = 1/2$.

Let $(\Omega, \mathcal{S}, P)$ be a probability space. From the above definitions, a few properties of P follow.

Theorem 3.1.2 (Properties of probability measures)**.** *Let $(\Omega, \mathcal{S}, P)$ be a probability space.*

(i) For any $A \in \mathcal{S}$ it holds that $P(A^c) = 1 - P(A)$. In particular, $P(\emptyset) = 0$.

(ii) Let $A \in \mathcal{S}$ and $B_1, B_2, \ldots \in \mathcal{S}$ a partition of Ω. Then $P(A) = \sum_{i=1}^{\infty} P(A \cap B_i)$ (law of total probability).

(iii) Suppose $A \subseteq B$. Then $P(A) \leq P(B)$ (monotonicity).

(iv) Let $A_1, A_2, \ldots \in \mathcal{S}$. Then $P\big(\bigcup_{i=1}^{\infty} A_i\big) \leq \sum_{i=1}^{\infty} P(A_i)$ (union bound).

Proof. (i) and (ii) are immediate.

(iii) Observe that $B = B \setminus A \cup A$. Then $P(B) = P(B \setminus A) + P(A) \geq P(A)$.

(iv) Set $B_i = A_i \setminus \bigcup_{j=1}^{i-1} A_j$ and observe that $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$. Hence, $P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = P\Big(\bigcup_{i=1}^{\infty} B_i\Big) = \sum_{i=1}^{\infty} P(B_i) \leq \sum_{i=1}^{\infty} P(A_i)$ as $B_i \subseteq A_i$ by construction and (iii). □

For a topological space $(\mathcal{X}, \tau_{\mathcal{X}})$, we denote by $\mathcal{B}(\mathcal{X})$ its Borel sigma-algebra, that is, the smallest sigma-algebra generated by all open sets. If $\mathcal{X} = \mathbb{R}^d$ and $d > 1$, we write $\mathcal{B}^d$ for the Borel sigma-algebra. If $d = 1$, we write $\mathcal{B}$. If $P : \mathcal{B}(\mathcal{X}) \to [0, 1]$, we call P Borel probability measure.

Definition 3.1.5 (Almost everywhere, almost surely)**.** *Let $(\Omega, \mathcal{S}, \mu)$ be a measure space. If a property p holds for all $\omega \in \Omega$ except for some that form at most a set of μ-measure zero, we say that p holds μ-almost everywhere (μ-a.e.). If μ is a probability measure, we also write that p holds μ-almost surely (μ-a.s.).*

We omit the measure if it is clear from the context, writing just a.e. or a.s.

3.2 Measurable functions and random variables

In this section, we define measurable functions and observe that in the language of measure-theory, random variables are nothing but real-valued measurable functions. We also review the properties of measurable functions that will allow us to later define the Lebesgue integral and therefore expectation.

Definition 3.2.1 (Measurable function)**.** *Let $(\Omega_1, \mathcal{S}_1)$ and $(\Omega_2, \mathcal{S}_2)$ be measurable spaces. A function $f : \Omega_1 \to \Omega_2$ is measurable (more precisely: $\mathcal{S}_1/\mathcal{S}_2$-measurable) if $f^{-1}(A) = \{\omega \in \Omega_1 : f(\omega) \in A\} \in \mathcal{S}_1$ for all $A \in \mathcal{S}_2$.*

Put differently, a function f is measurable if f^{-1} maps measurable sets to measurable sets.

Definition 3.2.2 (Generated sigma-algebra)**.** *Let Ω_1 be a set, $(\Omega_2, \mathcal{S}_2)$ a measure space, and $f : \Omega_1 \to \Omega_2$. Then $\{\{\omega : f(\omega) \in B\} : B \in \mathcal{S}_2\} = \{\{f \in B\} : B \in \mathcal{S}_2\}$ is a sigma-algebra, written as $\sigma(f)$, and the smallest sigma-algebra w.r.t. which f is measurable.*

If $(\Omega_1, \tau_{\Omega_1})$ and $(\Omega_2, \tau_{\Omega_2})$ are topological spaces and the function $f : \Omega_1 \to \Omega_2$ is $\mathcal{B}(\Omega_1)/\mathcal{B}(\Omega_2)$-measurable, we say that f is Borel measurable.

Definition 3.2.3 (Random variable, random vector)**.** *Let $(\Omega, \mathcal{S})$ be a measurable space. If $X : \Omega \to \mathbb{R}$ is $\mathcal{S}/\mathcal{B}$-measurable, we call X random variable. If $X : \Omega \to \mathbb{R}^d$ is $\mathcal{S}/\mathcal{B}^d$-measurable, we call X random vector.*

An important property of measurable functions is that they can be composed.

Theorem 3.2.1 (Composition of measurable functions)**.** *Let $(\Omega_1, \mathcal{S}_1)$, $(\Omega_2, \mathcal{S}_2)$, and $(\Omega_3, \mathcal{S}_3)$ be measurable spaces and suppose $f : \Omega_1 \to \Omega_2$ and $g : \Omega_2 \to \Omega_3$ are measurable. Then $g \circ f : \Omega_1 \to \Omega_3$ is $\mathcal{S}_1/\mathcal{S}_3$-measurable.*

Proof. Observe that

$$\Omega_1 \xrightarrow{f} \Omega_2 \xrightarrow{g} \Omega_3 \text{ and}$$
$$\mathcal{S}_1 \xleftarrow{f^{-1}} \mathcal{S}_2 \xleftarrow{g^{-1}} \mathcal{S}_3.$$ □

Further, random variables form an algebra. Suprema, infima, and limits of sequences of random variables also preserve measurability.[3]

Theorem 3.2.2 (More facts on random variables)**.** *Let $X, Y : \Omega \to \mathbb{R}$ be random variables and $c \in \mathbb{R}$. Then $f + g$, fg, and cf are measurable. Moreover, if (X_n) are random variables, then so are $\inf_n X_n$, $\sup_n X_n$, $\liminf_n X_n$, and $\limsup_n X_n$.*

So far, we did not consider the interplay of random variables and probability measures defined on the same space. Thus, before we end this section, we note that for a probability space $(\Omega, \mathcal{S}, P)$ carrying a random variable X one has

$$\Omega \xrightarrow{X} \mathbb{R},$$
$$[0,1] \xleftarrow{P} \mathcal{S} \xleftarrow{X^{-1}} \mathcal{B},$$
$$\text{or, more precisely, } [0,1] \xleftarrow{P} \sigma(X) \xleftarrow{X^{-1}} \mathcal{B}.$$

This observation underpins the following definitions.

Definition 3.2.4 (Law, distribution function)**.** *Let $(\Omega, \mathcal{S}, P)$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable. The law of X is $\mathcal{L}_X = P \circ X^{-1}$, $\mathcal{L}_X : \mathcal{B} \to [0,1]$ and it is a probability measure on $(\mathbb{R}, \mathcal{B})$. The distribution function $F_X : \mathbb{R} \to [0,1]$ of X is*

$$F_X(c) = \mathcal{L}_X(-\infty, c] = P(X \leq c) = P(\{\omega : X(\omega) \leq c\}).$$

Remark 3.2.1. *Note that the law of a random variable equals its so-called distribution. Do not mix this with its distribution* function*.*

3.3 Independence

Independence is arguably where general measure theory and probability theory really diverge. In general, one may define independence in terms of sigma-algebras as follows.

Definition 3.3.1 (Independent sub-sigma-algebras)**.** *Let $(\Omega, \mathcal{S}, P)$ be a probability space. Sub-sigma-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots$ of $\mathcal{S}$ are called independent if, whenever $G_i \in \mathcal{G}_i$ $(i \in \mathbb{N})$ and $i_1, \ldots, i_n$ are distinct,*

$$P(G_{i_1} \cap \cdots \cap G_{i_n}) = \prod_{k=1}^{n} P(G_{i_k}). \tag{5}$$

[3]Here one considers the extended real line, i.e., including $\pm\infty$ and the corresponding Borel sigma-algebra. We do not make this change explicit.

The definitions from elementary probability in terms of random variables and events can then be defined based on (5).

Definition 3.3.2 (Independent random variables)**.** *Let $(\Omega, \mathcal{S}, P)$ be a probability space carrying the random variables $X_1, X_2, \ldots$. Then $X_1, X_2, \ldots$ are independent if the sigma-algebras $\sigma(X_1), \sigma(X_2), \ldots$ are independent.*

Definition 3.3.3 (Independent events)**.** *Let $(\Omega, \mathcal{S}, P)$ be a probability space and $E_1, E_2, \ldots \in \mathcal{S}$. The events $E_1, E_2, \ldots$ are independent if the sigma-algebras $\sigma(\mathbf{1}_{E_1}), \sigma(\mathbf{1}_{E_2}), \ldots$ are independent.*

For our use case, the main thing to keep in mind is that *independence means multiply.*

What formally plays an important role (though only in the background) is that one can construct sequences of independent random variables with prescribed distribution functions. In other words, the notion of independent and identically distributed (i.i.d.) random variables is well-defined. We do not explore this topic here.

3.4 The Lebesgue integral

In this section, we construct the Lebesgue integral. Its construction proceeds in four steps:

(i) We define the integral for *indicator functions*,

(ii) then for *simple functions*,

(iii) followed by *non-negative functions*,

(iv) and finally for *integrable functions.*

The details are as follows, with $(\Omega, \mathcal{S}, \mu)$ a measure space.

(i) **Indicator functions.** Let

$$\text{Sim} \int \mathbf{1}_{\{A\}} \mathrm{d}\mu = \text{Sim} \int \mathbf{1}_{\{A\}}(\omega) \mathrm{d}\mu(\omega) = \mu(A),$$

with $\text{Sim} \int$ indicating that our current integral is only defined for indicator functions.

(ii) **Simple functions.** A simple function f is a function that can be written as

$$f(\omega) = \sum_{i=1}^{n} a_i \mathbf{1}_{\{A_i\}}(\omega) \text{ for } \omega \in \Omega,$$

with $a_i \geq 0$, disjoint $A_i \in \mathcal{S}$ and $\mu(A_i) < \infty$. We note that a simple function is measurable. We now define

$$\text{Sim} \int f \mathrm{d}\mu = \sum_{i=1}^{n} a_i \mu(A_i). \tag{6}$$

For (6), one can check various properties, such as linearity, monotonicity, independence of the A_i-choices, and that for any two simple functions f and g satisfying $\mu(\{x \in S : f(x) \neq g(x)\}) = 0$, it holds that $\mathrm{Sim} \int f \mathrm{d}\mu = \mathrm{Sim} \int g \mathrm{d}\mu$. In other words, if two simple functions agree almost everywhere, their integrals are equal. We do not pursue the corresponding proofs here.

Now, for a fixed $A \in \mathcal{S}$ and a simple function f, we define

$$\mathrm{Sim} \int_A f \mathrm{d}P = \mathrm{Sim} \int \mathbf{1}_A f \mathrm{d}P,$$

and note that $\mathbf{1}_A f$ is again simple.

(iii) **Non-negative functions.** For a non-negative measurable function f and $A \in \mathcal{S}$, we set

$$\int_A f \mathrm{d}\mu = \sup \left\{ \mathrm{Sim} \int_A h \mathrm{d}\mu : h \text{ simple}, \ h \leq f \right\} \leq \infty,$$

which is sensible given that every non-negative measurable function can be approximated by an increasing sequence of simple functions.

Theorem 3.4.1 (Approximation by simple functions)**.** *For any non-negative measurable function f, there exists a sequence of simple functions (f_n) such that $f_n \uparrow f$.*

Proof. Define the staircase function $\alpha_n : [0, \infty] \to [0, \infty]$ as $(n \in \mathbb{N})$

$$\alpha_n = \begin{cases} 0 & \text{if } x = 0, \\ (i-1)2^{-n} & \text{if } (i-1)2^{-n} < x \leq i2^{-n} \text{ for } i \in \mathbb{N}, \\ n & \text{if } x > n. \end{cases}$$

Then $f_n = \alpha_n \circ f$ is simple and $f_n \uparrow f$. □

(iv) **Measurable functions.** For measurable f, write $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$. Let $A \in \mathcal{S}$. If any of $\int_A f^+ \mathrm{d}\mu$ or $\int_A f^- \mathrm{d}\mu$ is finite, we define

$$\int_A f \mathrm{d}\mu = \int_A f^+ \mathrm{d}\mu - \int_A f^- \mathrm{d}\mu,$$

which can be $\pm\infty$. If both integrals, $\int_A f^+ \mathrm{d}\mu$ and $\int_A f^- \mathrm{d}\mu$, are finite, we say f is integrable on A. In other words, the integral can be defined but a function is called "integrable" iff the integral is finite. For f integrable on A, we write $f \in \mathcal{L}^1(\Omega, \mathcal{S}, \mu)$ on A; if $A = \Omega$, we omit "on A" and also write $f \in \mathcal{L}^1$ and $\int f \mathrm{d}\mu$. $\mathcal{L}^1$ is the space of Lebesgue integrable functions.

The Lebesgue integral has all the properties that we expect, collected in the following result.

Theorem 3.4.2 (Properties of Lebesgue integral)**.** *Let $f, g \in \mathcal{L}^1$ and $\lambda \in \mathbb{R}$. It holds that*

(i) $\int f + g \mathrm{d}\mu = \int f \mathrm{d}\mu + \int g \mathrm{d}\mu$ *and* $\int \lambda f \mathrm{d}\mu = \lambda \int f \mathrm{d}\mu$ *(linearity),*

(ii) if $f \leq g$*, then* $\int f \mathrm{d}\mu \leq \int g \mathrm{d}\mu$ *(monotonicity),*

(iii) if f *is non-negative and* $\int f \mathrm{d}\mu = 0$*, then* $f = 0$ μ*-a.e., and*

(iv) $\left| \int f \mathrm{d}\mu \right| \leq \int |f| \mathrm{d}\mu$.

We state here a few key results of integration theory, without proofs.

Theorem 3.4.3 (Monotone convergence)**.** *Let* $(\Omega, \mathcal{S}, \mu)$ *be a measure space and* $A \in \mathcal{S}$*. Suppose* (f_n) *is non-negative, measurable, and* $f_n \uparrow f$ *on* A*. Then* $\lim_{n\to\infty} \int_A f_n \mathrm{d}\mu = \int_A f \mathrm{d}\mu$.

Theorem 3.4.4 (Fatou's lemma)**.** *Let* $(\Omega, \mathcal{S}, \mu)$ *be a measure space and* (f_n) *a sequence of measurable functions. Then*

$$\int \liminf_{n\to\infty} f_n \, \mathrm{d}\mu \leq \liminf_{n\to\infty} \int f_n \, \mathrm{d}\mu.$$

Theorem 3.4.5 (Dominated convergence)**.** *Let* $(\Omega, \mathcal{S}, \mu)$ *be a measure space,* f*,* (f_n) *measurable,* $\lim_{n\to\infty} f_n = f$*, and suppose* $|f_n(\omega)| \leq g(\omega)$ *for some non-negative* $g \in \mathcal{L}^1$ *for all* $\omega \in \Omega$*,* $n \in \mathbb{N}$*. Then* $\lim_{n\to\infty} \int |f_n - f| \mathrm{d}\mu = 0$*. In particular,* $\lim_{n\to\infty} \int f_n \, \mathrm{d}\mu = \int f \mathrm{d}\mu$.

For the "in particular"-part, observe that

$$\left| \int f_n \, \mathrm{d}\mu - \int f \mathrm{d}\mu \right| = \left| \int f_n - f \mathrm{d}\mu \right| \leq \int |f_n - f| \mathrm{d}\mu.$$

Having defined integration, one may construct new measures from a given measure. Indeed, let $f \in \mathcal{L}^1$, $f \geq 0$ μ-a.e., and consider

$$\nu(A) = \int_A f \mathrm{d}\mu \text{ for } A \in \mathcal{S}.$$

Then ν is a measure on $(\Omega, \mathcal{S})$. We observe that, for any $A \in \mathcal{S}$ for which $\mu(A) = 0$, we have that $\nu(A) = 0$. The Radon-Nikodym theorem gives a converse.

Definition 3.4.1 (Absolute continuity)**.** *Let* μ, ν *be measures on the measurable space* $(\Omega, \mathcal{S})$*. If* $\nu(A) = 0$ *for any* $A \in \mathcal{S}$ *for which* $\mu(A) = 0$*, we say that* ν *is absolutely continuous w.r.t.* μ *and write* $\nu \ll \mu$.

Theorem 3.4.6 (Radon-Nikodym)**.** *Let* $(\Omega, \mathcal{S})$ *be a measure space,* μ, ν *sigma-finite measures on* $(\Omega, \mathcal{S})$*, and suppose that* $\nu \ll \mu$*. Then, there exists a non-negative measurable* f *such that*

$$\nu(A) = \int_A f \mathrm{d}\mu \text{ for any } A \in \mathcal{S}.$$

If h *is another such function, then* $f = h$ μ*-a.e.*

The function f is called the Radon-Nikodym derivative of ν w.r.t. μ and written as $f = \dfrac{\mathrm{d}\nu}{\mathrm{d}\mu}$.

3.5 Expectation

We are ready to define the expectation, the variance, and the covariance of random variables.

Definition 3.5.1 (Expectation)**.** *Let $(\Omega, \mathcal{S}, P)$ be a probability space. The expectation of $X \in \mathcal{L}^1(\Omega, \mathcal{S}, P)$ is*

$$\mathbb{E}X = \int X \mathrm{d}P = \int_\Omega X(\omega) \mathrm{d}P(\omega). \tag{7}$$

Let $X, Y \in \mathcal{L}^1$. Extending the above list, key properties of the expectation are

(iv) if X, Y are independent, then $\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y$ (independence).

Note that expectation takes precedence over sums but not over products, which should cause no confusion. Put differently, $\mathbb{E}X + Y = \mathbb{E}[X] + Y$ and $\mathbb{E}XY = \mathbb{E}[XY]$.

One frequently does not know how to compute (7) due to integrating over Ω. The following result addresses this issue and yields the well known expression for the expectation; it is frequently used with $F = \mathbb{R}^d$.

Theorem 3.5.1 (Change of variables formula)**.** *Let $X : (\Omega, \mathcal{S}, P) \to (F, \mathcal{F})$ be a random variable with law $\mathcal{L}_X(A) = P(X \in A)$. If $f : (F, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ is measurable and $f \geq 0$ or $f \in \mathcal{L}^1(\Omega, \mathcal{S}, P)$, then*

$$\mathbb{E}f(X) = \int_F f(y) \mathrm{d}\mathcal{L}_X(y).$$

Remark 3.5.1. *Regarding the name of the result, note that*

$$\int_\Omega f(X(\omega)) \mathrm{d}P(\omega) = \int_F f(y) \mathrm{d}(P \circ X^{-1})(y).$$

We remark that the expectation may very well be undefined. Indeed, let the random variable X have law $P(X = 2^i) = P(X = -2^i) = 2^{-i}$ for $i \geq 2$. Then $\int X^+ \mathrm{d}P = \int X^- \mathrm{d}P = \infty$.

In elementary probability, we say that a random variable X has *density* f_X if

$$F_X(c) = \int_{-\infty}^c f_X(x) \mathrm{d}x \text{ for all } c \in \mathbb{R}$$

and its expectation then takes the form $\mathbb{E}X = \int_{-\infty}^\infty x f_X(x) \mathrm{d}x$. Note that letting $f_X = \dfrac{\mathrm{d}\mathcal{L}_X}{\mathrm{d}\lambda}$, with λ the Lebesgue measure on $\mathbb{R}$, gives $\mathbb{E}X = \int x \mathrm{d}\mathcal{L}_X(x)$. Also, by the definition of F_X and Radon-Nikodym

$$F_X(c) = \mathcal{L}_X(-\infty, c] = \int_{(-\infty, c]} \mathrm{d}\mathcal{L}_X = \int_{(-\infty, c]} \frac{\mathrm{d}\mathcal{L}_X}{\mathrm{d}\lambda} \mathrm{d}\lambda = \int_{(-\infty, c]} f_X \mathrm{d}\lambda,$$

which ties everything together.

Definition 3.5.2 (Variance, covariance)**.** *Let $(\Omega, \mathcal{S}, P)$ be a probability space and $X, Y \in \mathcal{L}^1$. The variance of X is*

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

The covariance of X and Y is given by

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

We note that the covariance is bilinear and that $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$. Next, we give a short introduction to conditional expectation.

3.6 Conditional expectation

Conditional expectations can seem intimidating if presented in the abstract setting. We start with an intuitive account, viewing conditional expectations as projections, before presenting the abstract definition. As we explore later, the projection-based perspective is closely related to regression.

Let $(\Omega, \mathcal{S}, P)$ be a probability space with random variables $T, \{M : M \in \mathcal{M}\} \in \mathcal{L}^2 = \mathcal{L}^2(\Omega, \mathcal{S}, P) = \{X \text{ random variable} : \int X^2 \mathrm{d}P < \infty\}$, where $\mathcal{M} \subset \mathcal{L}^2$. We call $\mathcal{L}^2$ the space of square integrable random variables.

A random variable $\widehat{M}$ is a projection of T onto $\mathcal{M}$ if (i) $\widehat{M} \in \mathcal{M}$ and (ii)

$$M \mapsto \mathbb{E}(T - M)^2, \quad M \in \mathcal{M}$$

is minimized.

Assume that $\mathcal{M}$ is linear. Then the following result shows that $\widehat{M}$ is a projection of T iff $T - \widehat{M}$ is orthogonal to $\mathcal{M}$ w.r.t. the inner product $\langle X, Y \rangle = \mathbb{E}XY$ $(X, Y \in \mathcal{L}^2)$.

Theorem 3.6.1 (Projection is orthogonal)**.** *Assume the setup from the previous paragraphs. In particular, let $\mathcal{M} \subset \mathcal{L}^2$ be linear. Then $\widehat{M}$ is the projection of T onto $\mathcal{M}$ iff $\widehat{M} \in \mathcal{M}$ and*

$$\mathbb{E}(T - \widehat{M})M = 0,$$

for every $M \in \mathcal{M}$. Every two projections of T onto $\mathcal{M}$ agree a.s. If $\mathcal{M}$ contains constant functions, then $\mathbb{E}T = \mathbb{E}\widehat{M}$ and $\mathrm{Cov}(T - \widehat{M}, M) = 0$ for every $M \in \mathcal{M}$.

Proof. $\Longleftarrow$: For any $M, \widehat{M} \in \mathcal{M}$,

$$\mathbb{E}(T - M)^2 = \mathbb{E}(T - \widehat{M})^2 + 2\mathbb{E}(T - \widehat{M})(\widehat{M} - M) + \mathbb{E}(\widehat{M} - M)^2.$$

If $\widehat{M}$ satisfies the orthogonality condition, then the middle term is zero, and we conclude that $\mathbb{E}(T-M)^2 \geq \mathbb{E}(T-\widehat{M})^2$, with strict inequality unless $\mathbb{E}(\widehat{M}-M)^2 = 0$. Thus, the orthogonality condition implies that $\widehat{M}$ is a projection, and also that it is unique a.s.

$\Longrightarrow$: For any $\alpha \in \mathbb{R}$,

$$\mathbb{E}(T - \widehat{M} - \alpha M)^2 - \mathbb{E}(T - \widehat{M})^2 = -2\alpha\mathbb{E}(T - \widehat{M})M + \alpha^2\mathbb{E}M^2.$$

If $\widehat{M}$ is a projection, then this expression is nonnegative for every α. But the parabola $\alpha \mapsto \alpha^2 \mathbb{E}M^2 - 2\alpha \mathbb{E}(T - \widehat{M})M$ is nonnegative iff the orthogonality condition $\mathbb{E}(T - \widehat{M})M = 0$ is satisfied.

If the constants are in $\mathcal{M}$, then the orthogonality condition implies $\mathbb{E}(T - \widehat{M})c = 0$, which yields the last assertions of the theorem. □

Hence, $\mathbb{E}X$ minimizes the quadratic form $a \mapsto \mathbb{E}(X - a)^2$ over the real numbers a; it is the best predictor of X with a quadratic loss function when given no additional information.

The *conditional expectation* $\mathbb{E}(X \mid Y)$ of X given Y is defined as the best "prediction" of X given knowledge of Y. Formally, $\mathbb{E}(X \mid Y)$ is a measurable function $g_0(Y)$ of Y that minimizes

$$\mathbb{E}(X - g(Y))^2$$

over all measurable functions g. Accordingly, $\mathbb{E}(X \mid Y)$ is the projection of X onto the linear space of all measurable functions of Y. In light of the preceding theorem, $\mathbb{E}(X \mid Y)$ is the unique measurable function satisfying

$$\langle X - \mathbb{E}(X \mid Y), g(Y) \rangle = \mathbb{E}(X - \mathbb{E}(X \mid Y))g(Y) = 0,$$

for every g.

If $\mathbb{E}(X \mid Y) = g_0(Y)$, we write $\mathbb{E}(X \mid Y = y)$ for $g_0(y)$, the expected value of X given that $Y = y$ is observed.

Remark 3.6.1. *We collect a few properties.*

(i) Orthogonality with $g \equiv 1$ yields $\mathbb{E}X = \mathbb{E}\mathbb{E}(X \mid Y)$, which is called the "tower property" or "law of total expectation".

