Justus-Liebig-Universität Gießen | Winter 2023/24

Quantitative Foundations of Artificial Intelligence

Lecture Notes

Quantitative Foundations of Artificial Intelligence

Justus-Liebig-Universität Gießen | Winter 2023/24

Lecturer: Prof. Dr. Ludger Overbeck

Last edited on March 26, 2024

By Marvin Theiß

This work © 2024 by Marvin Theiß is licensed under CC BY-NC-SA 4.0 ⓒⓘⓢⓞ

You are free to distribute, remix, adapt, and build upon the material in any medium or format. You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. You may not use the material for commercial purposes. If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

This work is in no way affiliated with the Justus Liebig University (JLU), except for the fact that it is based on a series of lectures held at JLU in the winter of 2023/24.

Warning: These lectures notes have seen more all-nighters than a college coffee pot and were crafted with more love than grandma's secret cookie recipe. Should you nevertheless spot any mistakes (be it typos, grammar, logical errata, gaps, broken links, ...), feel free to send an email to mrvnthss@mail.de or open a pull request on GitHub and I will do my best to fix these as soon as possible.

Contents

Part I

Statistical Learning Theory

1. Formal Setting

In this first chapter, we briefly describe the formal setting that will accompany us for the remainder of this course. We will introduce some common machine learning terminology and we will fix some notation.

Throughout this course, we let (X, Y) denote a pair of random variables defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$, taking values in arbitrary spaces $(\mathcal{X}, \mathcal{F}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{F}_\mathcal{Y})$, respectively[1]. Most often, these spaces will be the Euclidean space $\mathbb{R}^d$ of dimension d. Also, we may sometimes just write Z for the pair (X, Y). We denote the joint distribution of X and Y by $\mathbb{P}_{(X,Y)}$, i.e.,

$$\mathbb{P}_{(X,Y)}(A, B) = \mathbb{P}(X \in A, Y \in B), \quad A \in \mathcal{F}_\mathcal{X},\ B \in \mathcal{F}_\mathcal{Y}.$$

Similarly, we write $\mathbb{P}_X$ and $\mathbb{P}_Y$ for the marginal distributions of X and Y, respectively. We will frequently be working with (hypothetical) data in this course. To this end, we let $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ denote a set of n pairs $Z_i = (X_i, Y_i)$, all of which are i.i.d. draws from the joint distribution of (X, Y).

1.1. Loss and Expected Error

In the context of machine learning, the random variable Y is generally some output/label resulting from or associated with a given input/observation X. In this setting, we often want to approximate or predict the value of Y given the input X. That is, we want to find a measurable function

$$f \colon (\mathcal{X}, \mathcal{F}_\mathcal{X}) \to (\mathcal{Y}, \mathcal{F}_\mathcal{Y})$$

mapping X to $f(X)$ in such a way that, ideally, $f(X) = Y$. Not surprisingly, perfectly predicting Y from X is practically impossible in real-world scenarios. This is in part due to the fact that the joint distribution of (X, Y) is *unknown* in practice. Hence, we will frequently be making errors. To quantify how far off our predictions are, we need a *loss function*

$$l \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

that let's us compare our prediction $f(x)$ with the true value y by evaluating $l(f(x), y)$. This works great for individual observations $(x, y) \in \mathcal{X} \times \mathcal{Y}$. However, most of the time, we are interested in the *expected error* or *risk*

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

of a function $f \colon \mathcal{X} \to \mathcal{Y}$ for a pair of random variables (X, Y). Clearly, $L(f)$ depends on the joint distribution of X and Y mentioned before. Once again, this joint distribution is unknown in practice. For this very reason, a large part of this course is devoted to finding *distribution-free* upper bounds[2] of $L(f)$. In Chapter 2, we will see that this takes considerable effort, even for seemingly simple tasks like binary classification.

[1] $\mathcal{F}_\mathcal{X}$ and $\mathcal{F}_\mathcal{Y}$ denote σ-algebras on $\mathcal{X}$ and $\mathcal{Y}$, respectively. We won't be concerned with measurability in this course, so these won't play an important role.

[2] i.e., bounds that are free of specific assumptions about the underlying probability distribution

Taking $\mathcal{Y} = \mathbb{R}^d$, one particularly popular choice of loss function in machine learning is the *squared error loss*[3]

$$l(y_1, y_2) = \|y_1 - y_2\|_2^2,$$

whose associated expected error is the well-known mean squared error (MSE)

$$\mathrm{MSE}(f, Y) = \mathrm{MSE}(f) = \mathbb{E}[\|f(X) - Y\|_2^2].$$

The mean squared error is closely related to the conditional expectation $\mathbb{E}[Y \mid X]$ of Y given X. The latter, in turn, plays a central role in binary classification, the topic of the next chapter. Hence, let's take a closer look at the conditional expectation $\mathbb{E}[Y \mid X]$.

1.2. Conditional Expectation

Generally speaking, this course assume a solid background in the basics of probability theory. Nonetheless, in this section, we do take the time to (very) briefly review the concept of *conditional expectation* for the reason mentioned above. To keep things easy, let's assume that we're working in the real numbers $\mathbb{R}$, and let X and Y be two real-valued random variables defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathbb{E}[Y^2] < \infty$. Then, the *conditional expectation* of Y given X is the unique[4] random variable $\mathbb{E}[Y \mid X]$ satisfying

(1) $\mathbb{E}[Y \mid X]$ is $\sigma(X)$-measurable,

(2) $\int_A \mathbb{E}[Y \mid X] \, \mathrm{d}\mathbb{P} = \int_A Y \, \mathrm{d}\mathbb{P}$ for all $A \in \sigma(X)$.