(ii) If $X = f(Y)$ for a measurable function f, then $\mathbb{E}(X \mid Y) = X$. Given knowledge of Y, X is perfectly predictable.

(iii) If X and Y are independent, then $\mathbb{E}(X \mid Y) = \mathbb{E}X$.

(iv) If f is measurable, then $\mathbb{E}(f(Y)X \mid Y) = f(Y)\mathbb{E}(X \mid Y)$ for any X and Y. In other words, given Y, $f(Y)$ behaves like a constant.

(v) If X and Y are independent, then $\mathbb{E}(f(X, Y) \mid Y = y) = \mathbb{E}f(X, y)$ for every measurable f.

Exercise 3.6.1. *Prove the above properties of conditional expectation.*

Note that we assumed square integrability of X in the preceding definition. In fact, $X \in \mathcal{L}^1(\Omega, \mathcal{S}, P)$ suffices.

Definition 3.6.1 (Conditional expectation)**.** *Let $(\Omega, \mathcal{S}_0, P)$ be a probability space, $\mathcal{S} \subset S_0$ a sigma-algebra, and $X \in \mathcal{L}^1(\Omega, \mathcal{S}_0, P)$. The conditional expectation of X given $\mathcal{S}$, denoted by $\mathbb{E}(X \mid \mathcal{S})$, is a random variable Y satisfying*

(i) Y is $\mathcal{S}$ measurable, and

(ii) for all $A \in \mathcal{S}$, $\int_A X \mathrm{d}P = \int_A Y \mathrm{d}P$.

Any such Y *is called a version of* $\mathbb{E}(X \mid \mathcal{S})$.

We note that any two versions of the conditional expectation agree a.s. If $\mathcal{S} = \sigma(Z)$ for some random variable Z, we abbreviate this as $\mathbb{E}(X \mid Z)$.

It can be shown that the definition of conditional expectation in terms of sigma-algebras matches the projection-based one if square integrability is assumed; see Durrett [2019, Theorem 4.1.15] for a proof.

3.7 Classic inequalities

This section is dedicated to three "classical" inequalities on measure spaces: Cauchy-Schwarz', Hölder's, and, on probability spaces, Jensen's.[4]

We start with the CBS inequality for real inner product spaces, which you already know from linear algebra. The inequality also holds for complex inner product spaces but we do not need this result in the lecture.

Theorem 3.7.1 (Cauchy-Schwarz inequality)**.** *Let* $(V, \langle \cdot, \cdot \rangle)$ *be a real inner product space. Then, for any* $u, v \in V$*, it holds that* $|\langle u, v \rangle|^2 \leq \|u\|^2 \|v\|^2$*, with equality iff* u *and* v *are linearly dependent.*

Proof. Let $t \in \mathbb{R}$ and consider

$$0 \leq \|tu + v\|^2 = t^2 \|u\|^2 + 2t|\langle u, v \rangle| + \|v\|^2, \tag{8}$$

which is a quadratic function in t, say, $p(t) = t^2 \|u\|^2 + 2t\langle u, v \rangle + \|v\|^2$. As (8) is non-negative, p can not have multiple real roots, which implies that

$$4|\langle u, v \rangle|^2 - 4\|u\|^2 \|v\|^2 \leq 0,$$

yielding the first part of the statement. For the second part, notice that unless u and v are linearly dependent, the inequality in (8) is strict. As then $|\langle u, v \rangle|^2 < \|u\|^2 \|v\|^2$, we must have $\|tu + v\|^2 = 0$ if $|\langle u, v \rangle|^2 = \|u\|^2 \|v\|^2$ by contraposition. □

The next result, Jensen's inequality, applies to convex functions.

Definition 3.7.1 (Convex function)**.** *Let* $\mathcal{X}$ *be a convex subset of a real vector space* $(V, \langle \cdot, \cdot \rangle)$ *and* $\varphi : \mathcal{X} \to \mathbb{R}$*.* φ *is convex iff*

$$\varphi(tx + (1-t)y) \leq t\varphi(x) + (1-t)\varphi(y) \quad \text{for all } t \in [0, 1] \text{ and all } x, y \in \mathcal{X}.$$

Theorem 3.7.2 (Jensen's inequality)**.** *Let* $(\Omega, \mathcal{S}, P)$ *be a probability space,* $X \in \mathcal{L}^1$*,* $\varphi : \mathbb{R} \to \mathbb{R}$ *convex, and* $\varphi(X) \in \mathcal{L}^1$*. Then* $\varphi(\mathbb{E}X) \leq \mathbb{E}\varphi(X)$*.*

The final inequality of this section, Hölder's inequality, is stated in terms of L^p norms.

[4] In fact, the Cauchy-Schwarz inequality—also called Cauchy-Bunyakovsky-Schwarz inequality (CBS)—holds on any inner product space.

Definition 3.7.2 (L^p norm)**.** *Let $(\Omega, \mathcal{S}, \mu)$ be a measure space, $f : \Omega \to \mathbb{R}$ measurable, and $p \in [1, \infty)$. Denote by $\|f\|_p = (\int |f|^p \mathrm{d}\mu)^{1/p}$.*

Notice that for a measurable function f, the statements (i) $\|f\|_1 < \infty$, (ii) f is integrable, and (iii) $f \in \mathcal{L}^1$ are equivalent. Also note that we have not shown that $\|\cdot\|_p$ is actually a norm.[5] We do not prove but use this fact.

Theorem 3.7.3 (Hölder's inequality)**.** *Let $(\Omega, \mathcal{S}, \mu)$ be a measure space and $f, g : \Omega \to \mathbb{R}$ measurable. If $p, q \in (1, \infty)$ with $1/p + 1/q = 1$, then $\int |fg| \mathrm{d}\mu \leq \|f\|_p \|g\|_q$.*

We remark that Hölder's inequality implies CBS on $(\Omega, \mathcal{S}, \mu)$ equipped with the inner product $\langle f, g \rangle = \int fg \, \mathrm{d}\mu$ by using $p = q = 2$.

3.8 Notes

The material in this section is standard and can be found in any measure-theoretic probability textbook. We mostly follow the presentation in the wonderful and highly recommended Williams [1991]. A nice expository proof of the Radon-Nikodym Theorem is Shapiro [2018]. Theorem 3.5.1 is taken from Durrett [2019, Theorem 1.6.9]. van der Vaart [1998] gives the projection-based introduction of conditional expectations (in particular Theorem 3.6.1) and also includes proofs for the related examples stated herein. The abstract definition of conditional expectation can be found in Durrett [2019], which also has proofs for CBS and the inequalities of Jensen and Hölder. An interesting non-standard presentation of all the above topics (and many more) is Pollard [2002]. A reference for more advanced theory is Ledoux and Talagrand [1991].

[5] In fact, $\|\cdot\|_p$ is only a norm if one considers equivalence classes of functions that differ on a set of measure zero.

4 Concentration inequalities

Suppose that X is an integrable random variable with mean $\mu = \mathbb{E}X < \infty$. Two-sided concentration inequalities are frequently of the form

$$P(|X - \mu| \geq t) \leq \text{something small}, \tag{9}$$

where $t > 0$. In other words, they assert that the probability that a random quantity strays far from its expectation is small. In our applications, X and/or t often depend on the sample size and the inequalities permit making quantitative claims about the likely performance of an algorithm depending on the number of observed samples.

To see why we can expect results of the form (9) to hold true, we recall the strong law of large numbers (SLLN).

Theorem 4.0.1 (SLLN)**.** *Let $X_1, X_2, \ldots$ be i.i.d. integrable random variables with $\mathbb{E}X_i = \mu$ and $S_N = \sum_{i=1}^{N} X_i$. Then $S_N/N \to \mu$ a.s. as $N \to \infty$.*

As a.s. convergence implies convergence in probability, Theorem 4.0.1 implies that for any $\epsilon, \delta > 0$ there exists $N_0 \in \mathbb{N}$ such that

$$P(|S_N/N - \mu| > \epsilon) < \delta$$

for any $N \geq N_0$.

The next sections introduce the 'classical' concentration inequalities, which differ in the assumptions they require and in the bounds they provide. We give a short comparison at the end of this section.

4.1 Markov's inequality and derivatives

One of the classical tools for bounding a probability in terms of an expectation is Markov's inequality.

Theorem 4.1.1 (Markov's inequality)**.** *Let X be a non-negative random variable. For any $t > 0$ it holds that*

$$tP(X \geq t) \leq \mathbb{E}X.$$

Proof. Fix $t > 0$. Note that one may write any $x \in \mathbb{R}$ as $x = x\mathbf{1}_{\{x \geq t\}} + x\mathbf{1}_{\{x < t\}}$. Substituting $X \geq 0$ for x and taking expectations yields that

$$\mathbb{E}X = \underbrace{\mathbb{E}X\mathbf{1}_{\{X \geq t\}}}_{\geq t\mathbb{E}\mathbf{1}_{\{X \geq t\}}} + \underbrace{\mathbb{E}X\mathbf{1}_{\{X < t\}}}_{\geq 0} \geq tP(X \geq t),$$

which is the stated result. □

Markov's inequality is often applied in the form stated next.

Corollary 4.1.1. *Let Z be a random variable and $g : \mathbb{R} \to \mathbb{R}_+$ measurable and non-decreasing. Then, for any $t > 0$, it holds that*

$$g(t)P(Z \geq t) \leq \mathbb{E}g(Z).$$

From this corollary, one also obtains the following well-known result.

Theorem 4.1.2 (Chebyshev's inequality)**.** *Let X be an integrable random variable with mean μ and variance σ^2. For any $t > 0$ it holds that*

$$t^2 P(|X - \mu| \geq t) \leq \sigma^2.$$

Exercise 4.1.1. *Prove Chebyshev's inequality.*

Recall that, for $p > 0$, the p-th moment (resp. absolute moment) of a random variable X is defined as $\mathbb{E}X^p$ (resp. $\mathbb{E}|X|^p$). The moment generating function (MGF) of X, if it exists, is $M_X(t) = \mathbb{E}e^{tX}$ for $t \in \mathbb{R}$. Comparing Markov's and Chebyshev's respective inequalities suggests that the finiteness of higher moments yields faster concentration. Indeed, later results exploit this observation.

Remark 4.1.1. *By the series expansion of the exponential function one has that*

$$M_X(t) = \mathbb{E}e^{tX} = \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{t^i X^i}{i!}\right] = \sum_{i=0}^{\infty} \frac{t^i \mathbb{E}X^i}{i!},$$

justifying the name MGF.

4.2 Sums of independent random variables

The proof of the next result, Chernoff's inequality, illustrates a powerful technique in deriving concentration inequalities by using that the MGF of a sum of independent random variables equals the product of their MGFs.

Theorem 4.2.1 (Chernoff's inequality)**.** *Let $X_1, \ldots, X_N$ be independent random variables having Bernoulli distributions with parameters p_i, $S_N = \sum_{i=1}^N X_i$, and $\mu = \mathbb{E}S_N$. For any $t > \mu$ it holds that*

$$P(S_N \geq t) \leq e^{-\mu}\Big(\frac{e\mu}{t}\Big)^t.$$

Proof. We start by introducing a new parameter $\lambda > 0$ (optimized later), multiply the inequality $S_N \geq t$ by λ, and exponentiate to obtain

$$P(S_N \geq t) = P\Big(e^{\lambda S_N} \geq e^{\lambda t}\Big) \leq e^{-\lambda t}\mathbb{E}e^{\lambda S_N} = e^{-\lambda t}\prod_{i=1}^{N} \mathbb{E}e^{\lambda X_i}, \tag{10}$$

where the inequality follows by an application of Markov's inequality (Theorem 4.1.1) and the last equality by the independence of the X_i-s. As the X_i-s are Bernoulli with parameter p_i, we have (using that $1 + x \leq e^x$)

$$\mathbb{E}e^{\lambda X_i} = p_i e^{\lambda} + (1 - p_i) = 1 + \big(e^{\lambda} - 1\big)p_i \leq \exp\Big[\big(e^{\lambda} - 1\big)p_i\Big].$$

Hence, $\prod_{i=1}^N \mathbb{E}e^{\lambda X_i} \leq \prod_{i=1}^N \exp\Big[\big(e^\lambda - 1\big)p_i\Big] = \exp\Big[\big(e^\lambda - 1\big)\mu\Big]$, which we use in (10) to obtain

$$P(S_N \geq t) \leq e^{-\lambda t} \exp\Big[\big(e^\lambda - 1\big)\mu\Big].$$

Minimizing the r.h.s. in λ yields $\lambda = \ln(t/\mu)$, which is positive as $t > \mu$, and rearranging concludes the proof. □

Exercise 4.2.1. *Show that* (2) *of Theorem 2.2.1 follows from Theorem 4.2.1. In particular, first show that in the setting of Theorem 2.2.1,* $P(S_N \geq (1+\delta)\mu) \leq e^{\delta\mu}/(1+\delta)^{(1+\delta)\mu}$. *Then use that* $2\delta/(2+\delta) \leq \log(1+\delta)$ *for* $\delta \geq 0$.

Exercise 4.2.2. *Show that, in the setting of Theorem 4.2.1 but for any* $t < \mu$, *one has that* $P(S_N \leq t) \leq e^{-\mu}(e\mu)^t/t^t$.

So far, this section only considered Bernoulli random variables. The next result applies to random variables that are a.s. bounded.

Theorem 4.2.2 (Hoeffding's inequality)**.** *Let* $X_1, \ldots, X_N$ *be independent random variables satisfying* $a_i \leq X_i \leq b_i$ *a.s. Let* $S_N = \sum_{i=1}^N X_i$ *and* $\mu = \mathbb{E}S_N$. *For any* $t > 0$ *it holds that*

$$P(S_N - \mu \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}, \text{ and}$$
$$P(S_N - \mu \leq -t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}.$$

In particular,

$$P(|S_N - \mu| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}.$$

The proof of Hoeffding's inequality relies on the following lemma.

Lemma 4.2.1 (Hoeffding's lemma)**.** *Let* X *be an integrable random variable satisfying* $a \leq X \leq b$ *a.s. and* $\mathbb{E}X = 0$. *Then, for all* $t \in \mathbb{R}$,

$$\mathbb{E}e^{tX} \leq e^{\frac{t^2(b-a)^2}{8}}.$$

For the proof of Hoeffding's lemma, we recall the AM-GM (arithmetic mean-geometric mean) inequality and Taylor's theorem.

Lemma 4.2.2 (AM-GM inequality)**.** *For any* $a, b \geq 0$ *it holds that* $\sqrt{ab} \leq (a+b)/2$.

Theorem 4.2.3 (Taylor's theorem)**.** *Let* $f : [a,b] \to \mathbb{R}$, $n \in \mathbb{N}$, *and* $\alpha, \beta \in [a,b]$ $(\alpha \neq \beta)$. *Suppose that* $f^{(n-1)}$ *is continuous on* $[a,b]$ *and* $f^{(n)}(t)$ *exists for* $t \in (a,b)$. *Then there exists* $x \in (\alpha, \beta)$ *such that*

$$f(\beta) = P(\beta) + \frac{f^{(n)}(x)}{n!}(\beta - \alpha)^n,$$

where $P(t) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(t-\alpha)^k$.

With the preliminaries stated, we now prove Hoeffding's lemma (Lemma 4.2.1).

Proof. For any $x \in [a,b]$, we have $x = \alpha b + (1-\alpha)a$, and solving for α gives $\alpha = (x-a)/(b-a)$. Hence, by the convexity of $x \mapsto e^{tx}$, we obtain

$$e^{tx} = e^{\alpha tb + (1-\alpha)ta} \leq \alpha e^{tb} + (1-\alpha)e^{ta} = \frac{x-a}{b-a}e^{tb} + \frac{b-x}{b-a}e^{ta}.$$

Setting $x = X$, taking expectations, and using that $\mathbb{E}X = 0$ was assumed, we then have

$$\mathbb{E}e^{tX} \leq \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} = e^{g(u)}, \tag{11}$$

with $g(u) = \frac{au}{b-a} + \log\left(1 + \frac{a-ae^u}{b-a}\right)$ and $u = t(b-a)$. The function g has first and second derivatives

$$g'(u) = \frac{a}{b-a} + \frac{ae^u}{b-ae^u} \qquad \text{and} \qquad g''(u) = -\frac{abe^u}{(b-ae^u)^2},$$

respectively, and one has $g(0) = g'(0) = 0$. Now, by Taylor's theorem, there is a $\xi \in (0,u)$ such that

$$g(u) = g(0) + g'(0)u + \frac{1}{2}g''(\xi)u^2,$$

and, consequently, $g(u) \leq u^2/8$ as, by using the AM-GM inequality,

$$g''(u) = -\frac{abe^u}{(b-ae^u)^2} = \frac{(-a)(be^u)}{(b-ae^u)^2} \leq \frac{1}{4}.$$

Combining the bound on $g(u)$ with (11) yields the claim. □

Exercise 4.2.3. *Prove Hoeffding's inequality (Theorem 4.2.2) using the bound from Lemma 4.2.1.*

As an example application of Hoeffding's inequality, consider boosting. Indeed, suppose one has an algorithm for a decision problem that returns the correct result with probability $1/2+\delta$, for some small $\delta > 0$. By running the algorithm N times and taking as answer the majority of the individual answers, one can achieve a probability of success of at least $1-\epsilon$ for any $\epsilon \in (0,1)$ if $N \geq (1/2)\delta^{-2}\log(1/\epsilon)$.

Exercise 4.2.4 (Exercise 2.2.8; Vershynin 2018)**.** *Prove the above claim on boosting decision algorithms.*

The following inequality also takes variance information into account.

Theorem 4.2.4 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_N$ be independent zero-mean random variables satisfying $\sup_i |X_i| \leq B$ a.s. Denote by $S_N = \sum_{i=1}^N X_i$ and by $\sigma^2 = \sum_{i=1}^N \mathbb{E}X_i^2$. Then, for all $t > 0$, it holds that*

$$P(S_N \geq t) \leq \exp\left(-\frac{\frac{1}{2}t^2}{\sigma^2 + \frac{1}{3}Bt}\right)$$

Exercise 4.2.5. *Prove Bernstein's inequality. Hint: Use the expansion* $\mathbb{E}e^{tX} = 1 + \sum_{k=1}^{\infty} \frac{t^k \mathbb{E}X^k}{k!}$.

To close this section, we note that Hoeffding's inequality can be generalized to so-called sub-Gaussian random variables, where a random variable X is called sub-Gaussian iff there exists $C_1 > 0$ such that for all $t > 0$

$$P(|X| \geq t) \leq \exp\Big(-t^2/C_1^2\Big). \tag{12}$$

Indeed, (12) can be shown to be equivalent to the condition on the MGF

$$\mathbb{E}e^{t(X-\mathbb{E}X)} \leq e^{\frac{C_2^2 t^2}{2}},$$

for some $C_2 > 0$, yielding a bound which one may apply with the Chernoff method.

In similar fashion, Bernstein's inequality can be shown to hold for so-called sub-exponential random variables, where a random variable X is said to be sub-exponential iff it satisfies the tail bound

$$P(|X| \geq t) \leq 2\exp(-t/C_3),$$

for some $C_3 > 0$ and any $t > 0$.

4.3 Functions of independent random variables

In the last section, we saw that sums of independent random variables exhibit exponential concentration around their mean. A similar behavior is shown by functions of independent random variables that satisfy a bounded differences property.

Definition 4.3.1 (Bounded differences property)**.** *Let* Ω *be a set,* $x, x' \in \Omega^N$*,* $k \in \{1, \ldots, N\}$*, and denote by* $x^{\setminus k} \in \Omega^N$ *the vector*

$$x_j^{\setminus k} = \begin{cases} x_j & \text{if } j \neq k, \\ x'_k & \text{else.} \end{cases}$$

Then, a function $f : \Omega^N \to \mathbb{R}$ *has the bounded differences property with parameters* $(c_1, \ldots, c_n)$ *if*

$$\sup_{x, x' \in \mathcal{X}^N} |f(x) - f(x^{\setminus k})| \leq c_k$$

for all $k \in \{1, \ldots, N\}$.

The concentration result is as follows.

Theorem 4.3.1 (McDiarmid's inequality)**.** *Let* $X_1, \ldots, X_N$ *be independent and integrable, with* $X_i \in \Omega$*. Suppose* $f : \Omega^N \to \mathbb{R}$ *has the bounded differences*

property with parameters $c_1, \ldots, c_N$. Then, for any $t > 0$, it holds that

$$P(Z - \mathbb{E}Z \geq t) \leq \exp\left(- \frac{2t^2}{\sum_{k=1}^N c_i^2} \right), \text{ and,}$$
$$P(Z - \mathbb{E}Z \leq -t) \leq \exp\left(- \frac{2t^2}{\sum_{k=1}^N c_i^2} \right),$$

where $Z = f(X_1, \ldots, X_N)$. In particular,

$$P(|Z - \mathbb{E}Z| \geq t) \leq 2\exp\left(- \frac{2t^2}{\sum_{k=1}^N c_i^2} \right).$$

The proof is similar to the proof of Hoeffding's inequality (Theorem 4.2.2) but additionally exploits conditional expectations.

Proof. For $i \in \{1, \ldots, N\}$, let

$$d_i = \mathbb{E}[Z \mid X_1, \ldots, X_i] - \mathbb{E}[Z \mid X_1, \ldots, X_{i-1}],$$

and observe that (i) $\mathbb{E}d_i = 0$, and (ii) by telescoping $Z - \mathbb{E}Z = \sum_{i=1}^N d_i$. Further, let

$$L_i = \inf_{x \in \Omega} \mathbb{E}[Z \mid X_1, \ldots, X_{i-1}, x] - \mathbb{E}[Z \mid X_1, \ldots, X_{i-1}], \text{ and}$$
$$U_i = \sup_{x \in \Omega} \mathbb{E}[Z \mid X_1, \ldots, X_{i-1}, x] - \mathbb{E}[Z \mid X_1, \ldots, X_{i-1}],$$

implying that $L_i \leq d_i \leq U_i$ and $U_i - L_i \leq c_i$ by the bounded differences property. As in the proof of the Chernoff inequality (Theorem 4.2.1), we obtain the upper bound, for any $\lambda > 0$,

$$P(Z - \mathbb{E}Z \geq t) \leq e^{-\lambda t} \mathbb{E} \prod_{i=1}^N e^{\lambda d_i}; \tag{13}$$

it remains to bound $\mathbb{E}\prod_{i=1}^N e^{\lambda d_i}$. By using the properties of conditional expectations (Remark 3.6.1),

$$\begin{aligned}
\mathbb{E}\prod_{i=1}^N e^{\lambda d_i} &= \mathbb{E}\mathbb{E}\left[\left(\prod_{i=1}^{N-1} e^{\lambda d_i} \right) e^{\lambda d_N} \mid X_1, \ldots, X_{N-1} \right] \\
&= \mathbb{E}\left(\prod_{i=1}^{N-1} e^{\lambda d_i} \right) \mathbb{E}\left[e^{\lambda d_N} \mid X_1, \ldots, X_{N-1} \right] \\
&\leq \mathbb{E}\left(\prod_{i=1}^{N-1} e^{\lambda d_i} \right) e^{\frac{\lambda^2 c_N^2}{8}},
\end{aligned}$$

where Hoeffding's lemma (Lemma 4.2.1) yields the inequality. Repeating the last two steps $n-1$ times, we have

$$\mathbb{E}\prod_{i=1}^{N} e^{\lambda d_i} \leq \prod_{i=1}^{N} e^{\frac{\lambda^2 c_i^2}{8}} = e^{\frac{\lambda^2 \sum_{i=1}^{N} c_i^2}{8}}. \tag{14}$$

Consequently, by combining (13) and (14),

$$P(Z - \mathbb{E}Z \geq t) \leq e^{-\lambda t} e^{\frac{\lambda^2 \sum_{i=1}^{N} c_i^2}{8}}.$$

Minimizing this expression in $\lambda > 0$ gives $\lambda = 4t/\sum_{i=1}^{N} c_i^2$ and implies the first stated claim. The second claim follows by applying the first result to $-Z$, and a union bound argument concludes the proof of all stated claims. □

Exercise 4.3.1. *Fill in any blanks in the proof of McDiarmid's inequality.*

As an example for the application of McDiarmid's inequality, suppose that $X_1, \ldots, X_N$ are independent random variables with $\sup_i |X_i| \leq b$. Let $S_N = \sum_{i=1}^{N} X_i$, $\mu = \mathbb{E}S_N$, and $f : [-b, b]^N \to \mathbb{R}$ be defined by $(x_1, \ldots, x_N) \mapsto \sum_{i=1}^{N} x_i$. Then

$$\begin{aligned} &\sup_{x_1,\ldots,x_N,x_i' \in [-b,b]} |f(x_1, \ldots, x_i, \ldots, x_N) - f(x_1, \ldots, x_i', \ldots, x_N)| \\ &\quad = \sup_{x_i, x_i' \in [-b,b]} |x_i - x_i'| \leq 2b, \end{aligned}$$

that is, f has the bounded differences property with parameters $(2b, \ldots, 2b)$. Hence, by McDiarmid's inequality, for any $t > 0$,

$$P(|S_N - \mu| \geq t) \leq 2 \exp\left(- \frac{t^2}{2\sum_{k=1}^{N} b^2} \right),$$

corresponding to the result that we would obtain from Hoeffding's inequality.