Note that, by setting $A = \Omega \in \sigma(X)$, the second property implies $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$, which is known as the *law of total expectation*. Moreover, the second property can be rewritten as

$$\mathbb{E}[\mathbf{1}_A \, \mathbb{E}[Y \mid X]] = \mathbb{E}[\mathbf{1}_A Y], \quad A \in \sigma(X),$$

and this can be generalized to

$$\mathbb{E}[Z \, \mathbb{E}[Y \mid X]] = \mathbb{E}[ZY]$$

for random variables Z that are $\sigma(X)$-measurable and integrable. Finally, this identity can be rewritten as

$$\mathbb{E}[Z(Y - \mathbb{E}[Y \mid X])] = 0, \tag{1.1}$$

with Z as before. To understand the geometric implications of this property, let's briefly review some basic linear algebra. Let V be a finite-dimensional real[5] vector space (think $\mathbb{R}^n$) with an inner product $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$, and let U be a linear subspace of V. The *orthogonal projection* onto U is the unique linear map $P \colon V \to V$ satisfying

(1) $P\mathbf{v} \in U$,

(2) $\mathbf{v} - P\mathbf{v} \in U^\perp$,

[3] $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbb{R}^d$

[4] This is to be understood in an almost sure sense, i.e., if Z is another random variable satisfying the two properties, then $Z = \mathbb{E}[Y \mid X]$ almost surely.

[5] This also works for vector spaces over the complex numbers, we're simply focusing on real vector spaces to keep it simple.

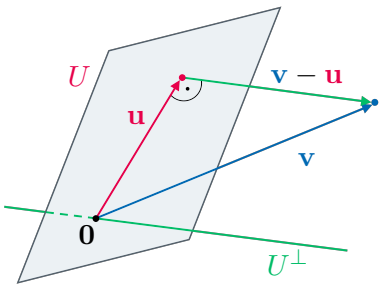

Figure 1.1. The orthogonal projection $\mathbf{u} = P(\mathbf{v})$ of a vector $\mathbf{v}$ onto the subspace U. Note that the vector difference $\mathbf{v} - \mathbf{u}$ is orthogonal to the subspace U, i.e., the vector $\mathbf{u}$ is a vector in U that minimizes the distance to $\mathbf{v}$.

for all $\mathbf{v} \in V$. Taken together, these two properties imply that, for every vector $\mathbf{v} \in V$, the projection $P\mathbf{v}$ is the vector in U that's closest to $\mathbf{v}$ with respect to the induced norm $\|\mathbf{v}\|^2 = \langle \mathbf{v}, \mathbf{v} \rangle$ on V. To see this, we choose two arbitrary vectors $\mathbf{v} \in V$ and $\mathbf{u} \in U$. We have

$$\|\mathbf{v} - \mathbf{u}\|^2 = \|(\mathbf{v} - P\mathbf{v}) + (P\mathbf{v} - \mathbf{u})\|^2 = \|\mathbf{v} - P\mathbf{v}\|^2 + 2\langle \mathbf{v} - P\mathbf{v}, P\mathbf{v} - \mathbf{u} \rangle + \|P\mathbf{v} - \mathbf{u}\|^2.$$

As $P\mathbf{v} - \mathbf{u} \in U$, and $\mathbf{v} - P\mathbf{v} \in U^\perp$, the term $\langle \mathbf{v} - P\mathbf{v}, P\mathbf{v} - \mathbf{u} \rangle$ vanishes, yielding

$$\|\mathbf{v} - \mathbf{u}\|^2 = \|\mathbf{v} - P\mathbf{v}\|^2 + \|P\mathbf{v} - \mathbf{u}\|^2 \geq \|\mathbf{v} - P\mathbf{v}\|^2,$$

which is exactly what we set out to prove.

Finally, to tie this in with our discussion of conditional expectation, we have to review some (very) fundamental functional analysis as well. For a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and $p > 0$, we denote[6] the set of measurable functions f for which $|f|^p$ is integrable with respect to $\mathbb{P}$ by $\mathcal{L}^p$, i.e.,

$$\mathcal{L}^p = \left\{ f \colon \Omega \to \mathbb{R} \;\middle|\; f \text{ is measurable, } \int_\Omega |f(\omega)|^p \, \mathrm{d}\mathbb{P}(\omega) < \infty \right\}, \quad p > 0.$$

On $\mathcal{L}^p$, we can define a seminorm via

$$\|f\|_p = \left(\int_\Omega |f(\omega)|^p \, \mathrm{d}\mathbb{P}(\omega) \right)^{1/p}, \quad f \in \mathcal{L}^p.$$

This does *not* yield a "true" norm since any two functions f and g that agree almost surely (but may differ on a null set) will share the same seminorm $\|f\|_p = \|g\|_p$. In particular, $\|f\|_p = 0$ does *not* imply $f = 0$ (but only $f = 0$ almost surely). However, this can easily be fixed by considering the set L^p of equivalence classes

$$[f] = \{ g \in \mathcal{L}^p \mid f \sim g \}, \qquad f \sim g \Leftrightarrow \mathbb{P}(f - g \neq 0) = 0,$$

i.e., we simply identify all functions that agree almost surely. On $L^p = \mathcal{L}^p / \sim$, we define $\|[f]\|_p = \|f\|_p$, and one can check that this indeed defines a norm on L^p, turning the latter into a *normed space*. Moreover, one can show (cf. Riesz–Fischer theorem) that the L^p-spaces are *complete* (i.e., every Cauchy sequence converges) with respect to this norm and hence, are *Banach spaces*.

[6] Note that $\mathcal{L}^p$ critically depends on the σ-algebra $\mathcal{A}$ (measurability) and the probability measure $\mathbb{P}$ (integrability), we only drop these in the notation for convenience!