Remark 4.3.1. *A* martingale *is a sequence of integrable random variables $X_1, X_2, \ldots$ that satisfies $\mathbb{E}[X_i \mid X_1, \ldots, X_{i-1}] = X_i$. The sequence (d_i) that we constructed in the proof of McDiarmid's inequality is a* martingale difference sequence*, characterized by the property that $\mathbb{E}[d_i \mid X_1, \ldots, X_{i-1}] = 0$.*

4.4 Comparison of concentration results

The previous section ended by showing that Hoeffding's inequality is a special case of McDiarmid's inequality. In this section, we give a quick comparison of Chebyshev's, Hoeffding's, and Bernstein's inequality.

Suppose that the random variables $X_1, \ldots, X_N$ are independent, have zero-mean, and satisfy $\sup_i |X_i| \leq B$ a.s. Let $S_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ (observe that we now

consider the empirical mean). Applying Hoeffding's inequality (Theorem 4.2.2), we then have for any $t > 0$ that

$$P(|S_N| \geq t) \leq 2e^{-\frac{t^2 N}{2B^2}}. \tag{15}$$

In the same setting, Bernstein's inequality (Theorem 4.2.4) reads

$$P(|S_N| \geq t) \leq 2\exp\left(-\frac{t^2 N}{2\frac{\sigma^2}{N} + \frac{2}{3}Bt}\right). \tag{16}$$

Using the trivial bound $\sigma^2 \leq NB^2$, the exponent on the r.h.s. of the last displayed equation becomes

$$-\frac{t^2 N}{2B^2 + \frac{2}{3}Bt}. \tag{17}$$

Comparing (17) with the exponent in (15), we see that in this case Hoeffding's inequality gives a tighter bound for any $t > 0$.

Suppose now that $\sigma^2 = 0$. In this case the exponent in (16) becomes $-3tN/2B$, which is smaller than the exponent in (15) for any $0 < t < 3B$. Notice that for $t \geq B$ the probability in both statements is zero. Bernstein's inequality gives a much better bound for random variables with small variances.

For small sample sizes N, Chebyshev's inequality gives a tighter bound than either the Hoeffding or the Bernstein inequality.

4.5 Notes

All of the results in this section are very well-known. The de facto reference on concentration inequalities is Boucheron et al. [2013] and goes way beyond what was covered in this section.

Taylor's theorem, used in the proof of Hoeffding's lemma, is copied from Rudin [1976]. An instructive alternative proof of Hoeffding's lemma (Lemma 4.2.1) with a probabilistic argument—but giving slightly weaker bounds—is in Duchi, Section 3.

Our proof of McDiarmid's inequality mostly follows the proof of Ledoux and Talagrand [1991, Lemma 1.5] but uses Hoeffding's lemma. For a proof of Bernstein's inequality, we refer to Vershynin [2018]. All the results are also developed in Wainwright [2019], using the sub-Gaussian (resp. sub-exponential) framework.

For concentration of random matrices, which we did not cover here, see, e.g., Tropp [2012] and Tao [2012].

5 Regression

Let the random variables $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ have joint distribution P. The goal is to find a (measurable) function $f : \mathbb{R}^p \to \mathbb{R}$ to predict Y given X and we assess the quality of f using a loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$; lower loss indicates better quality. Let us consider the squared loss $(x, y) \mapsto (x - y)^2$. Given some f, the expected prediction loss is

$$\mathrm{EPE}(f) = \mathbb{E}L(Y, f(X)) = \mathbb{E}(Y - f(X))^2 = \mathbb{E}\mathbb{E}[(Y - f(X))^2 \mid X].$$

From Section 3.6, it is immediate that $f(X) = \mathbb{E}(Y \mid X)$—the conditional expectation—is the minimizer of the preceding display. In this context, the conditional expectation is also known as *regression function*, regressing X onto Y.

In this section, within the framework of empirical risk minimization, we investigate linear regression and ridge regression, where $f = f_{\boldsymbol{\beta}}$ is assumed to be a function parameterized by some $\boldsymbol{\beta}$ and linear in its parameters.

5.1 Empirical risk minimization

The above display is an instance of a general learning paradigm, called empirical risk minimization (ERM), which we quickly elaborate before returning to linear regression. We begin by introducing the expected risk associated to a predictor, followed by associated quantities.

Definition 5.1.1 (Expected risk)**.** *Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y}), P)$ be a probability space and $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ a measurable function, called loss function. For a measurable function $f : \mathcal{X} \to \mathcal{Y}$, the expected risk associated to P and L is defined to be*

$$\mathcal{R}(f) = \mathbb{E}L(Y, f(X)),$$

where $(X, Y) \sim P$. We assume that lower values of L correspond to better predictions.

As we assumed L to be non-negative, the expected risk is well-defined but can be infinite.

A *Bayes predictor* $f^* : \mathcal{X} \to \mathcal{Y}$ associated to P and L is any function satisfying

$$f^*(x) \in \operatorname*{arg\,min}_{y \in \mathcal{Y}} \mathbb{E}\left[L(Y, y) \mid X = x\right] \tag{18}$$

for all $x \in \mathcal{X}$.[6] The Bayes predictor achieves the lowest expected risk among all measurable f.

Theorem 5.1.1 (Optimality of Bayes predictor)**.** *In the setting of Definition 5.1.1, $f \mapsto R(f)$ is minimized at f^* as defined in* (18)*.*

[6] For technical reasons, we require $L(Y, y) \in \mathcal{L}^1$ for all $y \in \mathcal{Y}$.

Proof. By (18), f^* satisfies for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ that

$$\mathbb{E}\left[L(Y, y) \mid X = x\right] \geq \mathbb{E}\left[L(Y, f^*(x)) \mid X = x\right]. \tag{19}$$

Hence, using twice the tower property of conditional expectations, for any measurable $f : \mathcal{X} \to \mathcal{Y}$,

$$\begin{aligned}\mathcal{R}(f) &= \mathbb{E}L(Y, f(X)) = \mathbb{E}\mathbb{E}\left[L(Y, f(x)) \mid X = x\right] \overset{(19)}{\geq} \mathbb{E}\mathbb{E}\left[L(Y, f^*(x)) \mid X = x\right] \\ &= \mathbb{E}L(Y, f^*(X)) = \mathcal{R}(f^*).\end{aligned}$$

□

Notice that $\mathcal{R}(f^*) = \mathbb{E}\inf_{y\in\mathcal{Y}} \mathbb{E}\left[L(Y, y) \mid X = x\right] =: R^*$, which we call *Bayes risk* and use as a baseline.

Definition 5.1.2 (Excess risk)**.** *In the setting of Definition* (5.1.1)*, the excess risk of a measurable function $f : \mathcal{X} \to \mathcal{Y}$ is $\mathcal{R}(f) - R^*$.*

Due to the optimality property of $\mathcal{R}^*$ (Theorem 5.1.1) the excess risk is non-negative.

In practice, our goal is to find some $f : \mathcal{X} \to \mathcal{Y}$ that is similar to f^* but we do not have access to P. Instead, we observe samples, which we use to estimate $\mathcal{R}(f)$.

Definition 5.1.3 (Empirical risk)**.** *In the setting of Definition 5.1.1, assume that $((X_i, Y_i))_{i=1}^N \overset{i.i.d.}{\sim} P$. The empirical risk takes the form*

$$\hat{\mathcal{R}}_N(f) = \frac{1}{N}\sum_{i=1}^{N} L(Y_i, f(X_i)).$$

Put differently, the empirical risk corresponds to the risk associated to the empirical measure $\hat{P}_N = \frac{1}{N}\sum_{i=1}^N \delta_{(x_i,y_i)}$, where δ_x denotes the Dirac measure centered at $x \in \mathcal{X} \times \mathcal{Y}$.

A trivial solution to minimize $f \mapsto \hat{R}_N(f)$ given samples $((X_i, Y_i))_{i=1}^N$ is given by the function

$$f(x) = \begin{cases} Y_i & \text{if } x = X_i \quad \text{for } i = 1, \ldots, N, \\ 0 & \text{else,} \end{cases}$$

where we assume that each X_i is only observed once. To mitigate this issue, ERM restricts the class of functions over which the empirical risk is minimized. In linear regression, one assumes that f is a linear function.

Remark 5.1.1. *By SLLN, for any fixed measurable $f : \mathcal{X} \to \mathcal{Y}$, $\hat{R}_N(f) \to R(f)$ as $N \to \infty$ a.s. In practice, f depends on the data and the SLLN is not sufficient. Instead, one requires uniform convergence over the class of functions that one considers.*

5.2 Linear regression

For $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$, the linear model has the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \tag{20}$$

that is, one assumes that the regression function $f(X) = \mathbb{E}(Y \mid X)$ is (roughly) linear.

In practice, we neither know $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1}$, nor P. Rather, we observe data $(x_1, y_1), \ldots, (x_N, y_N)$ from which we estimate $\hat{\boldsymbol{\beta}}$. The most popular method for estimating the parameter is *least squares*, minimizing the residual sum of squares

$$\begin{aligned} \mathrm{RSS}(\boldsymbol{\beta}) &= \sum_{i=1}^{N} L(y_i, f(x_i)) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \end{aligned}$$

using the data observed. Here, $\mathbf{X} \in \mathbb{R}^{N \times p+1}$, with each row an input vector and a 1 in the first position (to account for β_0), and $\mathbf{y} = (y_1, \ldots, y_N) \in \mathbb{R}^N$.

Theorem 5.2.1. *Let $N, p \in \mathbb{N}$, $\mathbf{y} \in \mathbb{R}^N$, $\mathbf{X} \in \mathbb{R}^{N \times p+1}$, and assume that $\mathbf{X}^\top \mathbf{X}$ has full rank. Then, the minimizer of*

$$\boldsymbol{\beta} \mapsto \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \tag{21}$$

is unique and takes the form

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{22}$$

Proof. Set the first derivative of (21) equal to 0, solve for $\boldsymbol{\beta}$, and observe that the second derivative of (21) is positive definite. □

We note that (i) $\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}$ is known as normal equation, (ii) $\mathbf{X}$ is called design matrix, and (iii) $\hat{\boldsymbol{\beta}}$ is the ordinary least squares estimator. The expectation and the variance of $\hat{\boldsymbol{\beta}}$, when assuming a specific model, are collected in the following result.

Theorem 5.2.2. *Let $Y = \mathbf{X}\boldsymbol{\beta}^* + \epsilon \in \mathbb{R}^N$, with $\mathbf{X} \in \mathbb{R}^{N \times p+1}$ nonrandom, $\epsilon = (\epsilon_i)_{i=1}^N \in \mathbb{R}^N$, ϵ_i uncorrelated ($\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for any $i \neq j$), $\mathbb{E}\epsilon_i = 0$, $\mathbb{E}\epsilon_i^2 = \sigma^2 < \infty$, and $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y$. Then*

(i) $\mathbb{E}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$, *and*

(ii) $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$.

Proof. For (i), direct calculation yields

$$\mathbb{E}\hat{\boldsymbol{\beta}} = \mathbb{E}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y = \mathbb{E}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}^* + \mathbb{E}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon = \boldsymbol{\beta}^*.$$

Regarding (ii), we observe that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}^* + \epsilon) - \boldsymbol{\beta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon$. Hence,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathbb{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top = \mathbb{E}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon\epsilon^\top\mathbf{X}(\boldsymbol{X}^\top\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

as $\mathbb{E}\epsilon\epsilon^\top = \mathrm{Var}(\epsilon) = \sigma^2\mathbf{I}_N$. □

The preceding result shows that the estimator $\hat{\boldsymbol{\beta}}$ is unbiased. Indeed, the estimate (22) is the *best linear unbiased estimator*, captured by the famous Gauss-Markov theorem. Here, *best* means minimum variance and *linear* means that the estimator is a linear function of Y; any linear estimator has the form $\mathbf{A}Y$ for some $\mathbf{A} \in \mathbb{R}^{p+1\times N}$.

Before stating the result, we recall a few facts from linear algebra. A linear operator $P : V \to V$ acting on a vector space V is a projection iff $P^2 = P$. For an inner product space $(V, \langle\cdot,\cdot\rangle)$, the projection is orthogonal iff $\langle Px, y\rangle = \langle x, Py\rangle$ for all $x, y \in V$. Further, for two symmetric matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p+1\times p+1}$, $\mathbf{A} \leq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive semi-definite (written as $\mathbf{B} - \mathbf{A} \geq \mathbf{0}$).

Theorem 5.2.3 (BLUE; Gauss-Markov)**.** *Let $Y = \mathbf{X}\boldsymbol{\beta}^* + \epsilon$, with $\mathbf{X} \in \mathbb{R}^{N\times p+1}$ nonrandom, ϵ_i uncorrelated, $\mathbb{E}\epsilon_i = 0$, and $\mathbb{E}\epsilon_i^2 = \sigma^2 < \infty$. Then*

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) \leq \mathrm{Var}\left(\tilde{\boldsymbol{\beta}}\right),$$

where $\hat{\boldsymbol{\beta}}$ is as in (22) *and $\tilde{\boldsymbol{\beta}}$ is any linear unbiased estimator.*

Proof. Recalling from (22) that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y$$

shows that $\hat{\boldsymbol{\beta}}$ is linear. Unbiasedness was shown in Theorem 5.2.2(i).

As $\tilde{\boldsymbol{\beta}}$ is assumed linear and unbiased, we must have

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}Y \text{ and } \mathbb{E}\tilde{\boldsymbol{\beta}} = \mathbb{E}\mathbf{A}Y = \mathbf{A}\mathbf{X}\boldsymbol{\beta}^* \stackrel{!}{=} \boldsymbol{\beta}^* \text{ for all } \boldsymbol{\beta}^*,$$

for some $\mathbf{A} \in \mathbb{R}^{p+1\times N}$, implying that $\mathbf{A}\mathbf{X} = \mathbf{I}$. Noting that $\mathrm{Var}(\tilde{\boldsymbol{\beta}}) = \sigma^2\mathbf{A}\mathbf{A}^\top$ and using Theorem 5.2.2(ii) gives $\mathrm{Var}(\tilde{\boldsymbol{\beta}}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{A}\mathbf{A}^\top - (\mathbf{X}^\top\mathbf{X})^{-1})$. As $\sigma^2 > 0$ it is sufficient to show that $\mathbf{A}\mathbf{A}^\top - (\mathbf{X}^\top\mathbf{X})^{-1} \geq \mathbf{0}$. We have

$$\begin{aligned}\mathbf{A}\mathbf{A}^\top - (\mathbf{X}^\top\mathbf{X})^{-1} &= \mathbf{A}\mathbf{A}^\top - \mathbf{A}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{A}^\top = \mathbf{A}\mathbf{A}^\top - \mathbf{A}\mathbf{H}\mathbf{A}^\top \\ &= \mathbf{A}(\mathbf{I} - \mathbf{H})\mathbf{A}^\top\end{aligned}$$

with the projection $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. Now, for any $\mathbf{x} \in \mathbb{R}^{p+1}$,

$$\langle \mathbf{x}, \mathbf{A}(\mathbf{I} - \mathbf{H})\mathbf{A}^\top\mathbf{x}\rangle = \langle (\mathbf{I} - \mathbf{H})\mathbf{A}^\top\mathbf{x}, (\mathbf{I} - \mathbf{H})\mathbf{A}^\top\mathbf{x}\rangle = \|(\mathbf{I} - \mathbf{H})\mathbf{A}^\top\mathbf{x}\|_2^2 \geq 0,$$

showing that $\mathbf{A}(\mathbf{I} - \mathbf{H})\mathbf{A}^\top \geq \mathbf{0}$ and proving the claim. □

Exercise 5.2.1. *The mean-squared error of an estimator $\boldsymbol{\beta} \in \mathbb{R}^p$ of some parameter $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is defined to be* $\mathrm{MSE}(\boldsymbol{\beta}) = \mathbb{E}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2$. *Show that one has the decomposition*

$$\mathrm{MSE}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}^* - \mathbb{E}\boldsymbol{\beta}\|_2^2 + \mathrm{tr}(\mathrm{Var}(\boldsymbol{\beta}))$$

which is known as bias variance decomposition.

5.3 Fixed design analysis

In the following, we analyze the expected excess risk of the ordinary least squares estimator in the *fixed design* setting. Here, the $\mathbf{x}_i$-s are assumed to be non-random and the only randomness is in the Y_i-s. Importantly, in the fixed design setting, we do not make any assertion on the generalization of the learned function, that is, its performance on unseen data.

We again assume that we have samples $((\mathbf{x}_i, Y_i))_{i=1}^N$, $Y_i = (1, \mathbf{x}_i)^\top \boldsymbol{\beta}^* + \epsilon_i$, and that the ϵ_i-s are i.i.d. with $\mathbb{E}\epsilon_1 = 0$ and $\mathbb{E}\epsilon_1^2 = \sigma^2 > 0$. Let $\epsilon = (\epsilon_i)_{i=1}^N \in \mathbb{R}^N$.

Additionally, we assume that we are interested in minimizing the risk

$$\mathcal{R}(\boldsymbol{\beta}) = \mathbb{E}\frac{1}{N}\|Y - \mathbf{X}\boldsymbol{\beta}\|_2^2,$$

over the class of *linear functions*, with $Y = (Y_i)_{i=1}^N \in \mathbb{R}^N$ and $\mathbf{X} = ((1, \mathbf{x}_i))_{i=1}^N \in \mathbb{R}^{N \times p+1}$ (assumed to have full column rank). Note that the expectation is over the randomness in ϵ.

Lemma 5.3.1 (Optimal linear predictor)**.** *Assume the above setting. For any $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, it holds that*

$$\mathcal{R}(\boldsymbol{\beta}) \geq \mathcal{R}(\boldsymbol{\beta}^*).$$

Hence, $\mathcal{R}^* = \mathcal{R}(\boldsymbol{\beta}^*)$. Moreover, $\mathcal{R}^* = \sigma^2$.

The proof is similar to that of Theorem 5.1.1 but uses that (i) the relationship of Y on $\mathbf{X}$ is known, and (ii) one minimizes over linear functions.

Proof. As $Y = \mathbf{X}\boldsymbol{\beta}^* + \epsilon$, we have for any $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ that

$$\begin{aligned}
\mathcal{R}(\boldsymbol{\beta}) &= \mathbb{E}\frac{1}{N}\|\mathbf{X}\boldsymbol{\beta}^* + \epsilon - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \frac{1}{N}\mathbb{E}\|\mathbf{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) + \epsilon\|_2^2 \\
&= \frac{1}{N}\mathbb{E}\big[\|\mathbf{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\|_2^2 + \|\epsilon\|_2^2 - 2\langle \mathbf{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}), \epsilon\rangle_2\big] \\
&= \frac{1}{N}\|\mathbf{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\|_2^2 + \sigma^2, \qquad (23)
\end{aligned}$$

by using that $\mathbb{E}\epsilon = \mathbf{0}$ and $\mathbb{E}\|\epsilon\|^2 = N\sigma^2$. As $\mathbf{X}$ has full column rank, the minimum is unique and attained at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. □

The above display (23) also implies that the risk of any linear predictor is at least σ^2, which is called *irreducible error*.

We can now establish the expected excess risk of the least squares predictor (22), which was given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y$.

Theorem 5.3.1 (Expected excess risk)**.** *It holds that*

$$\mathbb{E}\mathcal{R}(\hat{\boldsymbol{\beta}}) - \mathcal{R}^* = \frac{\sigma^2(p+1)}{N}.$$

Proof. By (23), $\mathcal{R}(\hat{\boldsymbol{\beta}}) - \mathcal{R}^* = \frac{1}{N}\|\mathbf{X}\boldsymbol{\Delta}\|^2$, with $\boldsymbol{\Delta} = \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}$. Hence,

$$\begin{aligned} N(\mathbb{E}\mathcal{R}(\hat{\boldsymbol{\beta}}) - \mathcal{R}^*) &= \mathbb{E}\|\mathbf{X}\boldsymbol{\Delta}\|^2 = \mathbb{E}\boldsymbol{\Delta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\Delta} = \operatorname{tr}\big[\mathbb{E}\boldsymbol{\Delta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\Delta}\big] \\ &= \mathbb{E}\operatorname{tr}\big[\boldsymbol{\Delta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\Delta}\big] = \mathbb{E}\operatorname{tr}\big[\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\mathbf{X}^\top\mathbf{X}\big] \\ &= \operatorname{tr}\big[\mathbb{E}\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\mathbf{X}^\top\mathbf{X}\big] = \operatorname{tr}\Big[\operatorname{Var}(\hat{\boldsymbol{\beta}})\mathbf{X}^\top\mathbf{X}\Big] \\ &= \sigma^2\operatorname{tr}(\mathbf{I}_{p+1}) = \sigma^2(p+1), \end{aligned}$$

by the 'trace trick', the cyclic invariance of the trace, Theorem 5.2.2(ii), and as $\operatorname{tr}(\mathbf{I}_{p+1}) = p+1$, where $\mathbf{I}_{p+1}$ is the $p+1$-dimensional identity matrix.

Dividing by N yields the stated result and concludes the proof. □

Regarding the trace trick, observe that for a random square matrix $\mathbf{A} = (a_{ij})_{i,j=1}^d$, it holds that

$$\operatorname{tr}[\mathbb{E}\mathbf{A}] = \operatorname{tr}\big[(\mathbb{E}a_{ij})_{i,j=1}^d\big] = \sum_{i=1}^d \mathbb{E}a_{ii} = \mathbb{E}\sum_{i=1}^d a_{ii} = \mathbb{E}\operatorname{tr}[\mathbf{A}].$$

Theorem 5.3.1 has two key take-aways. First, the expected excess risk converges with a rate that is linear in the sample size N, which is rather fast. Second, if p is close to N—or even $p = N$—the expected excess risk is close to σ^2. Note that $p > N$ implies that $\mathbf{X}$ does not have full column rank; hence, $\mathbf{X}^\top\mathbf{X}$ is singular and $\hat{\boldsymbol{\beta}}$ is not defined.

Ridge regression, detailed in the following section, performs regularization and allows to handle large p. Another possibility for handling high-dimensional data is reducing their dimensionality, detailed in later sections.

5.4 Ridge regression

Let us recall from Exercise 5.2.1 that the mean squared error, providing a notion of quality of the estimator $\boldsymbol{\beta}$ of $\boldsymbol{\beta}^*$, has the bias-variance decomposition

$$\mathrm{MSE}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}^* - \mathbb{E}\boldsymbol{\beta}\|_2^2 + \mathrm{tr}(\mathrm{Var}(\boldsymbol{\beta})).$$

Theorem 5.2.3 showed that the estimator (22), which we now denote by $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$, has minimum variance among all unbiased estimators. Hence, to reduce the MSE, one must consider biased estimators.

Independent of the aim of reducing the MSE, $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ also can have practical issues: (i) As already noted, the estimator is undefined in case $\mathbf{X}$ does not have full column rank (which is implied if $p > N$). (ii) If two inputs x_i, x_j $(i \neq j)$ are highly correlated, their associated coefficients β_i, β_j can exhibit high variance.

To mitigate these issues, we introduce *ridge regression*, which is a minimizer of the regularized empirical loss

$$\hat{\mathcal{R}}_N(f) + \lambda\Omega(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda\Omega(f), \quad \lambda \geq 0 \tag{24}$$

where $\Omega(f)$ measures the "complexity" of the predictor f. In other words, minimizing (24) allows trading-off loss and model complexity, controlled by $\lambda \geq 0$. Indeed, the ridge regression estimator is the minimizer of

$$\boldsymbol{\beta} \mapsto \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \tag{25}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathbf{y} = (y_i)_{i=1}^N \in \mathbb{R}^N$, $\mathbf{X} = (\mathbf{x}_i)_{i=1}^N \in \mathbb{R}^{N\times p}$, $\beta_0 = \frac{1}{N}\sum_{i=1}^N y_i = \bar{\mathbf{y}}$, and each x_{ij} is replaced by $x_{ij} - \bar{\mathbf{x}}_j$. In other words, β_0 is not part of the optimization—note that $\mathbf{X}$ accordingly has p columns only—and the $\mathbf{x}_i$-s have been centered.

The minimizer $\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}$ of the above display is collected in the following result.

Theorem 5.4.1. *Let $\lambda > 0$, $N, p \in \mathbb{N}$, $\mathbf{y} \in \mathbb{R}^N$, and $\mathbf{X} \in \mathbb{R}^{N\times p}$. Then, the minimizer of*

$$\boldsymbol{\beta} \mapsto \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{26}$$

is unique and takes the form

$$\hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{27}$$

Proof. Set the first derivative of (26) equal to 0, solve for $\boldsymbol{\beta}$, and observe that the Hessian of (26) is positive definite. □

Exercise 5.4.1. *Show that $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is invertible for any $\lambda > 0$.*

Another idea to restrict the sizes of the β_j-s is to consider the constrained optimization problem

$$\begin{aligned} &\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \\ &\text{subject to } \|\boldsymbol{\beta}\|_2^2 \leq t, \end{aligned} \tag{28}$$

for some $t > 0$, which makes the size constraints explicit. In fact, formulations (26) and (28) have the same solution.

Theorem 5.4.2. *In the setting of Theorem 5.4.1,* (27) *is the unique solution to* (28).

Exercise 5.4.2. *Prove Theorem 5.4.2.*

Besides establishing the Gauss-Markov theorem (Theorem 5.2.3), one interest in our analysis of $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ was understanding its bias and variance (see Theorem 5.2.2) when the true model is assumed to take the form $Y = \mathbf{X}\boldsymbol{\beta}^* + \epsilon$, with the only randomness in the noise ϵ. For $\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}$ we have the following result, allowing a direct comparison of both estimators.

Theorem 5.4.3. *Let $Y = \mathbf{X}\boldsymbol{\beta}^* + \epsilon \in \mathbb{R}^N$, with $\mathbf{X} \in \mathbb{R}^{N\times p}$ nonrandom, $\epsilon = (\epsilon_i)_{i=1}^N \in \mathbb{R}^N$, ϵ_i uncorrelated, $\mathbb{E}\epsilon_i = 0$, $\mathbb{E}\epsilon_i^2 = \sigma^2 < \infty$, and $\hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top Y$ for some $\lambda > 0$. Then*

(i) $\mathbb{E}\hat{\boldsymbol{\beta}}^{ridge} = \boldsymbol{\beta}^* - \lambda(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}^*$, *and*

(ii) $\mathrm{Var}(\hat{\boldsymbol{\beta}}^{ridge}) = \sigma^2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}$.

Proof. To obtain (i), we notice that $\mathbb{E}Y = \mathbf{X}\boldsymbol{\beta}^*$ and introduce $\pm\lambda\mathbf{I}$, yielding

$$\begin{aligned}\mathbb{E}\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} &\overset{(27)}{=} \mathbb{E}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}^* \\ &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I} - \lambda\mathbf{I})\boldsymbol{\beta}^* = (\mathbf{I} - \lambda(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1})\boldsymbol{\beta}^*,\end{aligned} \tag{29}$$

which, after simplifying, is the stated result. For (ii), we first observe that by using the r.h.s. of (29), one gets $\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} - \mathbb{E}\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\epsilon$. Now computing $\mathrm{Var}(\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}) = \mathbb{E}(\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} - \mathbb{E}\hat{\boldsymbol{\beta}}^{\mathrm{ridge}})(\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} - \mathbb{E}\hat{\boldsymbol{\beta}}^{\mathrm{ridge}})^\top$ using that $\mathbb{E}\epsilon\epsilon^\top = \sigma^2\mathbf{I}$ yields the claim. □

Exercise 5.4.3. *In the setting of Theorem 5.4.3, what happens w.r.t. the bias and variance of $\hat{\boldsymbol{\beta}}^{ridge}$ for $\lambda \to \infty$ (for $\lambda \to 0$)?*

Exercise 5.4.4. *Theorem 5.4.3(i) shows that for $\lambda > 0$ the ridge estimator is biased. Together with the expression of the variance in (ii), show that there exists $\lambda > 0$ such that $\mathrm{MSE}(\hat{\boldsymbol{\beta}}^{ridge}) \leq \mathrm{MSE}(\hat{\boldsymbol{\beta}}^{OLS})$.*

Naturally, the regularized estimator gives rise to the "reconstruction" $\hat{\mathbf{y}}^{\mathrm{ridge}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}$. To compare $\hat{\mathbf{y}}^{\mathrm{ridge}}$ with $\hat{\mathbf{y}}^{\mathrm{OLS}}$, we recall the singular value decomposition (SVD) of a matrix.

Theorem 5.4.4 (Singular value decomposition)**.** *Suppose $\mathbf{A} \in \mathbb{R}^{N\times p}$ has rank r. Then there exist orthogonal matrices $\mathbf{U} \in \mathbb{R}^{N\times N}$ and $\mathbf{V} \in \mathbb{R}^{p\times p}$ such that*

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top,$$

where $\mathbf{D} = \begin{bmatrix}\mathbf{S} & 0\\ 0 & 0\end{bmatrix}$, $\mathbf{S} = \mathrm{diag}(d_1, \ldots, d_r) \in \mathbb{R}^{r\times r}$, *and* $d_1 \geq \cdots \geq d_r > 0$.

Indeed, applying the SVD to the matrix $\mathbf{X}$ (assumed to have full column rank), we get $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, and find that

$$\hat{\mathbf{y}}^{\mathrm{OLS}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \mathbf{U}\mathbf{U}^\top\mathbf{y} = \sum_{j=1}^{p} \mathbf{u}_j\mathbf{u}_j^\top\mathbf{y}, \text{ and} \tag{30}$$

$$\hat{\mathbf{y}}^{\mathrm{ridge}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} = \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^\top\mathbf{y} \ = \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda}\mathbf{u}_j^\top\mathbf{y}. \tag{31}$$

Note that the $\mathbf{u}_j$-s, the orthogonal rows of $\mathbf{U}$, form an orthonormal basis (ONB) of $\mathbb{R}^p$. In the case of ordinary least squares, (30) shows that $\mathbf{y}$ is expressed in this new basis. For ridge regression, (31) shows that, while $\mathbf{y}$ is also expressed in this new basis, each coefficient $\mathbf{u}_j^\top\mathbf{y}$ is scaled depending on the j-th eigenvalue and the regularization parameter $\lambda > 0$. For $d_j^2 \gg \lambda$, the impact of the regularization is negligible; for $d_j^2 \ll \lambda$, the impact is large. We will see a correspondence of the eigenvalues and -vectors to the variance of the data in the section on principal component analysis, allowing us to interpret this observation.

By introducing regularization to least squares regression, we succeeded in handling $p \gg N$. However, note that the computation of $\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}$ requires inverting $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I} \in \mathbb{R}^{p\times p}$, which has a runtime complexity of $O(p^3)$ in practice. The matrix inversion lemma–a special case of the *Woodbury matrix identity*–allows "switching" to a runtime complexity of $O(N^3)$.

Lemma 5.4.1 (Matrix inversion)**.** *Let* $\mathbf{X} \in \mathbb{R}^{N\times p}$*. Then*

$$(\mathbf{X}^\top\mathbf{X} + \mathbf{I}_p)^{-1}\mathbf{X}^\top = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mathbf{I}_N)^{-1}.$$

Proof. Multiply $\mathbf{X}^\top\mathbf{X} + \mathbf{I}_p$ from the left and $\mathbf{X}\mathbf{X}^\top + \mathbf{I}_N$ from the right. □

Indeed, applying Lemma 5.4.1 to $\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}$ in (27) shows that

$$\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{y}.$$

We close this section by investigating the risk of ridge regression in the fixed design setting.

5.5 Fixed design analysis

Our goal in this section is to compare the expected excess risk of ordinary least squares regression with that of ridge regression. Hence, we again consider the setting of Section 5.3 but adapt the conventions of Section 5.4, that is, we assume that the data has been centered and we do not explicitly consider the intercept β_0.

We start by obtaining the expected excess risk of ridge regression in closed-form.

Theorem 5.5.1 (Expected excess risk)**.** *Let* $\mathbf{S} = \mathbf{X}^\top \mathbf{X}$ *and* $\mathbf{S}_\lambda = \mathbf{S} + \lambda \mathbf{I}$ *for* $\lambda > 0$*. It holds that*

$$\mathbb{E}\mathcal{R}(\hat{\boldsymbol{\beta}}^{ridge}) - \mathcal{R}^* = \frac{\lambda^2}{N} \boldsymbol{\beta}^{*\top} \mathbf{S} \mathbf{S}_\lambda^{-2} \boldsymbol{\beta}^* + \frac{\sigma^2}{N} \operatorname{tr}(\mathbf{S}^2 \mathbf{S}_\lambda^{-2}).$$

Proof. Abbreviating $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\mathrm{ridge}}$, we have the decomposition

$$\begin{aligned} N(\mathbb{E}\mathcal{R}(\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}) - \mathcal{R}^*) &\overset{(23)}{=} \mathbb{E}\|\mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})\|_2^2 \\ &= \|\mathbf{X}\boldsymbol{\beta}^*\|_2^2 + \mathbb{E}\|\mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 - 2\mathbb{E}\langle \mathbf{X}\boldsymbol{\beta}^*, \mathbf{X}\hat{\boldsymbol{\beta}} \rangle. \end{aligned} \tag{32}$$

The second term can be expressed as

$$\mathbb{E} \operatorname{tr}\Big(\hat{\boldsymbol{\beta}}^\top \mathbf{S} \hat{\boldsymbol{\beta}}\Big) = \operatorname{tr}\Big(\mathbf{S}\mathbb{E}\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^\top\Big) = \boldsymbol{\beta}^{*\top} \mathbf{S}^3 \mathbf{S}_\lambda^{-2} \boldsymbol{\beta}^* + \sigma^2 \operatorname{tr}(\mathbf{S}^2 \mathbf{S}_\lambda^{-2}),$$

by using the cyclic invariance property of the trace, $\mathbb{E}YY^\top = \mathbf{X}\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\top}\mathbf{X}^\top + \sigma^2\mathbf{I}$, and the fact that $\mathbf{S}$ and $\mathbf{S}_\lambda^{-1}$ commute (as is seen by invoking Lemma 5.4.1 twice). Using that $\mathbb{E}\hat{\boldsymbol{\beta}} = \mathbf{S}_\lambda^{-1}\mathbf{S}\boldsymbol{\beta}^*$, the third term is

$$-2\boldsymbol{\beta}^{*\top}\mathbf{S}\mathbb{E}\hat{\boldsymbol{\beta}} = -2\boldsymbol{\beta}^{*\top}\mathbf{S}\mathbf{S}_\lambda^{-1}\mathbf{S}\boldsymbol{\beta}^* = -2\boldsymbol{\beta}^{*\top}\mathbf{S}^2\mathbf{S}_\lambda^{-1}\boldsymbol{\beta}^*,$$

where we again swapped $\mathbf{S}$ and $\mathbf{S}_\lambda^{-1}$. Combining the terms yields

$$\begin{aligned} (32) &= \boldsymbol{\beta}^{*\top}\mathbf{S}\boldsymbol{\beta}^* + \boldsymbol{\beta}^{*\top}\mathbf{S}^3\mathbf{S}_\lambda^{-2}\boldsymbol{\beta}^* + \sigma^2 \operatorname{tr}(\mathbf{S}^2\mathbf{S}_\lambda^{-2}) - 2\boldsymbol{\beta}^{*\top}\mathbf{S}^2\mathbf{S}_\lambda^{-1}\boldsymbol{\beta}^* \\ &= \boldsymbol{\beta}^{*\top}(\mathbf{S} + \mathbf{S}^3\mathbf{S}_\lambda^{-2} - 2\mathbf{S}^2\mathbf{S}_\lambda^{-1})\boldsymbol{\beta}^* + \sigma^2 \operatorname{tr}(\mathbf{S}^2\mathbf{S}_\lambda^{-2}) \\ &= \lambda^2 \boldsymbol{\beta}^{*\top}\mathbf{S}\mathbf{S}_\lambda^{-2}\boldsymbol{\beta}^* + \sigma^2 \operatorname{tr}(\mathbf{S}^2\mathbf{S}_\lambda^{-2}), \end{aligned}$$

where introducing $\mathbf{I} = \mathbf{S}_\lambda^{-1}\mathbf{S}_\lambda$ and simplifying repeatedly gives the last equality. Dividing the result by N concludes the proof. □

By bounding the expected excess risk of ridge regression, we determine an optimal choice of the regularization parameter λ.

Theorem 5.5.2 (Expected excess risk bound)**.** *Let* $\lambda^* = \sigma\sqrt{\operatorname{tr}(\mathbf{S})}/\|\boldsymbol{\beta}^*\|_2$*. Then it holds that*

$$\mathbb{E}\mathcal{R}(\hat{\boldsymbol{\beta}}^{ridge}) - \mathcal{R}^* \leq \frac{\sigma\sqrt{\operatorname{tr}(\mathbf{S})}\|\boldsymbol{\beta}^*\|_2}{2N}.$$

Proof. Let λ_j denote the eigenvalues of $\mathbf{S}$. This implies that $\mathbf{S}_\lambda^{-1}\mathbf{S}\mathbf{S}_\lambda^{-1}$ has eigenvalues $\lambda_j/(\lambda_j + \lambda)^2$. Hence, using the C^* property of bounded linear operators and the self-adjointness of $\mathbf{S}_\lambda^{-1}\mathbf{S}\mathbf{S}_\lambda^{-1}$, we get

$$\lambda\|\mathbf{S}^{1/2}\mathbf{S}_\lambda^{-1}\|_{\mathrm{op}}^2 = \lambda\|\mathbf{S}_\lambda^{-1}\mathbf{S}\mathbf{S}_\lambda^{-1}\|_{\mathrm{op}} = \max_{j \in 1,\ldots,p} \frac{\lambda\lambda_j}{(\lambda_j + \lambda)^2} \leq \frac{1}{4}, \tag{33}$$

by using AM-GM (Lemma 4.2.2) in the last inequality. We continue by bounding both terms in the result of Theorem 5.5.1 separately.

- **Term 1.** The first term satisfies the bound
$$\frac{\lambda^2}{N}\boldsymbol{\beta}^{*\top}\mathbf{S}\mathbf{S}_\lambda^{-2}\boldsymbol{\beta}^* = \frac{\lambda^2}{N}\|\mathbf{S}^{1/2}\mathbf{S}_\lambda^{-1}\boldsymbol{\beta}^*\|_2^2 \leq \frac{\lambda^2}{N}\|\mathbf{S}^{1/2}\mathbf{S}_\lambda^{-1}\|_{\mathrm{op}}^2\|\boldsymbol{\beta}^*\|_2^2 \leq \frac{\lambda}{4N}\|\boldsymbol{\beta}^*\|_2^2,$$
where we used (33) for the last upper bound.

- **Term 2.** Denoting by $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm, the second term admits the upper bound
$$\begin{aligned}\frac{\sigma^2}{N}\operatorname{tr}(\mathbf{S}^2\mathbf{S}_\lambda^{-2}) &= \frac{\sigma^2}{N}\operatorname{tr}(\mathbf{S}\mathbf{S}_\lambda^{-1}\mathbf{S}_\lambda^{-1}\mathbf{S}) = \frac{\sigma^2}{N}\|\mathbf{S}_\lambda^{-1}\mathbf{S}\|_{\mathrm{F}}^2 \\ &\leq \frac{\sigma^2}{N}\|\mathbf{S}_\lambda^{-1}\mathbf{S}^{1/2}\|_{\mathrm{op}}^2\|\mathbf{S}^{1/2}\|_{\mathrm{F}}^2 \leq \frac{\sigma^2}{4\lambda N}\|\mathbf{S}^{1/2}\|_{\mathrm{F}}^2 = \frac{\sigma^2}{4\lambda N}\operatorname{tr}(\mathbf{S}),\end{aligned}$$
where we first used that for linear operators $\mathbf{A}, \mathbf{B} : \mathbb{R}^p \to \mathbb{R}^p$, it holds that $\|\mathbf{A}\mathbf{B}\|_{\mathrm{F}} \leq \|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{B}\|_{\mathrm{F}}$, and then used the bound obtained in (33).

Combining both upper bounds yields
$$\mathbb{E}\mathcal{R}(\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}) - \mathcal{R}^* \leq \frac{\lambda}{4N}\|\boldsymbol{\beta}^*\|_2^2 + \frac{\sigma^2}{4\lambda N}\operatorname{tr}(\mathbf{S}).$$
The r.h.s. attains its minimum at $\lambda^2 = \sigma^2\operatorname{tr}(\mathbf{S})/\|\boldsymbol{\beta}^*\|_2^2$, which is the stated result. □

To compare the risks of both estimators, we recall from Theorem 5.3.1 that
$$\mathbb{E}\mathcal{R}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}) - \mathcal{R}^* = O(N^{-1}).$$
Appropriately normalizing $\mathbf{X}^\top\mathbf{X}$, we have by Theorem 5.5.2 that
$$\mathbb{E}\mathcal{R}(\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}) - \mathcal{R}^* \leq \frac{\sigma\sqrt{\operatorname{tr}(\mathbf{S})}\|\boldsymbol{\beta}^*\|_2}{2N} = \frac{\sigma\sqrt{\operatorname{tr}(\frac{1}{N}\mathbf{X}^\top\mathbf{X})}\|\boldsymbol{\beta}^*\|_2}{2\sqrt{N}} = O(N^{-1/2}).$$
Comparing both rates suggests that the risk of ridge regression converges slower than that of ordinary least squares. Informally speaking, the reduction in MSE comes at the cost of requiring more samples.[7]

5.6 Notes

Our presentation of least squares and ridge regression essentially follows Hastie et al. [2009], where the Gauss-Markov theorem in the form presented here is stated as Ex. 3.3(b). A helpful resource for computing matrix derivatives is Minka [2000]. The formulation of SVD stated in this section is a part of Laub [2005, Theorem 5.1].

Our overview of empirical risk minimization and the fixed design analysis of ordinary least squares and ridge regression are taken from Bach [2024], which also details their random design analysis, among many other topics of learning theory. Notice that there the ridge regression optimization problem is formulated with an additional factor of N^{-1}, yielding slightly different results.

[7] Note that (i) we have not shown that this choice of λ^* implies a reduction of MSE and (ii) we do not know if the $\sqrt{N}$-rate is optimal.

6 High-dimensional data

The data one wants to analyze can often be represented as elements of $\mathbb{R}^d$, with d large. Examples include images, gene data, or long time series. Interestingly, the geometry of high-dimensional Euclidean space usually does not align with our intuition. We illustrate this statement by means of two examples in the following.

6.1 Peculiarities

Our first example concerns the amount of "room" available as the dimensionality increases.

Indeed, consider a unit cube in d dimensions and suppose that we are interested in a region within the cube that captures a a given fraction of the data. For simplicity, we assume that the data, consisting of N samples, is uniformly distributed within the unit cube. Let $r \in [0, 1]$. For $d = 1$, a region of length r will contain Nr samples in expectation. For $d = 2$, a region of size r^2 is expected to capture Nr^2 of the data. In the general case, a cube of side length r captures Nr^d of the data in expectation. Hence, the region required to cover a given number of samples increases *exponentially* in d. To make this concrete, suppose there are $N = 1000$ samples uniformly distributed in the hypercube in $d = 20$ dimensions. A region of side length $r = 1/2$ is expected to capture only $N2^{-d} \approx 0.095\%$ of the data. This abundance of space is one instance of the *curse of dimensionality*.

As a second example, let us now consider the position of random vectors in $\mathbb{R}^d$. Then, a key observation is that random vectors with d independent almost surely bounded components stay at a fixed distance $\sqrt{d}$ from the origin with high probability.[8] In other words, most realizations of such a random vector lie in an ϵ-neighborhood of the surface of a ball with radius $\sqrt{d}$. The following result makes this precise.

Theorem 6.1.1 (Norm concentration)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = 1$ ($i \in \mathbb{N}_{>0}$), and $\sup_i |X_i| \leq a$ a.s. for some $a > 0$. Denote $X = (X_1, \ldots, X_d)$ for some $d \in \mathbb{N}_{>0}$. Then, for any $\epsilon > 0$ and given that $d \geq \epsilon^2$, it holds that*

$$P\Big(\Big|\|X\|_2 - \sqrt{d}\Big| \geq \epsilon\Big) \leq 2\exp(-\epsilon^2/2a^2).$$

Proof. Notice that for any $b \geq 0$, if

$$|b - 1| \geq \epsilon, \text{ then } |b^2 - 1| \geq \max(\epsilon, \epsilon^2). \tag{34}$$

[8] In fact. the result holds more generally for random vectors with independent sub-Gaussian components.

Hence, we have

$$P\Big(\Big|\|X\|_2 - \sqrt{d}\Big| \geq \epsilon\Big) = P\bigg(\bigg|\frac{1}{\sqrt{d}}\|X\|_2 - 1\bigg| \geq \frac{\epsilon}{\sqrt{d}}\bigg)$$
$$\overset{(34)}{\leq} P\bigg(\bigg|\frac{1}{d}\|X\|_2^2 - 1\bigg| \geq \max\bigg(\frac{\epsilon}{\sqrt{d}}, \frac{\epsilon^2}{d}\bigg)\bigg) = P\bigg(\bigg|\frac{1}{d}\|X\|_2^2 - 1\bigg| \geq \frac{\epsilon}{\sqrt{d}}\bigg),$$

by using in the last equality that $d \geq \epsilon^2$ was assumed. Now, as $\mathbb{E}\|X\|_2^2 = \mathbb{E}\sum_{i=1}^d X_i^2 = d$, Hoeffding's inequality (Theorem 4.2.2) with $t = \epsilon d^{1/2}$ yields the stated claim. □

One way of handling high-dimensional data is by reducing their dimensionality.

6.2 Principal component analysis

Given centered data $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$, the goal of dimensionality reduction is finding a map $f : \mathbb{R}^D \to \mathbb{R}^d$, with $d \ll D$, such that $f(\mathbf{x}_i)$ (approximately) preserves some properties of interest. The most well-known dimensionality reduction method is principal component analysis (PCA), where f is a linear map, that is, $\mathbf{x} \mapsto f(\mathbf{x})$ takes the form $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ for some $\mathbf{A} \in \mathbb{R}^{d\times D}$, and the goal is to preserve the d directions in which the data exhibits the largest variance.

In other words, one wants to project $\mathbf{x}_1, \ldots, \mathbf{x}_N$ onto some orthogonal basis $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^D$ such that the maximum amount of variance is preserved. We will determine the $\mathbf{v}_i$-s one-by-one.

Indeed, consider any $\mathbf{v} \in \mathcal{S}^{D-1} = \{\mathbf{x} \in \mathbb{R}^D : \|\mathbf{x}\|_2 = 1\}$ and suppose the $\mathbf{x}_i$-s are realizations of a mean-zero random variable $X \in \mathbb{R}^D$. Then $\mathrm{Var}(\langle \mathbf{v}, X\rangle)$ quantifies the variance of X in the direction of $\mathbf{v}$. Hence, the goal is to find

$$\arg\max_{\mathbf{v}\in\mathcal{S}^{D-1}} \mathrm{Var}(\langle \mathbf{v}, X\rangle) = \arg\max_{\mathbf{v}\in\mathcal{S}^{D-1}} \mathbb{E}(\langle \mathbf{v}, X\rangle)^2. \tag{35}$$

Before we state the (approximate) solution to (35), we collect a property of the operator norm of self-adjoint operators.

Lemma 6.2.1. *Let $\mathbf{A} \in \mathbb{R}^{D\times D}$ be self-adjoint with spectral decomposition $\mathbf{A} = \sum_{i=1}^D \lambda_i \mathbf{s}_i \mathbf{s}_i^\top$ and the λ_i-s in non-increasing order. Then $\|\mathbf{A}\|_{op} = \lambda_1$.*

Proof. By the definition of the operator norm and the monotonicity of $x \mapsto x^2$ for $x \geq 0$, we have that

$$\|\mathbf{A}\|_{\mathrm{op}}^2 = \sup_{\mathbf{x}\in\mathcal{S}^{D-1}} \|\mathbf{A}\mathbf{x}\|_2^2 = \sup_{\mathbf{x}\in\mathcal{S}^{D-1}} \sum_{i=1}^D \langle \mathbf{A}\mathbf{x}, \mathbf{s}_i\rangle^2 = \sup_{\mathbf{x}\in\mathcal{S}^{D-1}} \sum_{i=1}^D \lambda_i^2 \langle \mathbf{x}, \mathbf{s}_i\rangle^2$$
$$\leq \lambda_1^2 \sup_{\mathbf{x}\in\mathcal{S}^{D-1}} \sum_{i=1}^D \langle \mathbf{x}, \mathbf{s}_i\rangle^2 = \lambda_1^2,$$

with equality for $\mathbf{x} = \mathbf{s}_1$. Taking the square root yields the claim. □

Coming back to PCA, let $\mathbf{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^\top \in \mathbb{R}^{D\times D}$ be the plug-in estimator of the sample variance. The matrix $\mathbf{\Sigma}$ is self-adjoint and we denote by $\mathbf{s}_1$ the eigenvector associated to its largest eigenvalue. Using these notations, we obtain

$$\begin{aligned}\mathbf{v}_1 &= \arg\max_{\mathbf{v}\in\mathcal{S}^{D-1}} \frac{1}{N}\sum_{i=1}^{N}(\langle \mathbf{v}, \mathbf{x}_i\rangle)^2 = \arg\max_{\mathbf{v}\in\mathcal{S}^{D-1}} \frac{1}{N}\sum_{i=1}^{N} \operatorname{tr}(\mathbf{v}^\top \mathbf{x}_i\mathbf{x}_i^\top \mathbf{v}) \\ &= \arg\max_{\mathbf{v}\in\mathcal{S}^{D-1}} \operatorname{tr}(\mathbf{v}^\top \mathbf{\Sigma}\mathbf{v}) = \arg\max_{\mathbf{v}\in\mathcal{S}^{D-1}} \|\mathbf{\Sigma}^{1/2}\mathbf{v}\|_2^2 = \mathbf{s}_1.\end{aligned}$$

where the equality in the last step holds by the proof of Lemma 6.2.1.