Among all L^p-spaces, the space L^2 of (equivalence classes of) square-integrable functions is special, as it is the only L^p-space that can be equipped with an inner product that induces the norm $\|\cdot\|_2$ we defined earlier (i.e., L^2 is a *Hilbert space*). Indeed, for functions[7] $f, g \in L^2$, defining

$$\langle f, g \rangle_2 = \int_\Omega f(\omega) g(\omega) \, \mathrm{d}\mathbb{P}(\omega)$$

yields an inner product on L^2 satisfying $\langle f, f \rangle_2 = \|f\|_2^2$. Using these newly introduced concepts, the property (1.1) of the conditional expectation $\mathbb{E}[Y \mid X]$ can be reformulated as

$$\langle Z, Y - \mathbb{E}[Y \mid X] \rangle_2 = 0,$$

i.e., the difference between Y and the conditional expectation $\mathbb{E}[Y \mid X]$ is orthogonal to every $\sigma(X)$-measurable, integrable random variable Z. Moreover, note that the space of $\sigma(X)$-measurable, square-integrable random variables is a subspace of all square-integrable random variables on $(\Omega, \mathcal{A}, \mathbb{P})$. Since the conditional expectation $\mathbb{E}[Y \mid X]$ is, by definition, also $\sigma(X)$-measurable, identity (1.1) thus implies that the conditional expectation of Y given X is the orthogonal projection of Y onto the subspace of $\sigma(X)$-measurable, square-integrable random variables on $(\Omega, \mathcal{A}, \mathbb{P})$. As discussed earlier, this also implies that $\mathbb{E}[Y \mid X]$ is the random variable in the space of $\sigma(X)$-measurable, square-integrable random variables that is closest to Y with respect to the norm $\|\cdot\|_2$. In other words, taking into account all the information provided by the input X, the conditional expectation $\mathbb{E}[Y \mid X]$ is our best guess at the value of the output Y.

Finally, note that the conditional expectation $\mathbb{E}[Y \mid X]$ depends on X only via the σ-algebra $\sigma(X)$ generated by X, and, indeed, it makes perfect sense to define the conditional expectation of Y given a sub-σ-algebra $\mathcal{F} \subset \mathcal{A}$. Also, one can prove existence of the conditional expectation $\mathbb{E}[Y \mid \mathcal{F}]$ assuming that $\mathbb{E}[|Y|] < \infty$.

To conclude this intermezzo on conditional expectation, we list some[8] important properties[9] of conditional expectation:

- $\mathbb{E}[\mathbb{E}[Y \mid \mathcal{F}]] = \mathbb{E}[Y]$ (law of total expectation)
- $\mathbb{E}[\lambda Y + Z \mid \mathcal{F}] = \lambda \, \mathbb{E}[Y \mid \mathcal{F}] + \mathbb{E}[Z \mid \mathcal{F}]$ (linearity)
- If $Y \leq Z$, then $\mathbb{E}[Y \mid \mathcal{F}] \leq \mathbb{E}[Z \mid \mathcal{F}]$ (monotonicity)
- If Y is independent of $\sigma(Z, \mathcal{F})$, then $\mathbb{E}[YZ \mid \mathcal{F}] = \mathbb{E}[Y] \, \mathbb{E}[Z \mid \mathcal{F}]$ (pulling out independent factors)
- In particular, if Y is independent of $\mathcal{F}$, then $\mathbb{E}[Y \mid \mathcal{F}] = \mathbb{E}[Y]$ (pulling out independent factors)
- If Y is $\mathcal{F}$-measurable, then $\mathbb{E}[Y \mid \mathcal{F}] = Y$ (stability)
- In particular, for sub-σ-algebras $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{A}$, we have $\mathbb{E}[\mathbb{E}[Y \mid \mathcal{F}_1] \mid \mathcal{F}_2] = \mathbb{E}[Y \mid \mathcal{F}_1]$ (stability)
- If Y is $\mathcal{F}$-measurable, then $\mathbb{E}[YZ \mid \mathcal{F}] = Y \, \mathbb{E}[Z \mid \mathcal{F}]$ (pulling out known factors)
- For sub-σ-algebras $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{A}$, we have $\mathbb{E}[\mathbb{E}[Y \mid \mathcal{F}_2] \mid \mathcal{F}_1] = \mathbb{E}[Y \mid \mathcal{F}_1]$ (tower rule)

[7] For convenience, from now on we will simply speak of functions instead of equivalence classes of functions, and we will also drop the brackets for the equivalence classes.

[8] Note that this is merely a selection. In fact, there are many other useful properties of conditional expectation, such as monotone convergence, dominated convergence, Fatou's lemma, Jensen's inequality, and many more.

[9] Again, these identities/inequalities are to be understood in an almost sure sense.

In the previous section, we claimed that mean squared error and conditional expectation are closely related to each other. The following result makes this relationship explicit. Before we state the result, note that the conditional expectation of $Y \in \mathbb{R}^d$ given X is simply the vector of conditional expectations of Y_i given X for $i = 1, \dots, d$, i.e.,

$$\mathbb{E}[Y \mid X] = (\mathbb{E}[Y_1 \mid X], \dots, \mathbb{E}[Y_d \mid X])^\top.$$

We have the following result:

Proposition 1.1. *Let $Y \colon \Omega \to \mathbb{R}^d$ be a random variable such that $\mathbb{E}[|Y_i|] < \infty$, $i = 1, \dots, d$, and let $X \colon \Omega \to \mathcal{X}$ be a random variable and $f \colon \mathcal{X} \to \mathbb{R}^d$ a measurable function such that the mean squared error $\mathrm{MSE}(f)$ exists. Then,*

$$\mathrm{MSE}(f) = \mathrm{MSE}(f, \mathbb{E}[Y \mid X]) + \mathbb{E}[\sigma_Y^2(X)],$$

where

$$\sigma_Y^2(X) = \mathbb{E}[\|Y - \mathbb{E}[Y \mid X]\|_2^2 \mid X]$$

is the conditional variance *of Y given X. In particular,*

$$\mathrm{MSE}(\mathbb{E}[Y \mid X]) = \mathbb{E}[\sigma_Y^2(X)] = \mathbb{E}[\|Y - \mathbb{E}[Y \mid X]\|_2^2].$$

Proof. Bilinearity of the inner product yields

$$\begin{aligned}\mathrm{MSE}(f) &= \mathbb{E}[\|f(X) - Y\|_2^2] = \mathbb{E}[\|(f(X) - \mathbb{E}[Y \mid X]) + (\mathbb{E}[Y \mid X] - Y)\|_2^2] \\ &= \mathbb{E}[\|f(X) - \mathbb{E}[Y \mid X]\|_2^2] + \mathbb{E}[\|\mathbb{E}[Y \mid X] - Y\|_2^2] + 2\sum_{i=1}^{d} \mathbb{E}[(f_i(X) - \mathbb{E}[Y_i \mid X])(\mathbb{E}[Y_i \mid X] - Y_i)].\end{aligned}$$

The terms $\mathbb{E}[(f_i(X) - \mathbb{E}[Y_i \mid X])(\mathbb{E}[Y_i \mid X] - Y_i)]$ vanish for $i = 1, \dots, d$. By the law of total expectation, we have

$$\mathbb{E}[\|\mathbb{E}[Y \mid X] - Y\|_2^2] = \mathbb{E}[\|Y - \mathbb{E}[Y \mid X]\|_2^2] = \mathbb{E}[\mathbb{E}[\|Y - \mathbb{E}[Y \mid X]\|_2^2 \mid X]] = \mathbb{E}[\sigma_Y^2(X)].$$

Altogether,

$$\mathrm{MSE}(f) = \mathbb{E}[\|f(X) - \mathbb{E}[Y \mid X]\|_2^2] + \mathbb{E}[\|\mathbb{E}[Y \mid X] - Y\|_2^2] = \mathrm{MSE}(f, \mathbb{E}[Y \mid X]) + \mathbb{E}[\sigma_Y^2(X)]. \qquad \square$$

Remark 1.2. *(1) The quantity $\mathrm{MSE}(f, \mathbb{E}[Y \mid X])$ measures the mean squared error of using $f(X)$ as an approximation of the conditional expectation $\mathbb{E}[Y \mid X]$.*

(2) The term $\mathbb{E}[\sigma_Y^2(X)]$ is independent of the choice of $f \colon \mathcal{X} \to \mathbb{R}^d$. In particular, this shows that the mean squared error $\mathrm{MSE}(f)$ is minimized by $f(X) = \mathbb{E}[Y \mid X]$.

(3) The property $\mathbb{P}_{(X,Y)} = \mathbb{P}_{Y|X}\mathbb{P}_X$ enables us to consider $\mathcal{X} \times \mathcal{Y}$ sequentially, i.e., as an input space $\mathcal{X}$ followed *by an output space $\mathcal{Y}$, since, for any measurable function $f \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we have*

$$\mathbb{E}[f(X, Y)] = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} f(x, y) \, \mathrm{d}\mathbb{P}_{Y|X=x}(y) \right) \mathrm{d}\mathbb{P}_X(x).$$

2. Binary Classification

In the first chapter, we described the formal setting that will accompany us for the remainder of this course as we tackle various problems that machine learning aims to solve. In this chapter, we will focus our attention on one of these problems, namely *binary classification*. The decision to focus on *binary* (rather than *non-binary*) classification is based on several reasons. On the one hand, binary classification covers much of what we want to accomplish in practice, and the response variables are bounded. On the other hand, we avoid some of the nasty surprises of non-binary classification.

Let us first recall the setup of binary classification: the data we observe is a sequence $(X_1, Y_1), \dots, (X_n, Y_n)$ of n *feature-label pairs*, where each (X_i, Y_i) is an independent draw from some (unknown) joint distribution $\mathbb{P}_{(X,Y)}$. The random variable Y takes values in $\{0, 1\}$ and X takes values in some *feature space* $\mathcal{X}$, which in many use cases will be $\mathbb{R}^d$. Because Y is supported on $\{0, 1\}$, the conditional random variable $Y \mid X$ follows a Bernoulli distribution, which we denote by $Y \mid X \sim \mathrm{Ber}(\eta(X))$, where

$$\eta(X) = \mathbb{P}(Y = 1 \mid X) = \mathbb{E}[Y \mid X]$$

is the *regression function*. For $x \in \mathcal{X}$, the value[1] $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$ is also called the *a posteriori probability* of Y given the observation $X = x$.

2.1. Bayes Classifier

First, we define an optimal classifier in the setting of binary classification, the so-called *Bayes classifier*. This is the classifier that (somehow) is aware of the regression function $\eta(X)$, and, based on an observation $X = x$, predicts the class Y that has the higher a posteriori probability $\mathbb{P}(Y = 1 \mid X = x)$. This is the classifier we *would* use *if* we knew the distribution of $Y \mid X$.

Definition 2.1. *The Bayes classifier $h^* \colon \mathcal{X} \to \{0, 1\}$ is defined by*

$$h^* \colon \mathcal{X} \to \{0, 1\}, \quad x \mapsto \begin{cases} 1, & \eta(x) > 1/2 \\ 0, & \eta(x) \leq 1/2 \end{cases}$$

It follows from the definition that h^* equals 1 whenever $\mathbb{P}(Y = 1 \mid X) > \mathbb{P}(Y = 0 \mid X)$ or, equivalently, $\eta(X) > 1 - \eta(X)$. To quantify the performance of a classifier $h \colon \mathcal{X} \to \{0, 1\}$ in a binary classification tasks, we use the 0-1 *loss function*

$$l(y_1, y_2) = \mathbf{1}_{\{y_1 \neq y_2\}} = \begin{cases} 1, & y_1 \neq y_2 \\ 0, & y_1 = y_2 \end{cases}$$

[1] Note that $\eta(X) \colon \Omega \to [0, 1]$ is a random variable defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, while $\eta \colon \mathcal{X} \to [0, 1]$ is a deterministic function.

Hence, the expected error of a classifier h is given by its *error probability*:

$$L(h) = \mathbb{E}[\mathbf{1}_{\{h(X)\neq Y\}}] = \mathbb{P}(h(X) \neq Y).$$

Further, note that the error probability equals the expected absolute deviation of the prediction $h(X)$ from the true label Y, i.e.,

$$L(h) = \mathbb{E}[|h(X) - Y|],$$

since $\{h(X) \neq Y\} = \{|h(X) - Y| = 1\}$, which in turn is a consequence of h and Y only taking values in $\{0, 1\}$.

The Bayes classifier is optimal in the sense that no other classifier can have a lower error probability.

Theorem 2.2. *The Bayes classifier satisfies the following two properties:*

(1) The error probability $L^* = L(h^*)$ of the Bayes classifier is given by

$$L^* = \mathbb{E}[\min(1 - \eta(X), \eta(X))] = 1/2 - 1/2\,\mathbb{E}[|2\eta(X) - 1|] \leq 1/2\,. \tag{2.1}$$

(2) For each classifier $h\colon \mathcal{X} \to \{0, 1\}$, we have

$$L(h) - L^* = \mathbb{E}[|2\eta(X) - 1|\mathbf{1}_{\{h(X)\neq h^*(X)\}}] = \int_{h\neq h^*} |2\eta(x) - 1|\,\mathrm{d}\mathbb{P}_X(x) > 0\,. \tag{2.2}$$

Proof. For every classifier $h\colon \mathcal{X} \to \{0, 1\}$, we have

$$\begin{aligned} L(h) = \mathbb{P}(h(X) \neq Y) &= \mathbb{P}(h(X) = 0, Y = 1) + \mathbb{P}(h(X) = 1, Y = 0) \\ &= \mathbb{E}[\mathbf{1}_{\{h(X)=0\}}\mathbf{1}_{\{Y=1\}}] + \mathbb{E}[\mathbf{1}_{\{h(X)=1\}}\mathbf{1}_{\{Y=0\}}] \\ &= \mathbb{E}[\mathbb{E}[Y\mathbf{1}_{\{h(X)=0\}} \mid X]] + \mathbb{E}[\mathbb{E}[(1 - Y)\mathbf{1}_{\{h(X)=1\}} \mid X]], \end{aligned}$$

where the last identity follows from the law of total expectation and the observation that $\mathbf{1}_{\{Y=1\}} = Y$ and $\mathbf{1}_{\{Y=0\}} = 1 - Y$. Since the function $\mathbf{1}_{\{h(X)=c\}}$ is $\sigma(X)$-measurable ($c = 0, 1$), we can rewrite the previous line as

$$\mathbb{E}[\mathbb{E}[Y \mid X]\mathbf{1}_{\{h(X)=0\}}] + \mathbb{E}[\mathbb{E}[1 - Y \mid X]\mathbf{1}_{\{h(X)=1\}}].$$

As $\eta(X) = \mathbb{E}[Y \mid X]$, we conclude that

$$L(h) = \mathbb{E}[\eta(X)\mathbf{1}_{\{h(X)=0\}} + (1 - \eta(X))\mathbf{1}_{\{h(X)=1\}}]. \tag{2.3}$$

By definition, $\{h^*(X) = 0\} = \{\eta(X) \leq 1/2\}$. Thus, for the Bayes classifier h^*, identity (2.3) yields

$$L^* = \mathbb{E}[\eta(X)\mathbf{1}_{\{\eta(X)\leq 1/2\}} + (1 - \eta(X))\mathbf{1}_{\{\eta(X)>1/2\}}] = \mathbb{E}[\min(\eta(X), 1 - \eta(X))],$$

which proves the first assertion of (1). The identity

$$\mathbb{E}[\min(\eta(X), 1 - \eta(X))] = 1/2 - 1/2\,\mathbb{E}[|2\eta(X) - 1|]$$

is left as an exercise. To prove (2), first observe that

$$\begin{aligned} L(h) - L^* &= \mathbb{E}[\eta(X)(\mathbf{1}_{\{h(X)=0\}} - \mathbf{1}_{\{h^*(X)=0\}}) + (1-\eta(X))(\mathbf{1}_{\{h(X)=1\}} - \mathbf{1}_{\{h^*(X)=1\}})] \\ &= \mathbb{E}[(2\eta(X)-1)(\mathbf{1}_{\{h(X)=0\}} - \mathbf{1}_{\{h^*(X)=0\}})] \end{aligned}$$

by (2.3), where we have again used the fact that the identity $\mathbf{1}_{\{h(X)=1\}} = 1 - \mathbf{1}_{\{h(X)=0\}}$ holds for any function $h\colon \mathcal{X} \to \{0,1\}$. Further, a straightforward case analysis shows that

$$\mathbf{1}_{\{h(X)=0\}} - \mathbf{1}_{\{h^*(X)=0\}} = \begin{cases} 0, & h(X) = h^*(X) \\ \operatorname{sgn}(\eta(X) - 1/2), & h(X) \neq h^*(X) \end{cases}$$