Having determined $\mathbf{v}_1$, we next aim to find a vector $\mathbf{v} \in \mathbb{R}^D$ in the direction of the largest variance not captured by $\mathbf{v}_1$. In other words, we require $\langle \mathbf{v}, \mathbf{v}_1\rangle = \langle \mathbf{v}, \mathbf{s}_1\rangle = 0$ and, arguing similarly to $\mathbf{v}_1$, we get

$$\begin{aligned}\mathbf{v}_2 &= \arg\max_{\substack{\mathbf{v}\in\mathcal{S}^{D-1}\\ \langle \mathbf{v},\mathbf{s}_1\rangle=0}} \|\mathbf{\Sigma}^{1/2}\mathbf{v}\|_2^2 = \arg\max_{\substack{\mathbf{v}\in\mathcal{S}^{D-1}\\ \langle \mathbf{v},\mathbf{s}_1\rangle=0}} \left\|\sum_{i=1}^{D} \lambda_i \mathbf{s}_i\mathbf{s}_i^\top \mathbf{v}\right\|_2^2 \\ &= \arg\max_{\mathbf{v}\in\mathcal{S}^{D-1}} \left\|\sum_{i=2}^{D} \lambda_i \mathbf{s}_i\mathbf{s}_i^\top \mathbf{v}\right\|_2^2 = \mathbf{s}_2,\end{aligned}$$

by using the orthogonality condition to simplify the optimization.

Continuing in this fashion for $\mathbf{v}_i, \ldots, \mathbf{v}_d$ $(d \le D)$, we find that the map $\mathbf{V} = (\mathbf{v}_i)_{i=1}^d \in \mathbb{R}^{d\times D}$ is the one which preserves most of the variance.

Remark 6.2.1. *Our approach can be shown to be equivalent to finding the d orthonormal eigenvectors spanning the subspace that best approximates—in squared distance sense—the given data when linearly projected onto that space.*

6.3 Johnson-Lindenstrauss lemma

The topic of this section is a result about approximate isometries. Let $\epsilon \in (0,1)$ be fixed and $R \subset \mathbb{R}^D$. We call a map $f : \mathbb{R}^D \to \mathbb{R}^d$ an ϵ-isometry on R if

$$(1-\epsilon)\|\mathbf{u}-\mathbf{v}\|_2^2 \le \|f(\mathbf{u}) - f(\mathbf{v})\|_2^2 \le (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|_2^2 \text{ for all } \mathbf{u}, \mathbf{v} \in R.$$

Put differently, f does not distort the geometry of R by more than ϵ.

The Johnson-Lindenstrauss lemma states that a linear map $\mathbf{A} : \mathbb{R}^D \to \mathbb{R}^d$ chosen randomly is an ϵ-isometry with high probability if d is large enough.

Theorem 6.3.1 (Johnson-Lindenstrauss)**.** *Let $A = \frac{1}{\sqrt{d}}(A_{ij})_{i,j=1}^{d,D} \in \mathbb{R}^{d\times D}$ with $A_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ $(N \ge 2)$ are distinct. Then, for any $\epsilon, \delta \in (0,1)$, it holds that*

$$(1-\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \le \|A\mathbf{x}_i - A\mathbf{x}_j\|_2^2 \le (1+\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2,$$

with probability at least $1-\delta$, given that $d > \frac{16}{\delta^2}\log(N/\epsilon)$.

Notice that the choice of d is independent of the ambient dimension D.

Before we embark on the proof of this result, let us recall a few more details on sub-exponential random variables (see also the end of Section 4.2). In particular, slightly extending our notion of sub-exponentiality, we say that a random variable X is (ν, α)-sub-exponential iff its MGF satisfies $\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq e^{\nu^2\lambda^2/2}$ for all $|\lambda| \leq 1/\alpha$.

Remark 6.3.1. *Many equivalent definitions of sub-exponentiality exist. For example, the preceding definition is equivalent to the existence of $c_1, c_2 > 0$ such that $P(|X - \mathbb{E}X| \geq t) \leq c_1 \exp(-c_2 t)$ for all $t > 0$.*

By the Chernoff method, a sum of independent (ν, α)-sub-exponential random variables satisfies the following concentration result.

Theorem 6.3.2. *Let $X_1, \ldots, X_N$ be independent random variables, $\mu_i = \mathbb{E}X_i$, and suppose that the X_i-s are (ν_i, α_i)-sub-exponential for $i = 1, \ldots, N$, respectively. Define*

$$\alpha_* = \max_{i=1,\ldots,N} \alpha_i \qquad \text{and} \qquad \nu_* = \sqrt{\sum_{i=1}^N \nu_i^2}.$$

Then

$$P\left(\left|\frac{1}{N}\sum_{i=1}^N (X_i - \mu_i)\right| \geq t\right) \leq \begin{cases} 2\exp\left(-\frac{Nt^2}{2(\nu_*^2/N)}\right) & \text{for } 0 \leq t \leq \frac{\nu_*^2}{N\alpha_*}, \\ 2\exp\left(-\frac{Nt}{2\alpha_*}\right) & \text{for } t > \frac{\nu_*^2}{N\alpha_*}. \end{cases}$$

Exercise 6.3.1. *Prove Theorem 6.3.2*

Remark 6.3.2. *Notice that the rate of decay of the tails in Theorem 6.3.2 depends on t. For t small, the decay is sub-Gaussian. For t large, the decay is sub-exponential.*

Exercise 6.3.2. *Let $Z, Z_1, \ldots, Z_N \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, i.e., Z and the Z_i-s are independent standard normal random variables. Show that Z^2 is $(2,4)$-sub-exponential and use this result to show that $Y = \sum_{i=1}^N Z_i^2$ is $(2\sqrt{N}, 4)$-sub-exponential. Conclude that*

$$P\left(\left|\frac{1}{N}\sum_{i=1}^N Z_i^2 - 1\right| \geq t\right) \leq 2\exp\left(-\frac{Nt^2}{8}\right) \text{ for all } t \in (0,1).$$

The random variable Y is said to have the chi-squared distribution with N degrees of freedom, denoted by $Y \sim \chi^2_N$.

We are ready to prove the Johnson-Lindenstrauss lemma (Theorem 6.3.1).

Proof. Let $\mathbf{u} \in \mathbb{R}^D \setminus \{\mathbf{0}\}$ be arbitrary. We first obtain a bound on

$$P\left(\frac{\|A\mathbf{u}\|_2^2}{\|\mathbf{u}\|_2^2} \notin [(1-\epsilon), (1+\epsilon)]\right) = P\left(\left|\frac{\|A\mathbf{u}\|_2^2}{\|\mathbf{u}\|_2^2} - 1\right| \leq \epsilon\right),$$

which we then extend to all $\binom{N}{2}$ choices of $\mathbf{u} = \mathbf{x}_i - \mathbf{x}_j$ $(i \neq j)$ by union bound. The details are as follows.

(Step 1.) Denoting by A_i the d rows of A, we have that

$$\frac{\|A\mathbf{u}\|_2^2}{\|\mathbf{u}\|_2^2} = \frac{1}{d}\sum_{i=1}^{d} \langle A_i, \mathbf{u}/\|\mathbf{u}\|_2\rangle_2^2 =: \frac{1}{d}\sum_{i=1}^{d} Y_i^2.$$

Notice that $Y_i = \frac{1}{\|\mathbf{u}\|_2}\sum_{j=1}^{D} A_{ij}u_j$ is standard normal by the independence of the A_i-s. Hence, $d\|A\mathbf{u}\|_2^2/\|\mathbf{u}\|_2^2 \sim \chi_d^2$, and, by the result in Exercise 6.3.2, we have

$$P\left(\left|\frac{\|A\mathbf{u}\|_2^2}{\|\mathbf{u}\|_2^2} - 1\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{d\epsilon^2}{8}\right). \tag{36}$$

(Step 2.) Using that $\binom{N}{2} = \frac{N(N-1)}{2}$, (36) and a union bound imply that

$$P\left(\left|\frac{\|A\mathbf{x}_i - \mathbf{x}_j\|_2^2}{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2} - 1\right| \geq \epsilon \text{ for all } i \neq j\right) \leq N(N-1)\exp\left(-\frac{d\epsilon^2}{8}\right).$$

As $N(N-1) \leq N^2$, the r.h.s. is smaller than δ for $d > \frac{8}{\epsilon^2}\log\frac{N^2}{\delta}$, which proves the claim. □

Exercise 6.3.3 (Achlioptas 2003)**.** *In the setting of the Theorem 6.3.1, let the A_{ij}-s be independent Rademacher random variables and show that a Johnson-Lindenstrauss-type result continues to hold, i.e., establish that $A = \frac{1}{\sqrt{d}}(A_{ij})_{i,j=1}^{d,D}$ with $P(A_{ij} = 1) = P(A_{ij} = -1) = 1/2$ and all A_{ij}-s independent is an ϵ-isometry on $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with high probability.*

6.4 Notes

Many illustrations of the curse of dimensionality exists. The abundance of room is for example illustrated in Steele [2004], Hastie et al. [2009]. Our result on the concentration of random vectors on the sphere is from Vershynin [2018].

For other perspectives as well as more information on PCA, see Hastie et al. [2009] and Shalev-Shwartz and Ben-David [2014, Section 23.1].

The Johnson-Lindenstrauss lemma originates from Johnson and Lindenstrauss [1982]. Matoušek [2008] gives an overview. Our description combines Vershynin [2018] and Wainwright [2019], with the proof taken from the latter.

7 Convex optimization

In this section, we review a few results from convex optimization, paving the way for handling classifiers that leverage optimal separating hyperplanes. In particular, we define convex optimization problems in their primal form, obtain their dual form, and derive sufficient conditions for their solutions. The concepts are exemplified by solving the ridge regression problem in dual form.

Let us start by stating a few definitions and a helpful lemma. In this section, we consider $\mathcal{X} = \mathbb{R}^n$, that is, we limit our definitions to (mostly open subsets of) Euclidean space. Recall that convex functions were defined in Definition 3.7.1. The epigraph of a function $f : \mathbb{R}^n \to \mathbb{R}$ is $\operatorname{epi} f = \{(\mathbf{x}, t) \mid \mathbf{x} \in \operatorname{dom} f,\, f(\mathbf{x}) \leq t\} \subset \mathbb{R}^{n+1}$.[9] Importantly, a function is convex iff. its epigraph is a convex set, defined as follows.

Definition 7.0.1 (Convex set)**.** *Let C be a subset of a real vector space V. The set C is convex if $tx + (1-t)y \in C$ for every $x, y \in C$ and $t \in [0, 1]$.*

Convex sets are closed w.r.t. intersections.

Lemma 7.0.1. *If all C_α with $\alpha \in \mathcal{A}$ are convex, then so is $C = \bigcap_{\alpha \in \mathcal{A}} C_\alpha$.*

Concave functions are closely related to convex functions.

Definition 7.0.2 (Concave function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is concave if $-f$ is convex.*

We are ready to talk about convex optimization problems.

7.1 Standard and dual form

The next definition captures what is meant by an optimization problem being convex.

Definition 7.1.1 (Convex optimization problem)**.** *Let $f_0, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ be convex, $\mathbf{A} \in \mathbb{R}^{p \times n}$, and $\mathbf{b} = (b_1, \ldots, b_p) \in \mathbb{R}^p$. A convex optimization problem in standard form is written as*

$$\begin{aligned} &\textit{minimize } f_0(\mathbf{x}) \\ &\textit{subject to } f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m, \\ &\qquad \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}. \end{aligned} \tag{37}$$

The function f_0 is called the objective function, the point $\mathbf{x} \in \mathbb{R}^n$ optimization variable, f_i ($i = 1, \ldots, m$) inequality constraint functions, and $\mathbf{x} \mapsto \mathbf{A}\mathbf{x} - \mathbf{b}$ equality constraint functions (identified as $\mathbf{x} \mapsto \mathbf{A}_i\mathbf{x} - \mathbf{b} = h_i(\mathbf{x})$; $i = 1, \ldots, p$).

The formulation (37) is also called primal problem, in contrast to the dual problem, which we introduce soon.

[9] For a function $f : \mathbb{R}^n \supset D \to \mathbb{R}$, we use $\operatorname{dom} f$ to denote D and write $f : \mathbb{R}^n \to \mathbb{R}$ (even though the function may not be defined on all of $\mathbb{R}^n$), as in Boyd and Vandenberghe [2004].

Remark 7.1.1. *Using the h_i-s, the equality constraints can equivalently be stated as $h_i(\mathbf{x}) = 0$ $(i = 1, \ldots, p)$. We use both formulations interchangeably in the following.*

Notice that the objective function and the inequality constraint functions are convex while the equality constraint functions are affine functions of $\mathbf{x} \in \mathbb{R}^n$.

Exercise 7.1.1. *Convince yourself that (i) ordinary least squares* (21) *and (ii) ridge regression in the formulations* (26) *and* (28) *are convex optimization problems, respectively.*

Remark 7.1.2. *Linear optimization problems (also called linear programs) are a special case of* (37)*, obtained when each f_i $(i = 0, \ldots, m)$ is affine.*

Let us introduce some terminology. Indeed, let $\mathcal{D}$ denote the intersection of the domains of the f_i-s.[10] A point $\mathbf{x} \in \mathcal{D}$ is called feasible if it satisfies all constraints and (37) is called feasible if at least one such $\mathbf{x} \in \mathcal{D}$ exists. Let $p^* = \inf_{\mathbf{x} \in \mathcal{D}} \{f_0(\mathbf{x}) \mid \mathbf{x} \text{ satisfies all constraints}\}$. If no solution exists, (37) is called infeasible and $p^* = \infty$ (by convention, being an infimum over an empty set). A point $\mathbf{x}^* \in \mathcal{D}$ is an optimal point of (37) if the problem is feasible and $f_0(\mathbf{x}^*) = p^*$.

A few transformations, collected in the following remark, permit bringing convex optimization problems into the standard form or into a "nicer" form. While these change the optimization problem itself, they do not influence its solution.

Remark 7.1.3.

(i) ***Scaling.*** *One may scale the objective functions and the inequality constraint functions by any positive and the equality constraints by any nonzero constant.*

(ii) ***Change of variables.*** *Let $\phi : \mathbb{R}^n \to \mathbb{R}^n$ be invertible with image covering $\mathcal{D}$. One may then (formally) substitute $\mathbf{x}$ in* (37) *with $\phi(\mathbf{z})$. If $\mathbf{z}$ solves this new problem, $\mathbf{x} = \phi(\mathbf{z})$ solves the original problem (and vice versa, that is, a solution to the original problem yields a solution to the new problem).*

(iii) ***Function transformations.*** *Let $\psi_0 : \mathbb{R} \to \mathbb{R}$ be monotonically increasing, $\psi_1, \ldots, \psi_m : \mathbb{R} \to \mathbb{R}$ satisfy $\psi_i(u) \leq 0$ iff. $u \leq 0$, and $\Psi_1, \ldots, \Psi_p : \mathbb{R} \to \mathbb{R}$ satisfy $\Psi_p(u) = 0$ iff. $u = 0$. Then, replacing f_0, $f_1, \ldots, f_m$, and $h_1, \ldots, h_p$ in* (37) *by $\psi_0 \circ f_0$, $\psi_1 \circ f_1, \ldots, \psi_m \circ f_m$, and $\Psi_1 \circ h_1, \ldots, \Psi_p \circ h_p$, respectively, preserves the feasible set and the optimal points. Notice that (i) is a special case hereof.*

(iv) ***Introducing/removing equality constraints.*** *If any of the f_i can be written as $\tilde{f}_i(\phi_i(\mathbf{x}))$, where ϕ_i is an affine transformation, one may replace $f_i(\mathbf{x})$ by $\tilde{f}_i(\mathbf{y}_i)$ and introduce the new equality constraint $\mathbf{y}_i = \phi_i(\mathbf{x})$. Likewise, one may remove equality constraints in the "reverse" fashion.*

[10] The equality constraints are well defined on all of $\mathbb{R}^n$ and thus not taken into account.

(v) ***Slack variables.*** *It holds that $f_i(\mathbf{x}) \leq 0$ iff. there is an $s_i \geq 0$ such that $f_i(\mathbf{x}) + s_i = 0$. We may thus replace every inequality constraint by an equality constraint and a non-negativity constraint on the slack variable s_i.*

(vi) ***Optimizing independently.*** *It holds that $\inf_{\mathbf{x},\mathbf{y}} f(\mathbf{x},\mathbf{y}) = \inf_{\mathbf{x}} \tilde{f}(\mathbf{x})$ with $\tilde{f}(\mathbf{x}) = \inf_{\mathbf{y}} f(\mathbf{x},\mathbf{y})$, which allows to optimize sequentially if the constraint functions are independent of any of $\mathbf{x}$ and $\mathbf{y}$.*

Further manipulations exist. We refer to Boyd and Vandenberghe [2004, Section 4.1.3] for these and for example applications of the ones listed above.

Of course, the goal of optimization is to find the optimal point (or multiple optimal points). One idea to achieve this goal is incorporating all constraints into the objective function.

Definition 7.1.2 (Lagrangian)**.** *Denote by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_p) \in \mathbb{R}^p$. The Lagrangian $L^* : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated to (37) is defined as*

$$L^*(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}).$$

A λ_i (resp. a ν_i) is called the Lagrange multiplier associated to f_i (resp. h_i). $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ are called dual variables or Lagrange multiplier vectors.

In the setting of Definition 7.1.1, L^* is convex in $\mathbf{x}$ if all $\lambda_i \geq 0$. This is evident as scaling a convex function by a nonnegative constant and taking the sum of convex functions give a convex function. As any affine function is convex, the claim follows.

Besides addition and (nonnegative) scaling, even the pointwise supremum of an infinite set of convex functions yields a convex function.

Lemma 7.1.1. *If, for each $\mathbf{y} \in \mathcal{A}$, $f(\mathbf{x},\mathbf{y})$ is convex in $\mathbf{x}$, then*

$$g(\mathbf{x}) = \sup_{\mathbf{y} \in \mathcal{A}} f(\mathbf{x},\mathbf{y})$$

is convex in $\mathbf{x}$.

Proof. The pointwise supremum of functions corresponds to the intersection of their epigraphs. We have

$$\operatorname{epi} g = \bigcap_{\mathbf{y} \in \mathcal{A}} \operatorname{epi} f(\cdot, \mathbf{y}).$$

The result follows from Lemma 7.0.1. □

Similarly, the pointwise infimum of concave functions is a concave function.

Definition 7.1.3 (Dual function)**.** *The (Lagrange) dual function $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is the minimum value of the Lagrangian L^* over $\mathbf{x} \in \mathcal{D}$:*

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x} \in \mathcal{D}} L^*(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

Remark 7.1.4. *As the dual function is the pointwise infimum of a family of affine functions of (λ, ν), it is concave. This claim is independent of whether the original optimization problem is convex.*

Let $\boldsymbol{\lambda} \geq 0$ be a shorthand for $\lambda_i \geq 0$ for all $i = 1, \ldots, m$. For the dual function, we have the following result.

Lemma 7.1.2. *The dual function satisfies $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$ for any $(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in \mathbb{R}^m \times \mathbb{R}^p$ with $\boldsymbol{\lambda} \geq 0$.*

Proof. Suppose that $\tilde{\mathbf{x}}$ is a feasible point of the primal problem. Then

$$\sum_{i=1}^{m} \lambda_i f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{\mathbf{x}}) \leq 0$$

as the first term is non-positive and the second term is zero. Hence, adding $f_0(\tilde{\mathbf{x}})$ to both sides of the inequality, we get

$$L^*(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{\mathbf{x}}) \leq f_0(\tilde{\mathbf{x}}).$$

As the infimum over all $\mathbf{x} \in \mathcal{D}$ is a lower bound of the l.h.s., this implies

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x} \in \mathcal{D}} L^*(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq L^*(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\tilde{\mathbf{x}}),$$

which holds for all feasible points $\tilde{\mathbf{x}}$, yielding the claim. □

By the preceding result, the dual function provides a lower bound on the optimal value of (37), which suggests the question of how tight the bound possibly is and leads to the corresponding dual problem.

$$\begin{aligned} &\text{maximize } g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &\text{subject to } \boldsymbol{\lambda} \geq 0. \end{aligned} \tag{38}$$

Remark 7.1.5. *The dual problem is a convex optimization problem as $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is concave by Remark 7.1.4 and as maximizing g is equivalent to minimizing the convex $-g$.*

The pair $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ is called dual optimal if it is optimal for the dual problem. Let d^* be the optimal value of the dual problem. The optimal duality gap is $p^* - d^*$; the case $p^* = d^*$ is called *strong duality*.

7.2 Karush-Kuhn-Tucker conditions

We are ready to state the main result of this section.

Theorem 7.2.1 (KKT conditions)**.** *In the setting of Definition 7.1.1, assume that $f_0, \dots, f_m$ are differentiable. If $\tilde{\mathbf{x}}$, $\tilde{\boldsymbol{\lambda}}$, and $\tilde{\boldsymbol{\nu}}$ satisfy the Karush-Kuhn-Tucker (KKT) conditions*

$$\begin{aligned} f_i(\tilde{\mathbf{x}}) &\le 0,\ i = 1, \dots, m, \\ h_i(\tilde{\mathbf{x}}) &= 0,\ i = 1, \dots, p, \\ \tilde{\lambda}_i &\ge 0,\ i = 1, \dots, m, \\ \tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) &= 0,\ i = 1, \dots, m, \\ \nabla f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^{p} \tilde{\nu}_i \nabla h_i(\tilde{\mathbf{x}}) &= 0, \end{aligned}$$

then $\tilde{\mathbf{x}}$ and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ are primal and dual optimal, with zero duality gap.

Before proving this result, we recall the first order condition for convexity. See Boyd and Vandenberghe [2004, Section 3.1.3] for a proof of this property.

Lemma 7.2.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then f is convex iff.* $\operatorname{dom} f$ *is convex and*

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \text{ for all } \mathbf{x}, \mathbf{y} \in \operatorname{dom} f.$$

Notice that if for a convex differentiable function f, $\nabla f(\mathbf{x}) = 0$, we immediately have that $f(\mathbf{y}) \ge f(\mathbf{x})$ for all $\mathbf{y} \in \operatorname{dom} f$. Put differently, if the gradient of a convex function is zero at a point $\mathbf{x}$, the point $\mathbf{x}$ is a "global" minimum. We are ready to prove Theorem 7.2.1.

Proof. By the discussion below Definition 7.1.2, $L^*(\mathbf{x}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ is convex in $\mathbf{x}$. The last KKT condition requires that

$$\nabla_{\mathbf{x}} L^*(\mathbf{x}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})\Big|_{\mathbf{x}=\tilde{\mathbf{x}}} = 0;$$

hence, $\tilde{\mathbf{x}}$ minimizes the Lagrangian by the preceding lemma. This observation, together with the other KKT conditions, implies

$$g(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) = L^*(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) = f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^{m} \tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^{p} \tilde{\nu}_i h_i(\tilde{\mathbf{x}}) = f_0(\tilde{\mathbf{x}}),$$

that is, the duality gap is zero. □

In other words, finding a solution $\tilde{\mathbf{x}}$ and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ satisfying the KKT conditions yields a solution to the primal problem if it is sufficiently regular (that is, has differentiable constraint functions). One may thus (numerically) solve (38) subject to the KKT conditions.

Remark 7.2.1. *Notice that when the KKT conditions are satisfied, $\lambda_i > 0$ requires that $f_i(\tilde{\mathbf{x}}) = 0$. We will later exploit this observation.*

Remark 7.2.2. *A natural question is whether any optimal solution satisfies the KKT conditions. Additionally imposing the so-called Slater's condition permits obtaining the converse in the statement above (necessity) but is omitted.*

We refer to the following section for an application of the KKT conditions.

7.3 Ridge regression dual form

In some cases, the dual problem can be solved analytically. To illustrate the concept, we solve the ridge regression problem (25) using its dual form and the KKT conditions.

Recall from (25) that the ridge regression problem is

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2,$$

for some $\lambda>0$. If $\lambda=0$, we recover the ordinary least squares problem, feasible only if $\mathbf{X}^\top\mathbf{X}$ has full rank.

By Remark 7.1.3(iv), we obtain the equivalent (constrained) problem

$$\begin{aligned}&\text{minimize } \|\mathbf{r}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2,\\ &\text{subject to } \mathbf{y}-\mathbf{X}\boldsymbol{\beta}=\mathbf{r},\end{aligned}$$

which has Lagrangian ($\mathbf{r},\boldsymbol{\nu}\in\mathbb{R}^N$, with N the number of samples)

$$L^*(\mathbf{r},\boldsymbol{\beta},\boldsymbol{\nu})=\|\mathbf{r}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2+\boldsymbol{\nu}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{r}), \tag{39}$$

and corresponding dual problem

$$\text{maximize } g(\boldsymbol{\nu})=\text{maximize } \inf_{\mathbf{r},\boldsymbol{\beta}} L^*(\mathbf{r},\boldsymbol{\beta},\boldsymbol{\nu}).$$

To solve the dual, we first minimize L^* in $\mathbf{r}$ and $\boldsymbol{\beta}$ and then maximize in $\boldsymbol{\nu}$.