and hence

$$L(h) - L^* = \mathbb{E}[(2\eta(X)-1)\operatorname{sgn}(\eta(X)-1/2)\mathbf{1}_{\{h(X)\neq h^*(X)\}}] = \mathbb{E}[|2\eta(X)-1|\mathbf{1}_{\{h(X)\neq h^*(X)\}}],$$

completing this proof. □

Example 2.3 (Bayes error)**.** *Let's consider the prediction of a student's performance in a final exam. We denote a passing grade by $Y = 1$ and a failing grade by $Y = 0$. Our only observation is the number $X\colon \Omega \to [0,\infty)$ of hours of study per week. It is (somewhat) reasonable to assume that the a posteriori probability*

$$\eta\colon [0,\infty) \to [0,1], \quad x \mapsto \mathbb{P}(Y = 1 \mid X = x),$$

is monotonically increasing in x. We further assume that η is given by

$$\eta(x) = \frac{x}{x+c}, \quad x \geq 0,$$

for some constant $c > 0$. In this case, the Bayes classifier is simply

$$h^*(x) = \begin{cases} 1, & x > c \\ 0, & x \leq c \end{cases}$$

By (2.1), the Bayes error is given by

$$L^* = \mathbb{E}[\min(\eta(X), 1-\eta(X))] = \mathbb{E}\left[\frac{\min(X,c)}{X+c}\right],$$

since $1 - \eta(x) = c/(x+c)$ for $x \geq 0$.

(1) Assume that $\mathbb{P}(X = c) = 1$, i.e., (almost surely) every student studies exactly c hours. In this scenario, the Bayes error is equal to $1/2$, i.e., as large as possible. This is due to the fact that $\eta(X)$ is constantly equal to $1/2$ and thus, completely uninformative.

(2) Assume that $X \sim U(0,4c)$. Then,

$$L^* = \frac{1}{4c}\int_0^{4c} \frac{\min(x,c)}{x+c}\,\mathrm{d}x = \frac{1}{4c}\left(\int_0^c \frac{x}{x+c}\,\mathrm{d}x + \int_c^{4c} \frac{c}{x+c}\,\mathrm{d}x\right) = \frac{1}{4}\log\left(\frac{5\mathrm{e}}{4}\right) \approx 0.305785.$$

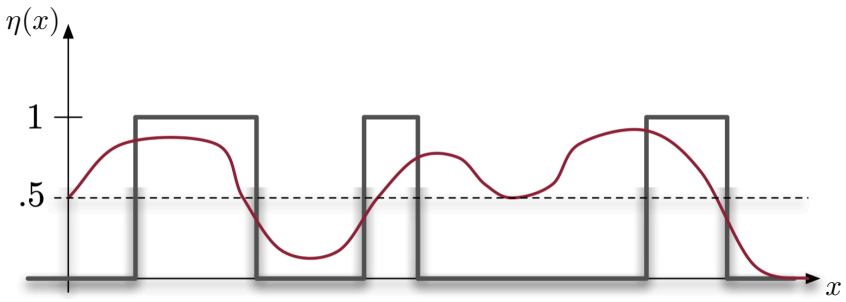

Figure 2.1. Illustration of the regression function η. **Gray**, the ideal case in which the observation X perfectly predicts Y, i.e., $\eta \in \{0,1\}$. **Red**, a more realistic scenario in which the regression function η takes values in the unit interval $[0,1]$.
Note. Adapted from "Mathematics of Machine Learning (18.657)," by P. Rigollet, Fall 2015, *Massachusetts Institute of Technology: MIT OpenCouseWare*, p. 2 (https://ocw.mit.edu/). CC BY-NC-SA 4.0.

Remark 2.4. *We can rewrite* (2.3) *as*

$$L(h) = 1 - \mathbb{E}[\eta(X)\mathbf{1}_{\{h(X)=1\}} + (1-\eta(X))\mathbf{1}_{\{h(X)=0\}}].$$

In particular, for the Bayes classifier, we have

$$L^* = 1 - \mathbb{E}[\eta(X)\mathbf{1}_{\{\eta(X)>1/2\}} + (1-\eta(X))\mathbf{1}_{\{\eta(X)\leq 1/2\}}].$$

Further, we observe that, by Proposition 1.1, the a posteriori probability

$$\eta(x) = \mathbb{P}(Y=1 \mid X=x) = \mathbb{E}[Y \mid X=x]$$

minimizes the mean squared error when Y is to be predicted by $h(X)$ for a function $h\colon \mathcal{X} \to \mathbb{R}$, i.e., the inequality

$$\mathbb{E}[(\eta(X)-Y)^2] \leq \mathbb{E}[(h(X)-Y)^2]$$

holds for all $h\colon \mathcal{X} \to \mathbb{R}$. In the exercises, we will see that the Bayes classifier is closely related to the mean absolute error

$$\mathbb{E}[|h(X)-Y|],$$

in the sense that h^* minimizes this error, i.e.,

$$L^* = \min_{h \in \mathbb{R}^{\mathcal{X}}} \mathbb{E}[|h(X)-Y|].$$

A function minimizing the mean absolute error is called the conditional median *of Y given X.*

Remark 2.5. *(1) For a classifier $h\colon \mathcal{X} \to \mathbb{R}$, the quantity*

$$R(h) = L(h) - L^* \geq 0$$

is called the excess risk *of h.*

(2) The error of the Bayes classifier equals $1/2$ if and only if $\eta(X) = 1/2$ almost surely. This is the case precisely when the feature variable X does not provide any *insight into the correct label Y. Essentially, the label Y is predicted by a coin flip, regardless of the information provided by the feature variable X.*

Further, (2.2) reveals that the excess risk weighs the discrepancy between the Bayes classifier h^* and any arbitrary classifier h based on how far η is from $1/2$. When η is close to $1/2$, any classifier will perform poorly and the excess risk is low. As η moves further away from $1/2$, the Bayes classifier performs well and classifiers that fail to do so are penalized more heavily.

So far, we have seen that the expected error of the Bayes classifier can be expressed as

$$L^* = \min_{h \in \mathbb{R}^{\mathcal{X}}} \mathbb{P}(h(X) \neq Y) = \mathbb{E}[\min(\eta(X), 1 - \eta(X))] = \frac{1}{2} - \frac{1}{2}\,\mathbb{E}[|2\eta(X) - 1|].$$

Next, we will look at special cases for which we can deduce other helpful ways of expressing the Bayes error. First, we assume that X has a density f with respect to the Lebesgue measure, i.e.,

$$\mathbb{P}(X \in A) = \int_A f(x)\,\mathrm{d}x.$$

Further, let f_i denote the conditional density of X given $Y = i$ for $i = 0, 1$. These are called *class-conditional densities*. The values $p = \mathbb{P}(Y = 1)$ and $1 - p = \mathbb{P}(Y = 0)$ are called *class probabilities*. Recall that, if X is continuous with density f_X and Y is discrete, Bayes rule states that

$$\mathbb{P}(Y = y \mid X = x) = \frac{f_{X|Y=y}(x)\,\mathbb{P}(Y = y)}{f_X(x)}\,.$$

In the setting above, this becomes

$$\eta(x) = \mathbb{P}(Y = 1 \mid X = x) = \frac{f_1(x)p}{f(x)}\,.$$

By the law of total probability, we can express the density $f(x)$ as $f_1(x)p + f_0(x)(1 - p)$ and thus obtain

$$\eta(x) = \mathbb{P}(Y = 1 \mid X = x) = \frac{f_1(x)p}{f_1(x)p + f_0(x)(1 - p)}\,. \tag{2.4}$$

By solving $\eta(x) > 1/2$, we see that the Bayes classifier is given by

$$h^*(x) = \begin{cases} 1, & f_1(x)p > f_0(x)(1 - p) \\ 0, & \text{else} \end{cases} \tag{2.5}$$

with error

$$L^* = \int_{\mathcal{X}} \min(\eta(x), 1 - \eta(x)) f(x)\,\mathrm{d}x = \int_{\mathcal{X}} \min(f_1(x)p, f_0(x)(1 - p))\,\mathrm{d}x,$$

since $\eta(x)f(x) = f_1(x)p$ and $(1 - \eta(x)) = f_0(x)(1 - p)$. Obviously, if f_1 and f_0 are non-overlapping, i.e., $\int f_0 f_1 = 0$, then $L^* = 0$. If we additionally assume that both classes are equally likely, i.e., $p = 1 - p = 1/2$, the Bayes classifier becomes

$$h^*(x) = \begin{cases} 1, & f_1(x) > f_0(x) \\ 0, & \text{else} \end{cases}$$

and its error is given by

$$L^* = \frac{1}{2}\int_{\mathcal{X}} \min(f_1(x), f_0(x))\,\mathrm{d}x = \frac{1}{2}\int_{\mathcal{X}} f_1(x) - (f_1(x) - f_0(x))^+\,\mathrm{d}x = \frac{1}{2} - \frac{1}{4}\int_{\mathcal{X}} |f_1(x) - f_0(x)|\,\mathrm{d}x,$$

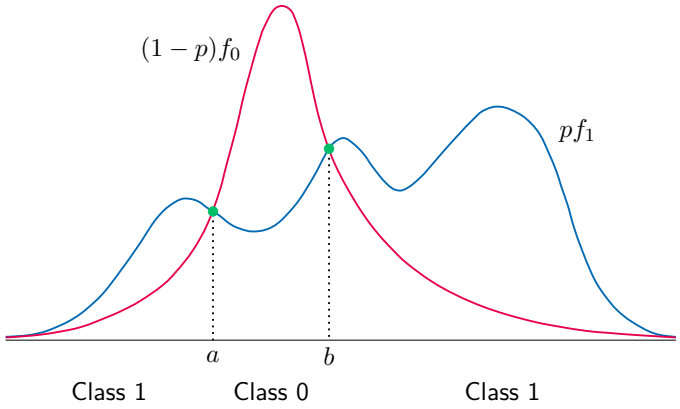

Figure 2.2. Illustration of the Bayes classifier in case the distribution of X is given by a density f and conditional densities f_0 and f_1 exist. The classifier equals 0 on the interval $[a, b]$ and 1 elsewhere.

where $g^+ = \max(g, 0)$ denotes the positive part of a function $g \in \mathbb{R}^{\mathcal{X}}$. This demonstrates that the Bayes error is directly related to the L^1-distance between the class densities f_1 and f_0.

2.2. Plug-In Decisions

We know that the Bayes classifier $h^*(x) = \mathbf{1}_{\{\eta(X)>1/2\}}$ is our best guess of Y based on the observation X. However, the function $\eta(x) = \mathbb{E}[Y \mid X = x]$ is often unknown. Assume that we have a function $\tilde{\eta}\colon \mathcal{X} \to [0, 1]$ approximating η. In this case, it is natural to use the *plug-in* classifier

$$\tilde{h}(x) = \begin{cases} 1, & \tilde{\eta}(x) > 1/2 \\ 0, & \tilde{\eta}(x) \le 1/2 \end{cases} \tag{2.6}$$

The error probability of this plug-in classifier is close to the Bayes error if $\tilde{\eta}$ is a good approximation of the regression function η.