(Step 1.) The Lagrangian has partial derivatives

$$\frac{\partial L^*}{\partial \mathbf{r}}=2\mathbf{r}-\boldsymbol{\nu}, \qquad \frac{\partial L^*}{\partial \boldsymbol{\beta}}=2\lambda\boldsymbol{\beta}-\mathbf{X}^\top\boldsymbol{\nu}.$$

Setting the derivatives to zero, we get that $\mathbf{r}=\boldsymbol{\nu}/2$ and $\boldsymbol{\beta}=\mathbf{X}^\top\boldsymbol{\nu}/2\lambda$. Substituting these results into (39) yields

$$\begin{aligned}L^*\left(\frac{\boldsymbol{\nu}}{2},\frac{\mathbf{X}^\top\boldsymbol{\nu}}{2\lambda},\boldsymbol{\nu}\right)&=\frac{1}{4}\|\boldsymbol{\nu}\|_2^2+\frac{1}{4\lambda}\|\mathbf{X}^\top\boldsymbol{\nu}\|_2^2+\boldsymbol{\nu}^\top\left(\mathbf{y}-\frac{1}{2\lambda}\mathbf{X}\mathbf{X}^\top\boldsymbol{\nu}-\frac{1}{2}\boldsymbol{\nu}\right)\\&=-\frac{1}{4}\|\boldsymbol{\nu}\|_2^2-\frac{1}{4\lambda}\|\mathbf{X}^\top\boldsymbol{\nu}\|_2^2+\boldsymbol{\nu}^\top\mathbf{y}.\end{aligned}$$

(Step 2.) Now maximizing in $\boldsymbol{\nu}$, we have

$$\frac{\partial L^*}{\partial \boldsymbol{\nu}}=-\frac{1}{2}\boldsymbol{\nu}-\frac{1}{2\lambda}\mathbf{X}\mathbf{X}^\top\boldsymbol{\nu}+\mathbf{y}=0 \iff \boldsymbol{\nu}=2\lambda(\mathbf{X}\mathbf{X}^\top+\lambda\mathbf{I})^{-1}\mathbf{y}$$

and, as $\boldsymbol{\beta}=\mathbf{X}^\top\boldsymbol{\nu}/2\lambda$, we obtain $\boldsymbol{\beta}=\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top+\lambda\mathbf{I})^{-1}\mathbf{y}$. The solution satisfies the KKT conditions, hence it also solves the primal problem (Theorem 7.2.1). Indeed, by Lemma 5.4.1, $\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top+\lambda\mathbf{I})^{-1}\mathbf{y}=(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$, with the r.h.s. corresponding to the solution of the primal problem, which independently verifies the derivation.

7.4 Notes

Our exposition of convex optimization closely follows Boyd and Vandenberghe [2004]. Rockafellar [1970] gives an account of convex analysis. As elaborated in Remark 7.1.2, linear programs (LPs) are a special case of the convex problems described in this section. The classic method of solving LPs is the simplex algorithm, detailed as literate program by Knuth [2005].

The KKT conditions are a generalization of the method of Lagrange. The latter applies when there are no inequality constraint functions ($m = 0$). Kalman [2009] gives an exposition that differs from the one found in most textbooks.

The dual form of ridge regression is due to Vapnik [2000]. The derivation in Section 7.3 puts the one from Saunders et al. [1998] into vector notation.

8 Classification

Section 5 focused on the regression setting, where the observed y_i-s are real-valued. In this section, we explore the case of $y_i \in \{0, 1\}$ (or $y_i \in \{-1, 1\}$), corresponding to *classification* tasks, already indicated in Section 1. In particular, we construct different classification algorithms using the ERM framework (Section 5.1).

8.1 Nearest neighbors

To set the stage, consider the ERM framework with $\mathcal{Y} = \{0, 1\}$ and the so-called zero-one loss L defined by $(y, \hat{y}) \mapsto \mathbf{1}_{\{y \neq \hat{y}\}}$. Given some predictor $f : \mathcal{X} \to \{0, 1\}$, by conditioning, the expected risk may be written as

$$\mathcal{R}(f) = \mathbb{E}L(Y, f(X)) = \mathbb{E}\mathbb{E}[L(Y, f(X)) \mid X]\,,$$

and the Bayes predictor f^* in (18) takes the form

$$\begin{aligned} f^*(x) &\in \operatorname*{arg\,min}_{y \in \{0,1\}} \mathbb{E}[L(Y, y) \mid X = x] \\ &= \operatorname*{arg\,min}_{y \in \{0,1\}} \{L(0, y)P(Y = 0 \mid X = x) + L(1, y)P(Y = 1 \mid X = x)\} \\ &= \operatorname*{arg\,min}_{y \in \{0,1\}} \{1 - P(Y = y \mid X = x)\} = \operatorname*{arg\,max}_{y \in \{0,1\}} P(Y = y \mid X = x), \quad (40) \end{aligned}$$

where the first equality follows by the properties of conditional expectations and the second one by comparing cases.

The r.h.s. of (40) suggests that a reasonable classifier assigns the most probable label, matching our intuition. Assume that $(\mathcal{X}, d_{\mathcal{X}})$ is a metric space. Considering the case where $x \in \mathcal{X}$ is not in the training set, we may approximate $P(Y = y \mid X = x)$ using the neighborhood of x, with the assumption that observations that are close in feature space should have the same label. Formally, the nearest neighbor classifier employs the approximation

$$P(Y = y \mid X = x) \approx \frac{1}{k} \sum_{x_i \in N_k(x)} y_i,$$

where $N_k(x)$ denotes the k-neighborhood (the k closest points) of x w.r.t. the metric $d_{\mathcal{X}}$ and k is a user-supplied parameter.

Remark 8.1.1. *Notice that a similar argument permits using nearest neighbors for regression problems.*

The decision of the nearest neighbor classifier hinges on k points only and we do not impose any assumption on how the so-called decision function looks like. It follows that this approach has a high variance but a low bias.

Exercise 8.1.1. *How does the choice of k influence the bias and variance of the nearest neighbor classifier?*

8.2 Logistic regression

Recall that we used the squared loss for our regression tasks. The squared loss is a continuously differentiable convex function, leading to nice solutions of the corresponding ERM problems (given that f is nice). The zero-one loss, defined in Section 8.1, is discontinuous and thus not differentiable. Indeed, optimizing the zero-one loss directly is an NP-hard problem.

To alleviate the issue, one uses "nice" approximations to the zero-one loss. This section presents two ways to arrive at the so-called (two-class) logistic regression classifier. The first involves approximating $P(Y \mid X)$ using ideas from linear regression while ensuring that the prediction stays in $[0, 1]$ so that it can be interpreted as a probability. The second, an ERM-based perspective, frames the problem in terms of minimizing a (given) loss function. Both approaches yield the same solution.

8.2.1 Maximum likelihood-based approach

We restrict ourselves to $\mathcal{X} = \mathbb{R}^p$. Following our ideas in Section 8.1, our goal is to predict

$$P(Y = y \mid X = \mathbf{x}) \text{ with } y \in \{0, 1\} = \mathcal{Y}$$

for a given $\mathbf{x} \in \mathbb{R}^p$.

As we consider two classes only ($|\mathcal{Y}| = 2$), it suffices to model $P(Y = 1 \mid X)$; the case $Y = 0$ follows by considering the complement. We could consider the linear model ($\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^p$)

$$P(Y = 1 \mid X) = \beta_0 + \boldsymbol{\beta}^\top X$$

but this allows $P(Y = 1 \mid X) \notin [0, 1]$ for certain choices of X (assuming that $\boldsymbol{\beta} \neq 0$—a reasonable assumption as otherwise the prediction is independent of the input), which is unsatisfactory. To solve this issue, we consider instead a function that takes values in $[0, 1]$, for example,

$$P(Y = 1 \mid X) = \frac{e^{\beta_0 + \boldsymbol{\beta}^\top X}}{1 + e^{\beta_0 + \boldsymbol{\beta}^\top X}},$$

the so-called *logistic function*. By an abuse of notation, we again prefix $\mathbf{x} \in \mathbb{R}^p$ by 1 and let $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, which lets us write

$$P(Y = 1 \mid X = \mathbf{x}) = \frac{e^{\boldsymbol{\beta}^\top \mathbf{x}}}{1 + e^{\boldsymbol{\beta}^\top \mathbf{x}}} =: S_{\boldsymbol{\beta}}(\mathbf{x}),$$

where $S_{\boldsymbol{\beta}} : \mathbb{R}^{p+1} \to [0, 1]$ denotes the logistic function with parameters $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$.

To estimate $P(Y = 1 \mid X = \mathbf{x})$, notice that Y given $X = \mathbf{x}$ is a Bernoulli random variable with parameter $S_{\boldsymbol{\beta}}(\mathbf{x})$ and probability mass function

$$p(y; S_{\boldsymbol{\beta}}(\mathbf{x})) = [S_{\boldsymbol{\beta}}(\mathbf{x})]^y [1 - S_{\boldsymbol{\beta}}(\mathbf{x})]^{1-y} \text{ for } y \in \{0, 1\}.$$

As the $(\mathbf{x}_i, y_i)$-pairs are independent by assumption, we have the likelihood function

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} p(y_i; S_{\boldsymbol{\beta}}(\mathbf{x}_i)) = \prod_{i=1}^{N} [S_{\boldsymbol{\beta}}(\mathbf{x}_i)]^{y_i} [1 - S_{\boldsymbol{\beta}}(\mathbf{x}_i)]^{1-y_i},$$

and the log-likelihood

$$\begin{aligned} l(\boldsymbol{\beta}) = \log \mathcal{L}(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \{y_i \log S_{\boldsymbol{\beta}}(\mathbf{x}_i) + (1 - y_i) \log[1 - S_{\boldsymbol{\beta}}(\mathbf{x}_i)]\} \\ &= \sum_{i=1}^{N} \left\{ y_i \boldsymbol{\beta}^\top \mathbf{x}_i - \log(1 + e^{\boldsymbol{\beta}^\top \mathbf{x}_i}) \right\}. \end{aligned} \tag{41}$$

Exercise 8.2.1. *Show that $l(\boldsymbol{\beta})$ is a concave function.*

To maximize (41), we set its gradient to zero, yielding the equations

$$\nabla l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \{\mathbf{x}_i [y_i - S_{\boldsymbol{\beta}}(\mathbf{x}_i)]\} = \mathbf{0},$$

which are $p + 1$ equations that are non-linear in $\boldsymbol{\beta}$. These can be solved by the Newton-Raphson algorithm (although convergence is not guaranteed).

Remark 8.2.1. *The general formulation of logistic regression applies to the case of $|\mathcal{Y}| \geq 2$ classes. For simplicity, our construction assumes that $\mathcal{Y}$ has two classes only, encoded as* 0 *and* 1.

The next section details an alternative view of logistic regression.

8.2.2 ERM-based approach

Let $\mathcal{Y} = \{-1, 1\}$ be the relabeling $y \mapsto 2y - 1$ of the y_i-s in the previous section (there encoded as 0 and 1). In this section, we write $y_i' = 2y_i - 1$ to make the difference clear. Consider the so-called logistic loss $L : \{-1, 1\} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ defined by

$$(y, x) \mapsto \log(1 + e^{-yx}). \tag{42}$$

Using the loss function (42) with $f : \mathbb{R}^{p+1} \to \mathbb{R}$ defined by $\mathbf{x} \mapsto \boldsymbol{\beta}^\top \mathbf{x}$, the empirical risk (Definition 5.1.3) given our data takes the form

$$\frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-y_i' f(\mathbf{x}_i)}), \tag{43}$$

which we aim to minimize over $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. When taking the different encodings of the y_i-s into account, this minimization problem is equivalent to maximizing the log-likelihood (41).

Indeed, notice that the minimization of (43) is agnostic to normalization. By combining this observation with the encoding $y_i' = 2y_i - 1$ for any $i = 1, \dots, N$, we obtain

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \log(1 + e^{-(2y_i-1)f(\mathbf{x}_i)}) = \max_{\boldsymbol{\beta}} \sum_{i=1}^{N} -\log(1 + e^{-(2y_i-1)f(\mathbf{x}_i)}) = \max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}),$$

where, in the last equation, we use that for every term in the sum

$$\begin{aligned} -\log(1 + e^{-(2y_i-1)f(\mathbf{x}_i)}) &= \begin{cases} -\log(1 + e^{f(\mathbf{x}_i)}) & \text{if } y_i = 0, \\ -\log(1 + e^{-f(\mathbf{x}_i)}) & \text{if } y_i = 1 \end{cases} \\ &= -(1 - y_i)\log(1 + e^{f(\mathbf{x}_i)}) \underbrace{-y_i \log(1 + e^{-f(\mathbf{x}_i)})}_{= y_i f(\mathbf{x}_i) - y_i \log(1 + e^{f(\mathbf{x}_i)})} \\ &= y_i f(\mathbf{x}_i) - \log(1 + e^{f(\mathbf{x}_i)}). \end{aligned}$$

Summing up all those terms from $i = 1$ to N gives back (41) and thereby supports the ERM-based approach.

8.3 Optimal separating (affine) hyperplanes

The goal of this section is to prepare the derivation of the support vector machine, which is a powerful kernel-based classification algorithm. However, to simplify, we do not use kernel functions yet and assume that the data is linearly separable.

Consider $\mathcal{Y} = \{-1, 1\}$ and suppose that the observed $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}$-pairs are linearly separable, that is, we assume that there exists an affine hyperplane separating the space $\mathbb{R}^d$ into two half-spaces, such that all samples $\mathbf{x}_i$ with $y_i = -1$ lie in one half-space and all samples $\mathbf{x}_i$ with $y_i = 1$ lie in the other half space. The goal of this section is to construct such an affine hyperplane that separates the points as good as possible. In other words, we want the hyperplane to have a maximal margin. We may then classify unseen objects depending on the half space to which they belong.

Given a distance $\beta_0 \in \mathbb{R}$ and a normal vector $\boldsymbol{\beta} \in \mathcal{S}^{d-1}$, an affine hyperplane can be defined as the set $M = \{\mathbf{x} \in \mathbb{R}^d \mid \langle \boldsymbol{\beta}, \mathbf{x} \rangle + \beta_0 = 0\}$. Observe that the shortest signed distance of any point $\mathbf{x} \in \mathbb{R}^d$ to M is $\langle \boldsymbol{\beta}, (\mathbf{x} - \mathbf{x}_0) \rangle$ with $\mathbf{x}_0 = -\beta_0 \boldsymbol{\beta} \in M$.

Exercise 8.3.1. *(i) Show that* $\{\mathbf{x} - \mathbf{x}_0 \mid \mathbf{x} \in \mathbb{R}^d, \langle \boldsymbol{\beta}, \mathbf{x} \rangle = 0\} = M$. *(ii) Show that* $\langle \boldsymbol{\beta}, (\mathbf{x} - \mathbf{x}_0) \rangle = \langle \boldsymbol{\beta}, \mathbf{x} \rangle + \beta_0$.

By the previous observations and the preceding exercise, $f(\mathbf{x}) = \beta_0 + \langle \boldsymbol{\beta}, \mathbf{x} \rangle$ is an affine function proportional to the signed distance of $\mathbf{x}$ to the hyperplane defined by $f(\mathbf{x}) = 0$. Consequently, a solution to the optimization problem

$$\begin{aligned} &\max_{\beta_0, \boldsymbol{\beta}} M \\ &\text{subject to } y_i(\beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle) \geq M, \ i = 1, \dots, N, \\ &\qquad \|\boldsymbol{\beta}\|_2 = 1 \end{aligned} \tag{44}$$

yields a separating hyperplane with maximal margin.

Exercise 8.3.2. *Show that $f(\mathbf{x}_i) < 0$ if a sample $(\mathbf{x}_i, y_i)$ with $y_i = 1$ is misclassified (resp. $f(\mathbf{x}_i) > 0$ if a sample $(\mathbf{x}_i, y_i)$ with $y_i = -1$ is misclassified).*

We now transform (44) to the standard form of a convex optimization problem. Indeed, incorporating the $\|\boldsymbol{\beta}\| = 1$ constraint and rescaling β_0 accordingly, we have the equivalent set of constraints ($i = 1, \ldots, N$)

$$\frac{y_i}{\|\boldsymbol{\beta}\|_2}(\beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle) \geq M.$$

Setting $\|\boldsymbol{\beta}\|_2 = 1/M$, we obtain the equivalent

$$\begin{aligned} &\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 \\ &\text{subject to } y_i(\beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle) \geq 1,\ i = 1, \ldots, N, \end{aligned} \tag{45}$$

seen to be a convex optimization problem.

The Lagrangian corresponding to (45) is

$$L^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 - \sum_{i=1}^{N} \lambda_i [y_i(\beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle) - 1],$$

and setting its derivatives w.r.t. β_0 and $\boldsymbol{\beta}$ to 0, respectively, gives

$$\frac{\partial L^*}{\partial \beta_0} = \sum_{i=1}^{N} \lambda_i y_i = 0 \qquad \text{and} \qquad \frac{\partial L^*}{\partial \boldsymbol{\beta}} = \boldsymbol{\beta} - \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i = 0. \tag{46}$$

Combining (46) with L^* allows us to obtain the dual problem

$$\begin{aligned} &\text{maximize } \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ &\text{subject to } \boldsymbol{\lambda} \geq 0, \\ &\qquad \sum_{i=1}^{N} \lambda_i y_i = 0. \end{aligned}$$

Remark 8.3.1. *As the constraints of the optimization problem are affine, the refined Slater's condition [Boyd and Vandenberghe, 2004, (5.27)] is satisfied and the converse to the KKT conditions holds, that is, the equations stated in Theorem 7.2.1 must be satisfied at an optimal point.*

In particular, for all $i = 1, \ldots, N$, it holds that

$$\lambda_i [y_i(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle + \beta_0) - 1] = 0,$$

which implies that $y_i(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle + \beta_0) = 1$ if $\lambda_i > 0$, and that $\lambda_i = 0$ if $\lambda_i(\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle + \beta_0) > 0$ (see Remark 7.2.1). Put differently, the Lagrange multipliers are non-zero only for points on the boundary of the "slab". Those $\mathbf{x}_i$-s are called support points.

To classify a new observation $\mathbf{x} \in \mathbb{R}^p$, we consider the half-space in which $\mathbf{x}$ lies, indicated by the sign of $f(\mathbf{x})$ (see Exercise 8.3.2). Formally, the corresponding prediction is

$$\hat{y} = \operatorname{sign} \hat{f}(\mathbf{x}) = \operatorname{sign}(\hat{\beta}_0 + \langle \hat{\boldsymbol{\beta}}, \mathbf{x} \rangle),$$

where, for $x \in \mathbb{R}$,

$$\operatorname{sign} x = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

8.4 Notes

Hastie et al. [2009] has additional information on all the contents of this section, except for the ERM-based perspective of logistic regression. However, it is well-known that using the logistic loss function in ERM leads to the same solution as the maximum likelihood-based approach, as shown in Section 8.2.2.

More details on maximal margin classifiers and their optimization can be found in Schölkopf and Smola [2002], Steinwart and Christmann [2008].

One important topic that is missing from this section are tree-based classifiers. In particular gradient boosted trees and random forests typically show very good results in practice; Hastie et al. [2009] gives an introduction.

9 Kernel methods

Ridge regression and the optimal affine hyperplane classifier suffer from strong assumptions. In ridge regression, we assumed that the relationship between the features and the output is affine. For the optimal affine hyperplane classifier, we assumed that the data is linearly separable.

In this section, we first weaken those assumptions using basis expansions. Here, the underlying idea is to map the data to a space in which the assumptions are satisfied, and to apply the respective algorithm in this space. The dual formulations of ridge regression and the optimal affine hyperplane classifier permit performing these calculations implicitly, using a *kernel function*, which is called the *kernel trick*. We close the section by stating a few key results of the rich theory surrounding kernel functions.

9.1 Kernel ridge regression

In Section 5.2, we assumed the model to be of the form (20). An alternative in the same setting, that is, $X = (X_1, \dots, X_p) \in \mathbb{R}^p$ and $Y \in \mathbb{R}$, is considering the model

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X), \tag{47}$$

with $M \in \mathbb{N}$, $\beta_m \in \mathbb{R}$, and $h_m : \mathbb{R}^p \to \mathbb{R}$ $(m = 1, \dots, M)$ adequately chosen. For example, typical choices include (i) $M = p$ and $h_m(X) = X_m$ for $m = 1, \dots, p$, which recovers (20) with $\beta_0 = 0$, (ii) $h_m(X) = X_j^2$, or (iii) $h_m(X) = X_i X_j$; the latter allows taking interactions between features into account. Similarly to (ii), one can consider other non-linear transformations, like the square root or the logarithm.

Notice that (47) is still linear in its parameters $\beta_1, \dots, \beta_M$.

Remark 9.1.1. *Technically, we do not require that $X \in \mathbb{R}^p$, that is, X is not required to be Euclidean. Assuming we are provided with $h_m : \mathcal{X} \to \mathbb{R}$ and that $X \in \mathcal{X}$, we can compute* (47)*. Still, we state the case of $X \in \mathbb{R}^p$ to highlight the similarity to the original linear model.*

For reasons that will become clear later, we collect the h_m-s using the *feature map*

$$\phi : \mathbb{R}^p \to \mathbb{R}^M, \qquad \mathbf{x} \mapsto (h_m(\mathbf{x}))_{m=1}^M, \tag{48}$$

let $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be defined by $(\mathbf{x}, \mathbf{y}) \mapsto \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$, and write $\mathbf{K} = [k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^N \in \mathbb{R}^{N \times N}$.

Exercise 9.1.1. *Show that* $\mathbf{K}$ *is positive definite.*

Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_M) \in \mathbb{R}^M$. Using these notations, we write the ridge regression problem as

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^M} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2, \tag{49}$$

where now $\mathbf{X} = [\phi(\mathbf{x}_i)]_{i=1}^N \in \mathbb{R}^{N\times M}$. Following similar steps as in Section 7.3 shows that $\boldsymbol{\beta} = \mathbf{X}^\top(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{y}$ by using that $\mathbf{X}\mathbf{X}^\top = \mathbf{K}$. Hence, the prediction of an unseen $\mathbf{x} \in \mathbb{R}^p$ takes the form

$$\begin{aligned}\hat{f}(\mathbf{x}) &= \langle \boldsymbol{\beta}, \phi(\mathbf{x})\rangle = \langle \mathbf{X}^\top(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{y}, \phi(\mathbf{x})\rangle = \langle (\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{y}, \mathbf{X}\phi(\mathbf{x})\rangle \\ &= \mathbf{y}^\top(\mathbf{K}+\lambda\mathbf{I})^{-1}\begin{pmatrix} k(\mathbf{x}_1,\mathbf{x}) \\ \vdots \\ k(\mathbf{x}_N,\mathbf{x})\end{pmatrix}, \end{aligned} \tag{50}$$

where we used that

$$\mathbf{X}\phi(\mathbf{x}) = \begin{pmatrix} \langle\phi(\mathbf{x}_1),\phi(\mathbf{x})\rangle \\ \vdots \\ \langle\phi(\mathbf{x}_N),\phi(\mathbf{x})\rangle\end{pmatrix} = \begin{pmatrix} k(\mathbf{x}_1,\mathbf{x}) \\ \vdots \\ k(\mathbf{x}_N,\mathbf{x}),\end{pmatrix}$$

and that $(\mathbf{K}+\lambda\mathbf{I})^{-1}$ is self-adjoint.

Notice that $\phi(\cdot)$ never occurs explicitly in (50) but only implicitly (through k). We may use this observation to allow infinite-dimensional feature maps.

Indeed, let $\mathcal{X}$ be a set, $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ a (separable) Hilbert space, $\phi : \mathcal{X} \to \mathcal{H}$ a feature map, $k(x,y) = \langle\phi(x),\phi(y)\rangle_{\mathcal{H}}$, and denote the sampling operator by $S_N : \mathcal{H} \to \mathbb{R}^N$, which acts as $h \mapsto (\langle h, \phi(x_i)\rangle)_{i=1}^N$.

Exercise 9.1.2. *Show that (i) the sampling operator has the adjoint $S_N^* : \mathbb{R}^N \to \mathcal{H}$, $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_N) \mapsto \sum_{i=1}^N \alpha_i\phi(x_i)$ and (ii) $S_N S_N^* = \mathbf{K}$.*

Consider the ERM problem

$$\inf_{f\in\mathcal{H}} \|\mathbf{y} - S_N f\|_2^2 + \lambda\|f\|_{\mathcal{H}}^2, \tag{51}$$

which recovers (49) if ϕ is chosen as in (48) (then $\mathcal{H} = \mathbb{R}^M$).

We will now show, in a more general setting, that the $f \in \mathcal{H}$ solving (51)—although (51) is an optimization problem over a potentially infinite-dimensional Hilbert space—lies in the finite-dimensional subspace $\mathcal{H}_N = \operatorname{span}\{\phi(x_i) \mid i = 1,\dots,N\}$.

Theorem 9.1.1 (Representer theorem)**.** *Let $((x_i,y_i))_{i=1}^N \in \mathcal{X}\times\mathbb{R}$, $L : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ a Hilbert space, $\phi : \mathcal{X} \to \mathcal{H}$, and $\lambda > 0$. Then*

$$\inf_{f\in\mathcal{H}} \sum_{i=1}^N L(y_i, \langle f, \phi(x_i)\rangle_{\mathcal{H}}) + \lambda\|f\|_{\mathcal{H}}^2$$

is attained at $f \in \mathcal{H}_N = \operatorname{span}\{\phi(x_i) \mid i = 1,\dots,N\}$.

Proof. Write $f = f_N + f_N^\perp$, with $f_N \in \mathcal{H}_N$ and $f_N^\perp \in \mathcal{H}_N^\perp$, the latter denoting

the orthogonal complement of $\mathcal{H}_N$ in $\mathcal{H}$. Then

$$\begin{aligned}\sum_{i=1}^{N} L(y_i, \langle f, \phi(x_i)\rangle) + \lambda\|f\|_{\mathcal{H}}^2 &= \sum_{i=1}^{N} L(y_i, \langle f_N, \phi(x_i)\rangle) + \lambda(\|f_N\|_{\mathcal{H}}^2 + \|f_N^{\perp}\|_{\mathcal{H}}^2) \\ &\geq \sum_{i=1}^{N} L(y_i, \langle f_N, \phi(x_i)\rangle) + \lambda\|f_N\|_{\mathcal{H}}^2,\end{aligned}$$

where the first equality holds by the orthogonality of $f_N^{\perp}$ and $\phi(x_i)$ $(i = 1, \ldots, N)$, and the Pythagorean theorem. □

Using this result in the setting of (51), we obtain

$$\inf_{f\in\mathcal{H}} \|\mathbf{y}-S_N f\|_2^2+\lambda\|f\|_{\mathcal{H}}^2 = \inf_{f\in\mathcal{H}_N} \|\mathbf{y}-S_N f\|_2^2+\lambda\|f\|_{\mathcal{H}}^2 = \inf_{\boldsymbol{\beta}} \|\mathbf{y}-\mathbf{K}\boldsymbol{\beta}\|_2^2+\lambda\boldsymbol{\beta}^\top\mathbf{K}\boldsymbol{\beta},$$

which has derivative

$$\frac{\partial}{\partial\boldsymbol{\beta}}\|\mathbf{y} - \mathbf{K}\boldsymbol{\beta}\|_2^2 + \lambda\boldsymbol{\beta}^\top\mathbf{K}\boldsymbol{\beta} = 2\mathbf{K}(\mathbf{K}\boldsymbol{\beta} - \mathbf{y} + \lambda\boldsymbol{\beta}).$$

Setting the derivative to zero shows that an optimum occurs at $\boldsymbol{\beta} = (\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{y}$. If $\phi(\mathbf{x}) = \mathbf{x} \in \mathbb{R}^p$, it holds that $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$ and we recover the dual solution of the ridge regression problem in Section 7.3.

To obtain a prediction for an unseen observation $x \in \mathcal{X}$, we have

$$\hat{f}(x) = \langle f, \phi(x)\rangle_{\mathcal{H}} = \langle S_N^*\boldsymbol{\beta}, \phi(x)\rangle_{\mathcal{H}} = \langle\boldsymbol{\beta}, S_N\phi(x)\rangle_2 = \sum_{i=1}^{N}\beta_i k(x, x_i),$$

which shows that each prediction is a weighted linear combination of the kernel function evaluated at the training points and the new observation. This result matches (50).

Remark 9.1.2. *We have seen how we may incorporate a (given) feature map into the ridge regression problem, which is helpful for building intuition. In practice, one instead typically chooses the kernel function such that the feature map associated to that function fits the problem at hand. Put differently, in the regression setting one rarely works directly with feature maps.*[11] *Section 9.3 provides more details on kernels and their feature maps.*

9.2 Support vector machine

In this section, we first extend the optimal separating hyperplane classifier (Section 8.3) to allow for misclassifications by introducing *slack variables*. We then use basis expansions, as in the previous section, to obtain non-linear decision boundaries, leading to the support vector machine (SVM). We close this section by pointing out how the SVM arises from an ERM perspective. The details are as follows.

[11] The same holds for classification, tackled in the next section.

9.2.1 Slack variables

Recall from (44) that the optimization problem for the optimal hyperplane is

$$\begin{aligned} &\max_{\beta_0,\boldsymbol{\beta}} M \\ &\text{subject to } y_i(\beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle) \geq M,\ i = 1, \ldots, N, \\ &\qquad \|\boldsymbol{\beta}\|_2 = 1. \end{aligned}$$

If the data is not linearly separable, (44) has no solution. The idea to tackle overlapping classes is to introduce the slack variables $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N) \in \mathbb{R}^N$ and to consider the optimization problem

$$\begin{aligned} &\max_{\beta_0,\boldsymbol{\beta}} M \\ &\text{subject to } y_i(\beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle) \geq M(1 - \xi_i),\ i = 1, \ldots, N, \\ &\qquad \boldsymbol{\xi} \geq \mathbf{0}, \\ &\qquad \mathbf{1}^\top \boldsymbol{\xi} \leq c, \\ &\qquad \|\boldsymbol{\beta}\|_2 = 1, \end{aligned} \tag{52}$$

where $c \geq 0$ is a user-specified constant, controlling the amount of allowed misclassifications.

Remark 9.2.1. *An alternative to* (52) *is introducing the constraints* $y_i(\beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle) \geq M - \xi_i$ *for all* $i = 1, \ldots, N$. *The choice* (52) *leads to a convex optimization problem, while the alternative results in a non-convex optimization problem. The former measures the overlap in relative distance from the margin; the latter measures the absolute distance from the margin.*

Arguing as for (44), that is, normalizing $\boldsymbol{\beta}$ by $\|\boldsymbol{\beta}\|_2$ (and adjusting β_0 and M) and setting $\|\boldsymbol{\beta}\|_2 = 1/M$, gives the optimization problem

$$\begin{aligned} &\min_{\beta_0,\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2 \\ &\text{subject to } y_i(\beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle) \geq 1 - \xi_i,\ i = 1, \ldots, N, \\ &\qquad \boldsymbol{\xi} \geq \mathbf{0}, \\ &\qquad \mathbf{1}^\top \boldsymbol{\xi} \leq c. \end{aligned}$$

This problem can equivalently be expressed as ($C \geq 0$)

$$\begin{aligned} &\min_{\beta_0,\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + C\mathbf{1}^\top \boldsymbol{\xi} \\ &\text{subject to } y_i(\beta_0 + \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle) \geq 1 - \xi_i,\ i = 1, \ldots, N, \\ &\qquad \boldsymbol{\xi} \geq \mathbf{0}, \end{aligned} \tag{53}$$

which has Lagrangian[12]

$$L^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + C\mathbf{1}^\top\boldsymbol{\xi} - \sum_{i=1}^{N} \lambda_i \left[y_i(\beta_0 + \langle\boldsymbol{\beta}, \mathbf{x}_i\rangle) - (1 - \xi_i)\right] - \boldsymbol{\mu}^\top\boldsymbol{\xi}.$$

Setting the derivatives w.r.t. β_0, $\boldsymbol{\beta}$, and $\boldsymbol{\xi}$ to zero, respectively, gives

$$\boldsymbol{\beta} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i, \qquad 0 = \boldsymbol{\lambda}^\top\mathbf{y}, \qquad \text{and} \qquad \boldsymbol{\lambda} = C\mathbf{1} - \boldsymbol{\mu}, \tag{54}$$

resulting in the dual problem

$$\begin{aligned} &\max \mathbf{1}^\top\boldsymbol{\lambda} - \frac{1}{2}\sum_{i,j=1}^{N} \lambda_i\lambda_j y_i y_j \langle\mathbf{x}_i, \mathbf{x}_j\rangle \\ &\text{subject to } 0 \le \boldsymbol{\lambda} \le C, \\ &\qquad \boldsymbol{\lambda}^\top\mathbf{y} = 0. \end{aligned} \tag{55}$$

In addition to the equations in (54), the KKT conditions include the constraints

$$\lambda_i \left[y_i(\langle\mathbf{x}_i, \boldsymbol{\beta}\rangle + \beta_0) - (1 - \xi_i)\right] = 0, \tag{56}$$

$$\mu_i\xi_i = 0, \tag{57}$$

$$y_i(\langle\mathbf{x}_i, \boldsymbol{\beta}\rangle + \beta_0) - (1 - \xi_i) \ge 0,$$

for $i = 1, \ldots, N$, which uniquely characterize the solution.

As for the separating hyperplane classifier, we call the observations $\mathbf{x}_i$ for which $\lambda_i > 0$ *support vectors*; by (54), these determine $\boldsymbol{\beta}$. Among the support vectors, some will have $\xi_i = 0$, which by (56) implies that they are on the margin. Further, by the last equation in (54), they satisfy $0 < \lambda_i < C$. If, instead, the support vectors satisfy $\xi_i > 0$, (57) implies that $\mu_i = 0$. Hence, $\lambda_i = C$ by the last equation in (54). The value of β_0 can be obtained from any support vector (or all) by using (56).

Having obtained $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$, the decision function takes the form

$$\hat{y} = \operatorname{sign} \hat{f}(\mathbf{x}) = \operatorname{sign}(\hat{\beta}_0 + \langle\hat{\boldsymbol{\beta}}, \mathbf{x}\rangle).$$

9.2.2 Basis expansion

The solution of the previous section allows us to handle non-linearly separable data but still results in an affine hyperplane, that is, a linear decision boundary. As for ridge regression in (48), let us replace each $\mathbf{x}_i$ by a feature vector $\phi(\mathbf{x}_i)$ and let k and $\mathbf{K}$ be as defined there. The objective function of the dual problem (55) then takes the form

$$\max\ \mathbf{1}^\top\boldsymbol{\lambda} - \frac{1}{2}\sum_{i,j=1}^{N} \lambda_i\lambda_j y_i y_j \underbrace{\langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle}_{=k(\mathbf{x}_i, \mathbf{x}_j)};$$

[12] Introducing the Lagrange multiplier vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ (instead of a single one) simplifies referring to the respective constraints.

it follows as in (54) that $\boldsymbol{\beta} = \sum_{i=1}^N \lambda_i y_i \phi(\mathbf{x}_i)$. The prediction involves

$$\hat{f}(\mathbf{x}) = \langle \phi(\mathbf{x}), \beta \rangle + \beta_0 = \sum_{i=1}^N \lambda_i y_i \underbrace{\langle \phi(\mathbf{x}), \phi(\mathbf{x}_i) \rangle}_{=k(\mathbf{x}, \mathbf{x}_i)} + \beta_0.$$

To obtain β_0, one may solve $y_i \hat{f}(\mathbf{x}_i) = 1$ for any $\mathbf{x}_i$ for which $0 < \lambda_i < C$.

9.2.3 ERM-based approach

It remains to fit SVM into our ERM-based perspective on learning. Indeed, in the setting of Theorem 9.1.1, let $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $(y, \hat{y}) \mapsto \max(1 - y\hat{y}, 0) =: [1 - y\hat{y}]_+$ denote the *hinge loss* and consider the ERM problem

$$\min_{\beta_0 \in \mathbb{R}, h \in \mathcal{H}} \sum_{i=1}^N [1 - y_i(\beta_0 + \langle h, \phi(x_i) \rangle_{\mathcal{H}})]_+ + \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 \tag{58}$$

with $y_i \in \{-1, 1\}$. By optimizing separately in β_0 and h (Remark 7.1.3(vi)), Theorem 9.1.1 yields that $h(x) = \sum_{i=1}^N \beta_i k(x, x_i)$. Then, $f(x) = \beta_0 + h(x)$ and we have the optimization problem

$$\min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^N} \sum_{i=1}^N [1 - y_i f(x_i)]_+ + \frac{\lambda}{2} \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta}.$$

Exercise 9.2.1. *Show that the constrained optimization problem* (53) *is equivalent to the ERM problem* (58) *with* $\phi(\mathbf{x}) = \mathbf{x}$.

9.3 Kernel theory

In the last section, without naming them explicitly, we employed so-called kernel functions. We now make explicit what constitutes a kernel function and explore a few of their properties. Key results of this section are the Riesz representation theorem, which allows expressing a linear functional by an inner product, and the Moore-Aronszajn theorem, which relates kernel functions to reproducing kernel Hilbert spaces.

Definition 9.3.1 (Kernel function)**.** *Let $\mathcal{X}$ be a nonempty set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric. If*

$$\sum_{i,j=1}^{N} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

holds for all $x_1, \ldots, x_N \in \mathcal{X}$ and any $(\alpha_i)_{i=1}^N \in \mathbb{R}^N$ ($N \in \mathbb{N}$), k is positive definite and called a kernel function on $\mathcal{X}$.

In the following, we write kernel for kernel function.

Remark 9.3.1. *Equivalently, a symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with $\mathcal{X}$ a nonempty set is a kernel iff. for any choice of $x_1, \ldots, x_N \in \mathcal{X}$ ($N \in \mathbb{N}$) it holds that $\mathbf{K} = [k(x_i, x_j)]_{i,j=1}^N \in \mathbb{R}^{N \times N}$ is a positive definite matrix.*

Suppose now that $\mathcal{X}$ is a nonempty set, $\mathcal{H}$ a Hilbert space, and $\phi : \mathcal{X} \to \mathcal{H}$ a feature map. By Exercise 9.1.1, $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ is a kernel on $\mathcal{X}$.

In general, the feature map associated to a kernel is not unique and even feature maps mapping to different Hilbert spaces can induce the same kernel. To illustrate, let $\mathcal{X} = \mathbb{R}^2$,

$$\phi_1 : \mathbb{R}^2 \to \mathbb{R}^3,\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \end{pmatrix},\ \text{and } \phi_2 : \mathbb{R}^2 \to \mathbb{R}^4,\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \\ x_1 x_2 \end{pmatrix}.$$

Then, for any $\mathbf{x} = (x_1, x_2)^\top \in \mathbb{R}^2$ and $\mathbf{y} = (y_1, y_2)^\top \in \mathbb{R}^2$,

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi_1(\mathbf{x}), \phi_1(\mathbf{y}) \rangle = \langle \phi_2(\mathbf{x}), \phi_2(\mathbf{y}) \rangle = x_1^2 y_1^2 + x_2^2 y_2^2 + 2 x_1 x_2 y_1 y_2.$$

Still, to each kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, one can associate a unique space, its reproducing kernel Hilbert space (RKHS), and a unique feature map, its canonical feature map. Before we establish the correspondence, we give a definition and investigate a few direct consequences.

Definition 9.3.2 (RKHS)**.** *Let $\mathcal{X}$ be a nonempty set and $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ a Hilbert space of functions $\mathcal{X} \to \mathbb{R}$. If the linear evaluation functional $\delta_x : \mathcal{H} \to \mathbb{R}$ defined by $h \mapsto h(x)$ is bounded for every $x \in \mathcal{X}$, we call $\mathcal{H}$ a reproducing kernel Hilbert space.*

We start our investigation by recalling a few facts about linear functionals. Indeed, let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces and recall that a linear operator $A : U \to V$ is bounded if $\|A\|_{\text{op}} \coloneqq \sup\{\|Au\|_V \mid u \in U,\ \|u\|_U = 1\} < \infty$. A linear functional $f : (U, \|\cdot\|_U) \to (\mathbb{R}, |\cdot|)$ is a linear operator. We have the following results.

Lemma 9.3.1. *If the linear functional $f : (U, \|\cdot\|_U) \to (\mathbb{R}, |\cdot|)$ is continuous at any $x_0 \in U$, it is continuous everywhere in U.*

Proof. Suppose that f is continuous at x_0, i.e., $f(x_n) \to f(x_0)$ as $x_n \to x_0$. Let $y_n \to y$. Then

$$f(y_n) = f(y_n - y + x_0 + y - x_0) = f(y_n - y + x_0) + f(y) - f(x_0) \to f(y)$$

as $y_n \to y$, by using that in this case $y_n - y + x_0 \to x_0$. □

For linear functionals, the terms 'bounded' and 'continuous' can be used interchangeably.

Theorem 9.3.1. *For any linear functional $f : (U, \|\cdot\|_U) \to (\mathbb{R}, |\cdot|)$, the conditions of continuity and boundedness are equivalent.*

Proof. ($\implies$) We argue by contraposition. Suppose that f is not bounded. Then, for every $n \in \mathbb{N}$, there exists $x_n \in U$ such that $|f(x_n)| > n\|x_n\|_U$, and we set $y_n = x_n/(n\|x_n\|_U)$. We have $\|y_n\|_U = 1/n$, i.e., $y_n \to 0$ as $n \to \infty$. However,

$$|f(y_n)| = \frac{1}{n\|x_n\|_U}|f(x_n)| > 1.$$

Hence, $f(x)$ is not continuous at $x = 0$.

($\impliedby$) Let $N = \|f\|_{\text{op}}$. Then, for any $x \in U$, $|f(x)| \leq N\|x\|_U$. For an arbitrary sequence $x_n \to 0$, it holds that

$$|f(x_n)| \leq N\|x_n\|_U \to 0$$

as $n \to \infty$. Hence, f is continuous at 0 and thus everywhere by Lemma 9.3.1. □

Remark 9.3.2. *In light of Theorem 9.3.1, an RKHS has a continuous linear evaluation functional.*

As an example of a linear functional, let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be a Hilbert space and fix an arbitrary nonzero $h \in \mathcal{H}$. Then $f(x) = \langle x, h\rangle_{\mathcal{H}}$ $(x \in \mathcal{H})$ is a linear functional on $\mathcal{H}$. Indeed, linearity follows by the linearity of the inner product. An application of the CBS inequality (Theorem 3.7.1) shows boundedness

$$|f(x)| \leq \|x\|_{\mathcal{H}}\|h\|_{\mathcal{H}},$$

with equality if $x = h/\|h\|_{\mathcal{H}}$. The last observation implies that $\|f\|_{\text{op}} = \|h\|_{\mathcal{H}}$. If $h = 0$, f is the zero linear functional and $\|f\|_{\mathcal{H}} = \|h\|_{\mathcal{H}} = 0$ as well.

The previous example is representative. In fact, *every* linear functional on a Hilbert space can be written as an inner product.

Theorem 9.3.2 (Riesz)**.** *For any bounded linear functional $f : (\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}}) \to (\mathbb{R}, |\cdot|)$ there exists a unique $h \in \mathcal{H}$ such that $f(x) = \langle x, h\rangle_{\mathcal{H}}$ for every $x \in \mathcal{H}$. Moreover, $\|f\|_{op} = \|h\|_{\mathcal{H}}$.*

We recall an elementary fact regarding the elements of a Hilbert space before proving the result.

Lemma 9.3.2. *Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be a Hilbert space and $\{e_n\}_{n=1}^{\infty}$ an orthonormal set in $\mathcal{H}$. The series $\sum_{n=1}^{\infty} \alpha_n e_n$ converges iff. $\sum_{n=1}^{\infty} |\alpha_n|^2 < \infty$.*

Proof. ($\implies$) Suppose that $\sum_{j=1}^{n} \alpha_j e_j$ converges to some $x \in \mathcal{H}$ as $n \to \infty$; then

$$\left\| \sum_{j=1}^{n} \alpha_j e_j \right\|^2 = \sum_{j=1}^{n} |\alpha_j|^2$$

converges to $\|x\|_{\mathcal{H}}^2$ as $n \to \infty$ by the continuity of $x \mapsto \|x\|_{\mathcal{H}}^2$. Hence, we have that $\sum_{n=1}^{\infty} |\alpha_n|^2 < \infty$.

($\impliedby$) Suppose that $\sum_{n=1}^{\infty} |\alpha_n|^2 < \infty$. Then $(\sum_{j=1}^{n} |\alpha_j|^2)_n$ is Cauchy. Let $x_n = \sum_{j=1}^{n} \alpha_j e_j$. Then, for any $m > n$,

$$\|x_m - x_n\|_{\mathcal{H}}^2 = \left\| \sum_{j=n+1}^{m} \alpha_j e_j \right\|^2 = \sum_{j=n+1}^{m} |\alpha_j|^2,$$

which shows that (x_n) is Cauchy. As $\mathcal{H}$ is complete, (x_n) thus converges to some $x \in \mathcal{H}$. □

We are ready to prove Theorem 9.3.2.

Proof. Let $(e_i)_{i\in I}$ be a countable orthonormal basis of $\mathcal{H}$, define $\alpha_i = f(e_i)$ $(i \in I)$, and $h = \sum_{i\in I} \alpha_i e_i$. We first show that $h \in \mathcal{H}$.

Let $J \subset I$ be any finite subset and denote by $h_J = \sum_{i\in J} \alpha_i e_i \in \mathcal{H}$. Then

$$f(h_J) = \sum_{i\in J} \alpha_i f(e_i) = \sum_{i\in J} |\alpha_i|^2,$$

and, consequently, by the properties of the operator norm

$$\sum_{i\in J} |\alpha_i|^2 \le \|f\|_{\mathrm{op}} \|h_J\|_{\mathcal{H}} = \|f\|_{\mathrm{op}} \sqrt{\sum_{i\in J} |\alpha_i|^2}.$$

Rearranging the inequality shows that $\sqrt{\sum_{i\in J} |\alpha_i|^2} \le \|f\|_{\mathrm{op}}$, which implies

$$\sqrt{\sum_{i\in I} |\alpha_i|^2} \le \|f\|_{\mathrm{op}}.$$

Hence, $h \in \mathcal{H}$ by Lemma 9.3.2.

The equivalence now follows by noting that for any $x = \sum_{i\in I} \beta_i e_i \in \mathcal{H}$, we have

$$f(x) = \sum_{i\in I} \beta_i f(e_i) = \sum_{i\in I} \alpha_i\beta_i = \left\langle \sum_{i\in I} \beta_i e_i, \sum_{i\in I} \alpha_i e_i \right\rangle_{\mathcal{H}} = \langle x, h\rangle_{\mathcal{H}}.$$

For uniqueness, suppose that there exist $g, h \in \mathcal{H}$ such that $f(x) = \langle x, g\rangle_{\mathcal{H}} = \langle x, h\rangle_{\mathcal{H}}$ for any $x \in \mathcal{H}$. Then $0 = \langle x, g\rangle_{\mathcal{H}} - \langle x, h\rangle_{\mathcal{H}} = \langle x, g-h\rangle_{\mathcal{H}}$ for every $x \in \mathcal{H}$, that is, $g = h$.

By CBS, we obtain the last claim as $|f(x)| = |\langle x, h\rangle_{\mathcal{H}}| \le \|x\|_{\mathcal{H}}\|h\|_{\mathcal{H}}$ with equality if $x = h/\|h\|_{\mathcal{H}}$. □

Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be an RKHS on $\mathcal{X} \neq \emptyset$. As the evaluation functional $\delta_x : \mathcal{H} \to \mathbb{R}$ is bounded for every $x \in \mathcal{X}$, by the Riesz representation theorem (Theorem 9.3.2), there exists a unique $k_x \in \mathcal{H}$ such that

$$f(x) = \delta_x(f) = \langle f, k_x\rangle_{\mathcal{H}}$$

holds for any $f \in \mathcal{H}$. This motivates the following definition.

Definition 9.3.3. *In the setting of Definition 9.3.2, we call k_x the reproducing kernel for the point $x \in \mathcal{X}$ and $k(x, y) = k_y(x)$ the reproducing kernel for $\mathcal{H}$ $(x, y \in \mathcal{X})$.*

Remark 9.3.3. *Notice that $k(x, y) = \langle k_x, k_y\rangle = \langle k_y, k_x\rangle = k(y, x)$, that is, the reproducing kernel is symmetric.*

The next results resolve the relationship of the reproducing kernel and a kernel function in the sense of Definition 9.3.1.

Lemma 9.3.3. *Let X be a nonempty set and $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ an RKHS on $\mathcal{X}$ with reproducing kernel k. Then k is a kernel function.*

Proof. Fix $x_1, \ldots, x_N \in \mathcal{X}$ and $(\alpha_i)_{i=1}^N \in \mathbb{R}^N$ $(N \in \mathbb{N})$. Then

$$\sum_{i,j=1}^N \alpha_i\alpha_j k(x_i, x_j) = \left\langle \sum_{i=1}^N \alpha_i k_{x_i}, \sum_{i=1}^N \alpha_i k_{x_i} \right\rangle_{\mathcal{H}} = \left\| \sum_{i=1}^N \alpha_i k_{x_i} \right\|_{\mathcal{H}}^2 \ge 0. \qquad \square$$

The proof of the converse statement reveals a lot about the structure of an RKHS.

Theorem 9.3.3 (Moore-Aronszajn)**.** *Let $\mathcal{X}$ be a nonempty set and $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ a kernel function. Then there exists an RKHS $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ on $\mathcal{X}$ of which k is the reproducing kernel.*

Let us collect a few results that we use in the proof of this statement.

Lemma 9.3.4. *Let U and V be vector spaces. A linear operator $T : U \to V$ is injective iff.* $\operatorname{null} T = \{0\}$.

Proof. See, for example, Robinson [2020, Lemma 1.2.1]. □

For a normed space $(U, \|\cdot\|_U)$, a *completion* is a complete normed space $(\mathcal{U}, \|\cdot\|_\mathcal{U})$ along with an isometric isomorphism $i : U \to \mathcal{U}$ onto a dense subspace of $\mathcal{U}$. We write $(i, \mathcal{U})$ for a completion.

Theorem 9.3.4. *Every normed space $(U, \|\cdot\|_U)$ has a completion $(i, \mathcal{U})$.*

Proof. See, for example, Robinson [2020, Theorem 7.3]. □

Lemma 9.3.5. *Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_\mathcal{H})$ be a (pre-)Hilbert space, $U \subset \mathcal{H}$ a dense subspace, and $x \in \mathcal{H}$. If $\langle x, u\rangle_\mathcal{H} = 0$ for all $u \in U$, then $x = 0$.*

Proof. See, for example, Werner [2018, V.1.5]. □

We are ready to prove Theorem 9.3.3.

Proof. We proceed by the following steps. (1.) We define a candidate reproducing kernel k_x for every point $x \in \mathcal{X}$ and a bilinear form B on the span $\mathcal{H}_0$ of the k_x-s. (2.) B is independent of the representation of $f, g \in \mathcal{H}_0$ and thus well-defined. (3.) We prove that B is an inner product. (4.) We construct a space $\hat{\mathcal{H}}$ isomorphic to the completion $\mathcal{H}$ of $\mathcal{H}_0$. (5.) We show that $\hat{\mathcal{H}}$ is an RKHS on $\mathcal{X}$.

(Step 1.) Let $k_x = k(\cdot, x)$ $(x \in \mathcal{X})$,[13] $\mathcal{H}_0 = \operatorname{span}\{k_x \mid x \in \mathcal{X}\}$, and define $B : \mathcal{H}_0 \times \mathcal{H}_0 \to \mathbb{R}$ by

$$(f, g) = \left(\sum_{i=1}^{n} \alpha_i k_{x_i}, \sum_{j=1}^{m} \beta_j k_{x_j}\right) \mapsto \sum_{i,j=1}^{n,m} \alpha_i \beta_j k(x_i, x_j).$$

(Step 2.) To show that B is independent of the representation of f it suffices to show that $B(f, g) = B(g, f) = 0$ iff. $f = \sum_{i=1}^{n} \alpha_i k_{x_i} \equiv 0$. In fact, it is sufficient to prove this statement for $g = k_y$, and we have $B(f, k_y) = \sum_{i=1}^{n} \alpha_i k(x_i, y) = f(y) = 0$. For the other direction, suppose that $B(f, h) = 0$ for every $h \in \mathcal{H}_0$. Taking $h = k_y$ shows that $B(f, k_y) = f(y) = 0$.

(Step 3.) The bilinearity of B follows directly from its definition. For any $f \in \mathcal{H}_0$, $B(f, f) \geq 0$ as k is a kernel. Hence, B is a semi-inner product. It remains to show that $B(f, f) = 0$ iff. $f \equiv 0$. Necessity follows by linearity. For sufficiency, notice that by CBS for semi-inner products, for any $y \in \mathcal{X}$,

$$|f(y)| = \left|\sum_{i=1}^{n} \alpha_i k(y, x_i)\right| = |B(f, k_y)| \leq B(f, f)^{1/2} B(k_y, k_y)^{1/2} = 0,$$

and $f \equiv 0$.

(Step 4.) Let $\mathcal{H}$ be the completion of $\mathcal{H}_0$ w.r.t. the norm induced by the inner product $B(\cdot,\cdot)$, which we now write as $\langle\cdot,\cdot\rangle_\mathcal{H}$. We first show that $\mathcal{H}$ can be identified with a function space on $\mathcal{X}$. To this end, define

$$\hat{h}(x) = \langle h, k_x\rangle_\mathcal{H} \text{ for } h \in \mathcal{H},$$

[13] $k(\cdot, x)$ denotes the function $x' \mapsto k(x', x)$ with $x \in \mathcal{X}$ fixed.

$\hat{\mathcal{H}} = \{\hat{h} \mid h \in \mathcal{H}\}$, and the map $L(h) = \hat{h}$. As L is linear, $\hat{\mathcal{H}}$ is a vector space of functions on $\mathcal{X}$. Notice that L is surjective. We now show that L is injective; by Lemma 9.3.4, it is sufficient to show that $\operatorname{null} L = \{0\}$, that is, $h \equiv 0$ iff. $L(h) \equiv 0$. Sufficiency is immediate. For necessity, suppose that $L(h) \equiv 0$. Then $\hat{h} = \langle h, k_x \rangle_{\mathcal{H}} = 0$ for every $x \in \mathcal{X}$, that is, $h \perp \mathcal{H}_0$. As $\mathcal{H}_0$ is dense in $\mathcal{H}$, $h = 0$ by Lemma 9.3.5. As L is surjective and injective, it follows that $\mathcal{H}$ and $\hat{\mathcal{H}}$ are isomorphic.

(Step 5.) Equip $\hat{\mathcal{H}}$ with the inner product $\langle \hat{h}_1, \hat{h}_2 \rangle_{\hat{\mathcal{H}}} = \langle h_1, h_2 \rangle_{\mathcal{H}}$. Then $\hat{\mathcal{H}}$ is a Hilbert space of functions on $\mathcal{X}$. Moreover, for any $x \in \mathcal{X}$,

$$\delta_x(\hat{h}) = \hat{h}(x) = \langle h, k_x \rangle_{\hat{\mathcal{H}}} = \langle \hat{h}, \hat{k}_x \rangle_{\hat{\mathcal{H}}} \leq \|\hat{h}\|_{\hat{\mathcal{H}}} \|\hat{k}_x\|_{\hat{\mathcal{H}}},$$

that is, the evaluation functional is bounded and $\hat{k}_x = k_x$ is the reproducing kernel for the point $x \in \mathcal{X}$. In other words, $\hat{\mathcal{H}}$ is an RKHS on $\mathcal{X}$ with reproducing kernel $\hat{k}_y(x) = \langle k_y, k_x \rangle_{\mathcal{H}} = k(x, y)$, with $\hat{k}_y$ the reproducing kernel for the point $y \in \mathcal{X}$. □

Remark 9.3.4. *It can be shown that the RKHS corresponding to a given kernel function is unique.*

Combining Lemma 9.3.3 and Theorem 9.3.3, we have the following result.

Corollary 9.3.1. *There is a one-to-one correspondence between an RKHS on a set and a kernel function on the set.*

Definition 9.3.4 (Canonical feature map)**.** *Let $\mathcal{H}$ be an RKHS on a set $\mathcal{X} \neq \emptyset$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We call $\mathcal{X} \ni x \mapsto k(\cdot, x) \in \mathcal{H}$ the canonical feature map.*

It will be convenient to have at our disposal a few tools to show that a symmetric function is a kernel.

Lemma 9.3.6. *Let k, k_1, and k_2 be kernel functions on a set $\mathcal{X}$, and $c \geq 0$. Then (i) $\tilde{k}(x, x') = k_1(x, x') + k_2(x, x')$ and (ii) $\tilde{k}(x, x') = ck(x, x')$ are kernels on $\mathcal{X}$.*

Lemma 9.3.7. *Let k_1 and k_2 be kernels on $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Then $k((x_1, x_2), (x_1', x_2')) = k_1(x_1, x_1')k_2(x_2, x_2')$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$.*

Lemma 9.3.8. *Let (k_n) be a sequence of kernels on $\mathcal{X}$. Then $k(x, x') = \lim_{n\to\infty} k_n(x, x')$ is a kernel on $\mathcal{X}$.*

Lemma 9.3.9. *Let k be a kernel on $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$. Then $\tilde{k}(x, x') = f(x)k(x, x')f(x')$ is a kernel on $\mathcal{X}$.*

Exercise 9.3.1. *Prove the preceding four lemmas.*

Well-known and frequently employed kernel functions on $\mathbb{R}^d$ include the linear kernel ($k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_2$), polynomial kernel ($k(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle_2 + c_0)^{c_1}$ with $c_0 \geq 0,\ c_1 \in \mathbb{N}$), and the Gaussian/RBF kernel ($k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2}$ with $\gamma > 0$).

Exercise 9.3.2. *Show that the Gaussian kernel is a kernel on $\mathbb{R}^d$.*

Remark 9.3.5. *The (canonical) feature map associated to a Gaussian kernel is not immediate but Theorem 9.3.3 assures its existence. In fact, the feature map associated to a Gaussian kernel is infinite-dimensional—still, by use of the kernel function, the inner product between feature maps can be computed in finite time.*

9.4 Notes

Introducing feature maps and kernels through basis expansions and ridge regression is inspired by Cristianini and Shawe-Taylor [2000], Hastie et al. [2009]. Our derivation of the SVM follows the latter; it is also detailed in Schölkopf and Smola [2002].

The original representer theorem is due to Kimeldorf and Wahba [1971]. Schölkopf et al. [2001] extend the result.

A systematic treatment of reproducing kernels is by Aronszajn [1950], which also states the Moore-Aronszajn theorem. However, our proof follows Paulsen and Raghupathi [2016]. More properties of kernels and their RKHSs can be found in Steinwart and Christmann [2008, Chapter 4].

A key result on kernel functions that we have omitted is Mercer's theorem. See Steinwart and Scovel [2012].

Kreyszig [1989] is a classic introductory text for the functional-analytic background. Good references—also including more advanced material—are Conway [1990], Werner [2018], Robinson [2020]. Our proofs regarding the linearity and boundedness/continuity of linear functionals are from Kolmogorov and Fomin [1957]. Regarding the Riesz representation theorem, to quote Halmos [1982]:

> The statement is [...] "coordinate-free", and therefore, according to current mathematical ethics, it is mandatory that the proof be such. The trouble is that most coordinate-free proofs ([...]) are so elegant that they conceal what is really going on.

Accordingly, we give a coordinatized proof of the result. See Robinson [2020] for an illustrated coordinate-free proof.

10 Information-theoretical measures

Many questions in data science and statistics revolve around comparing distributions, for example, answering if observed data matches assumed data, whether two distinct data sets have the same characteristics, or if different observed effects are independent. This section gives a quick introduction to the so-called kernel mean embedding, which is a powerful idea that allows answering these questions (and many other ones) in a principled fashion. The section also serves as an illustration of the "RKHS method", that is, choosing an RKHS as the function space in which one works to obtain computationally tractable quantities.

We start with a few results from integration theory.

10.1 The Bochner integral

Let (Ω, Σ, μ) be a finite measure space and $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ a Banach space. A function $f : \Omega \to \mathcal{X}$ is simple if it can be written as

$$f(\omega) = \sum_{i=1}^{n} x_i \mathbf{1}_{\{E_i\}}(\omega),$$

for some $x_1, \ldots, x_n \in \mathcal{X}$ and $E_1, \ldots, E_n \in \Sigma$. A function $f : \Omega \to \mathcal{X}$ is μ-measurable if there exists a sequence of simple functions (f_n) with $\lim_{n\to\infty} \|f_n - f\|_{\mathcal{X}} = 0$ μ-almost everywhere.

With these notations, we can define an integral for vector-valued functions.

Definition 10.1.1 (Bochner integral)**.** *A μ-measurable function $f : \Omega \to \mathcal{X}$ is Bochner integrable if there exists a sequence of simple functions (f_n) such that*

$$\lim_{n\to\infty} \int_{\Omega} \|f_n - f\|_{\mathcal{X}} \mathrm{d}\mu = 0.$$

In this case, $\int_E f \mathrm{d}\mu$ is defined for each $E \in \Sigma$ by

$$\int_E f \mathrm{d}\mu = \lim_{n\to\infty} \int_E f_n \mathrm{d}\mu,$$

with $\int_E f_n \mathrm{d}\mu$ defined in the obvious way.

We recall two results on Bochner integrals without proof.

Theorem 10.1.1. *A μ-measurable function $f : \Omega \to \mathcal{X}$ is Bochner integrable iff. $\int_\Omega \|f\|_{\mathcal{X}} \mathrm{d}\mu < \infty$. Further, $\| \int_E f \mathrm{d}\mu\|_{\mathcal{X}} \leq \int_E \|f\|_{\mathcal{X}} \mathrm{d}\mu$ for all $E \in \Sigma$.*

Theorem 10.1.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces and $T : \mathcal{X} \to \mathcal{Y}$ bounded linear. If f and Tf are Bochner integrable w.r.t. μ, then*

$$T\left(\int_E f \mathrm{d}\mu\right) = \int_E Tf \mathrm{d}\mu \text{ for all } E \in \Sigma.$$

Remark 10.1.1. *Theorem 10.1.2—due to Hille—holds for closed linear operators, which, for example, allows flipping differentiation and integration. For our use cases, the statement for bounded linear operators suffices.*

With the preliminaries defined, let us return to kernel methods.

10.2 Maximum mean discrepancy

For a topological space $(\mathcal{X}, \tau_{\mathcal{X}})$, we write $\mathcal{M}_1^+(\mathcal{X})$ for the set of all Borel probability measures on $\mathcal{X}$ (meant w.r.t. the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$). Suppose that we want to compare two probability measures.

Definition 10.2.1 (Integral probability metric; IPM)**.** *Let $P, Q \in \mathcal{M}_1^+(\mathcal{X})$ and $\mathcal{F}$ a class of real-valued measurable functions. We call*

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \int f \mathrm{d}P - \int f \mathrm{d}Q \right| \tag{59}$$

integral probability metric.

The IPM formulation captures many well-known distances on probability measures by appropriately choosing $\mathcal{F}$. (59) specializes to the Kantorovich metric ($\mathcal{F} = \{f : \|f\|_L \leq 1\}$, with $\|\cdot\|_L$ the Lipschitz-seminorm of bounded continuous functions on the metric space $(\mathcal{X}, d)$) and known to be the dual of the Wasserstein distance if $(\mathcal{X}, d)$ is separable, the total variation distance ($\mathcal{F} = \{f : \|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)| \leq 1\}$), the Kolmogorov-Smirnov distance ($\mathcal{F} = \{\mathbf{1}_{\{(-\infty, t]\}} : t \in \mathbb{R}^d\}$), and the maximum mean discrepancy ($\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$)—with $\mathcal{H}$ an RKHS on $\mathcal{X}$—, among others.

We elaborate how the maximum mean discrepancy (MMD) arises as an IPM after defining bounded kernel functions.

Definition 10.2.2. *A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on a set $\mathcal{X} \neq \emptyset$ is bounded if there exists a finite $c \geq 0$ such that $\sup_{x \in \mathcal{X}} k(x, x) \leq c$.*

Remark 10.2.1. *Notice the sensibility of this definition, as by the CBS inequality*

$$\sup_{x, x' \in \mathcal{X}} k(x, x') \leq \sup_{x, x' \in \mathcal{X}} \sqrt{k(x, x)} \sqrt{k(x', x')} = \sup_{x \in \mathcal{X}} k(x, x) \leq c.$$

Similar reasoning shows that a kernel is bounded if and only if it has bounded feature maps.

Lemma 10.2.1. *Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be an RKHS on $\mathcal{X}$ with associated bounded kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. If $\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$, then*

$$d_{\mathcal{F}}(P, Q) = \left\| \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}P(x) - \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}Q(x) \right\|_{\mathcal{H}},$$

for any $P, Q \in \mathcal{M}_1^+(\mathcal{X})$, where the integrals are meant in Bochner's sense.

Proof. By linearity, the reproducing property, Theorem 10.1.2, and CBS, it holds that

$$\begin{aligned} d_{\mathcal{F}}(P, Q) &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \int f(x) \mathrm{d}[P - Q](x) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \int \langle f, k(\cdot, x) \rangle_{\mathcal{H}} \mathrm{d}[P - Q](x) \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \left\langle f, \int k(\cdot, x) \mathrm{d}[P - Q](x) \right\rangle_{\mathcal{H}} = \left\| \int k(\cdot, x) \mathrm{d}[P - Q](x) \right\|_{\mathcal{H}}. \end{aligned}$$

□

We call the quantity on the r.h.s. of Lemma 10.2.1 *maximum mean discrepancy* (MMD), that is,

$$\mathrm{MMD}(P,Q) = \|\mu(P) - \mu(Q)\|_{\mathcal{H}},$$

where $\mu(P)$ and $\mu(Q)$ are the mean embeddings of P and Q, respectively, taking the form

$$\mu : \mathcal{M}_1^+(\mathcal{X}) \to \mathcal{H}, \qquad P \mapsto \int_{\mathcal{X}} k(\cdot, x)\mathrm{d}P(x),$$

with the integral meant in Bochner's sense.

Remark 10.2.2. *By Theorem 10.1.1, $\mu(P)$ exists for any $P \in \mathcal{M}_1^+(\mathcal{X})$ if the kernel k is bounded.*

If μ is injective on $\mathcal{M}_1^+(\mathcal{X})$, MMD induces a metric, and k is called characteristic.

To estimate MMD from samples $(X_i)_{i=1}^N \overset{\text{i.i.d.}}{\sim} P$ and $(Y_i)_{i=1}^M \overset{\text{i.i.d.}}{\sim} Q$, one may consider the plugin estimator, replacing P and Q by their empirical counterparts $\hat{P}_N = N^{-1}\sum_{i=1}^N \delta_{X_i}$ and $\hat{Q}_M = M^{-1}\sum_{i=1}^M \delta_{Y_i}$ and squaring, to obtain

$$\mathrm{MMD}^2(\hat{P}_N, \hat{Q}_M) = \frac{1}{N^2}\sum_{i,j=1}^N k(X_i, X_j) - \frac{2}{MN}\sum_{i,j=1}^{N,M} k(X_i, Y_j) + \frac{1}{M^2}\sum_{i,j=1}^M k(Y_i, Y_j).$$

Let us close the section by applying MMD to the two-sample testing problem, testing

$$H_0 : P = Q \qquad \text{versus} \qquad H_1 : P \neq Q,$$

for some $P, Q \in \mathcal{M}_1^+(\mathcal{X})$ only known through samples. Let us assume that k is characteristic, such that the testing problem is equivalent to

$$H_0 : \mathrm{MMD}(P,Q) = 0 \qquad \text{versus} \qquad H_1 : \mathrm{MMD}(P,Q) \neq 0.$$

Our goal is to find a threshold $t > 0$ to reject H_0 if $\mathrm{MMD}(\hat{P}_N, \hat{Q}_M) \geq t$ while bounding the probability $\alpha \in (0,1)$ that we reject H_0 incorrectly (called a level α test).

We collect a possible solution in the next result.

Theorem 10.2.1. *Suppose that $0 \leq k(x,y) \leq c$ for all $x, y \in \mathcal{X}$, k is characteristic, and $N = M$. Then choosing $t = \sqrt{2c/N}(1 + \sqrt{2\log(1/\alpha)})$ yields a level $\alpha \in (0,1)$ test.*

Proof. Our goal is to apply the bounded differences inequality (Theorem 4.3.1). Indeed, notice that $\mathrm{MMD}(\hat{P}_N, \hat{Q}_N)$ has the bounded differences property with $c_i = 2\sqrt{c}/N$ $(i = 1, \ldots, 2N)$ as can be seen by the reverse triangle inequality.

Hence, $\sum_{i=1}^{2N} c_i^2 = 8c/N$. It remains to bound $\mathbb{E}\,\mathrm{MMD}(\hat{P}_N, \hat{Q}_N)$:

$$\begin{aligned}\mathbb{E}\,\mathrm{MMD}(\hat{P}_N, \hat{Q}_N) &\leq \frac{1}{N}\left[\sum_{i,j=1}^{N} \mathbb{E}\left(k(X_i, X_j) + k(Y_i, Y_j) - 2k(X_i, Y_j)\right)\right]^{1/2} \\ &= \frac{1}{N}\left[2N\mathbb{E}k(X,X) + 2N(N-1)\mathbb{E}k(X,Y) - 2N^2\mathbb{E}k(X,Y)\right]^{1/2} \\ &= \frac{1}{N}\left[2N\mathbb{E}k(X,X) - 2N\mathbb{E}k(X,Y)\right]^{1/2} \leq \left(\frac{2c}{N}\right)^{1/2},\end{aligned}$$

where we used Jensen's inequality (Theorem 3.7.2). The application of the bounded differences inequality yields

$$P^{2N}\Big(\mathrm{MMD}(\hat{P}_N, \hat{Q}_N) - \mathbb{E}\,\mathrm{MMD}(\hat{P}_N, \hat{Q}_N) \geq t\Big) \leq e^{-\frac{t^2 N}{4c}},$$

and the combination with the expectation bound gives

$$P^{2N}\Big(\mathrm{MMD}(\hat{P}_N, \hat{Q}_N) - (2c/N)^{1/2} \geq t\Big) \leq e^{-\frac{t^2 N}{4c}}.$$

Solving for $\alpha = e^{-\frac{t^2 N}{4c}}$ and rearranging, we obtain the stated result. □

10.3 f-divergences

Next to IPMs, f-divergences provide another option for comparing probability measures, included here to provide a more comprehensive picture. We quickly recall the abstract setting, before providing a few examples.

Definition 10.3.1 (f-divergence)**.** *Let $P, Q \in \mathcal{M}_1^+(\mathcal{X})$ and $f : [0,\infty) \to (-\infty, \infty]$ a convex function.*[14] *The f-divergence of P and Q is*

$$D_f = \int_{\mathcal{X}} f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right) \mathrm{d}Q(x) \tag{60}$$

if P is absolutely continuous w.r.t. Q and $+\infty$ otherwise.

Well-known instances of (60) are the Kullback-Leibler divergence ($f(t) = t\log t$), Hellinger distance ($f(t) = (\sqrt{t}-1)^2$), χ^2-divergence ($f(t) = (t-1)^2$), or total variation distance ($f(t) = |t-1|$) [Sriperumbudur et al., 2012]. But, in practice, expression (60) can be challenging to estimate from samples, especially if one puts no additional assumptions on P and Q [Rubenstein et al., 2019].

Remark 10.3.1. *The total variation is the only non-trivial distance that is an IPM and an f-divergence.*

[14] Following Sriperumbudur et al. [2012], we do not require the usual $f(1) = 0$ condition.

10.4 Notes

Our definitions and the results on the Bochner integral are from Diestel and Uhl [1977, Chapter II]. Zolotarev [1983], Müller [1997] give a systematic treatment of IPMs and their generating classes of functions. A deeper discussion on IPMs and f-divergences is in Sriperumbudur et al. [2012].

Kernel mean embeddings [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007] have been known in statistics for more than 20 years. Their application to two-sample testing is in Gretton et al. [2012]. Notice that tighter bounds for MMD can be obtained by taking variance information into account [Kalinke and Gavioli-Akilagun, 2026]. See Muandet et al. [2017] for a survey on the kernel mean embedding.

References

Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

Francis Bach. *Learning Theory from First Principles*. MIT Press, 2024.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities*. Oxford University Press, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Sourav Chakraborty, N. V. Vinodchandran, and Kuldeep S. Meel. Distinct elements in streams: An algorithm for the (text) book. In *European Symposium on Algorithms (ESA)*, pages 34:1–34:6, 2022.

John B. Conway. *A Course in Functional Analysis*. Springer, second edition, 1990.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, 2000.

J. Diestel and J. J. Uhl, Jr. *Vector Measures*. American Mathematical Society, 1977.

John Duchi. CS229 supplemental lecture notes Hoeffding's inequality. Technical report, Stanford University. `https://cs229.stanford.edu/extra-notes/hoeffding.pdf`.

Rick Durrett. *Probability—Theory and Examples*. Cambridge University Press, 5th edition, 2019.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Paul Richard Halmos. *A Hilbert Space Problem Book*. Springer, second edition, 1982.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26(1):189–206, 1982.

Florian Kalinke and Shakeel Gavioli-Akilagun. Optimal online change detection via random Fourier features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. (to appear; preprint: `https://arxiv.org/abs/2505.17789`).

Dan Kalman. Leveling with Lagrange: an alternate view of constrained optimization. *Mathematics Magazine*, 82(3):186–196, 2009.

George Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

Donald E. Knuth. Simplex algorithm. Technical report, Stanford Computer Science Department, 2005. `https://cs.stanford.edu/~knuth/programs/lp.w`; typeset at `https://github.com/shreevatsa/knuth-literate-programs/blob/9b46afe/programs/lp.pdf`.

Donald E. Knuth. The CVM algorithm for estimating distinct elements in streams. Technical report, Stanford Computer Science Department, 2023. `https://cs.stanford.edu/~knuth/papers/cvm-note.pdf`.

A. N. Kolmogorov and S. V. Fomin. *Elements of the Theory of Functions and Functional Analysis. Vol. 1. Metric and Normed Spaces*. Graylock Press, 1957.

Erwin Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, Inc., 1989.

Alan J. Laub. *Matrix Analysis for Scientists & Engineers*. Society for Industrial and Applied Mathematics (SIAM), 2005.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer, 1991.

Jiří Matoušek. On variants of the Johnson-Lindenstrauss lemma. *Random Structures & Algorithms*, 33(2):142–156, 2008.

Thomas P. Minka. Old and new matrix algebra useful for statistics. Technical report, 2000. `https://tminka.github.io/papers/matrix/minka-matrix.pdf`.

Krikamol Muandet, Kenji Fukumizu, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016.

David Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2002.

James C. Robinson. *An Introduction to Functional Analysis*. Cambridge University Press, 2020.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Paul K. Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme, and Ilya O. Tolstikhin. Practical and consistent estimation of f-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4072–4082, 2019.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill Book Co., third edition, 1976.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. In *International Conference on Machine Learning (ICML)*, pages 515–521, 1998.

Bernhard Schölkopf and Alexander Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory (COLT)*, pages 416–426, 2001.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning*. Cambridge University Press, 2014.

Joel H. Shapiro. The Radon-Nikodym theorem "made easy". Technical report, Portland State University, 2018. `https://www.joelshapiro.org/Pubvit/Downloads/RN_HS/rnhs.pdf`.

Alexander Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

J. Michael Steele. *The Cauchy-Schwarz Master Class*. Mathematical Association of America; Cambridge University Press, 2004.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

Terence Tao. *Topics in Random Matrix Theory*. American Mathematical Society, 2012.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, 2000.

Roman Vershynin. *High-dimensional Probability*. Cambridge University Press, 2018.

Martin J. Wainwright. *High-dimensional Statistics*. Cambridge University Press, 2019.

Dirk Werner. *Funktionalanalysis*. Springer, eigth edition, 2018.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

V. Zolotarev. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1983.