Theorem 2.6. *The excess risk $R(\tilde{h})$ of the plug-in classifier $\tilde{h}$ defined in (2.6) satisfies*

$$R(\tilde{h}) = \mathbb{E}[|2\eta(X) - 1|\mathbf{1}_{\{\tilde{h}(X)\neq h^*(X)\}}] \le 2\,\mathbb{E}[|\eta(X) - \tilde{\eta}(X)|].$$

Proof. The first identity is simply an application of Theorem 2.2. Thus, all we need to demonstrate is that the inequality

$$|\eta(x) - 1/2| \le |\eta(x) - \tilde{\eta}(x)|$$

holds on the set $A = \{\tilde{h} \neq h^*\}$. First, assume that $\tilde{h}(x) = 0$ for $x \in A$. This implies that $h^*(x) = 1$. By definition of $\tilde{h}$ and h^*, we know that $\tilde{\eta}(x) \le 1/2$ and $\eta(x) > 1/2$, and thus

$$\tilde{\eta}(x) \le 1/2 < \eta(x).$$

If $\tilde{h}(x) = 1$, then $h^*(x) = 0$, and consequently $\tilde{\eta}(x) > 1/2$ and $\eta(x) \le 1/2$. Altogether, we have

$$\eta(x) \le 1/2 < \tilde{\eta}(x).$$

This proves the inequality stated at the beginning of this proof, and we're done. □

Since $\eta\colon \mathcal{X} \to [0,1]$ takes values in the unit interval, the condition $\eta(x) > 1/2$ is equivalent to the inequality $\eta(x) > 1 - \eta(x)$. Assuming that we have two functions $\tilde{\eta}_1\colon \mathcal{X} \to [0,1]$ and $\tilde{\eta}_0\colon \mathcal{X} \to [0,1]$ such that $\tilde{\eta}_1$ approximates η and $\tilde{\eta}_0$ approximates $1-\eta$ with the constraint that $\tilde{\eta}_1$ and $\tilde{\eta}_0$ do *not* sum to 1, we can still build a plug-in classifier $\tilde{h}$ by plugging $\tilde{\eta}_1$ and $\tilde{\eta}_0$ into the condition $\eta(x) > 1 - \eta(x)$:

$$\tilde{h}(x) = \begin{cases} 1, & \tilde{\eta}_1(x) > \tilde{\eta}_0(x) \\ 0, & \text{else} \end{cases} \tag{2.7}$$

While this puts us into a slightly different situation compared to Theorem 2.6, a bound similar to that in Theorem 2.6 still holds.

Theorem 2.7. *The excess risk $R(\tilde{h})$ of the plug-in classifier $\tilde{h}$ defined in (2.7) satisfies*

$$R(\tilde{h}) \leq \mathbb{E}[|\eta(X) - \tilde{\eta}_1(X)|] + \mathbb{E}[|(1-\eta(X)) - \tilde{\eta}_0(X)|].$$

The proof is left to the reader. Finally, assume that the class-conditional densities f_0 and f_1 exist and are approximated by $\tilde{f}_0$ and $\tilde{f}_1$. Further, let the class probabilities $p = \mathbb{P}(Y=1)$ and $1-p = \mathbb{P}(Y=0)$ be approximated by $\tilde{p}_1$ and $\tilde{p}_0$, respectively. In this case, one might use the plug-in decision

$$\tilde{h}(x) = \begin{cases} 1, & \tilde{f}_1(x)\tilde{p}_1 > \tilde{f}_0(x)\tilde{p}_0 \\ 0, & \text{else} \end{cases} \tag{2.8}$$

which is based on the Bayes classifier established in (2.5).

Proposition 2.8. *The excess risk $R(\tilde{h})$ of the plug-in classifier $\tilde{h}$ defined in (2.8) satisfies*

$$R(\tilde{h}) \leq \int_{\mathcal{X}} |p f_1(x) - \tilde{p}_1 \tilde{f}_1(x)| \,\mathrm{d}x + \int_{\mathcal{X}} |(1-p) f_0(x) - \tilde{p}_0 \tilde{f}_0(x)| \,\mathrm{d}x.$$

The proof is left to the reader once again.

2.3. Empirical Classification

In reality, we do *not* know the true distribution of (X,Y). Instead, we are often working with finite sets of data $\mathcal{D}_n = \{(X_1,Y_1),\dots,(X_n,Y_n)\}$ (e.g., a training set), where the (X_i,Y_i) are hypothetical future i.i.d. draws from $\mathbb{P}_{(X,Y)}$. Note that, if we build a classifier $h_n(X)$ based on this data, the classifier is random in two senses: it is a function of a random variable X, and it depends implicitly on the random data $\mathcal{D}_n$. Further, the expected error of h_n, i.e.,

$$L(h_n) = \mathbb{E}[\mathbf{1}_{\{h_n(X) \neq Y\}}] = \mathbb{P}(h_n(X) \neq Y),$$

is a *random variable*, as it depends on the random data $\mathcal{D}_n$! Nevertheless, the excess risk

$$R(h_n) = L(h_n) - L^* \geq 0$$

is always non-negative (in a deterministic sense, *not* only almost surely). One approach of building a classifier based on the observed data $\mathcal{D}_n$ is the following: based on the observations $(X_1,Y_1),\dots,(X_n,Y_n)$, we estimate

η by η_n and then construct the plug-in classifier $h_n = \mathbf{1}_{\{\eta_n > 1/2\}}$. However, the (true) excess risk of h_n is *not* computable based on the sampled data alone and we cannot use any results for upper bounds on the excess risk $R(h_n)$ discussed in Section 2.2. A naive attempt to solve this problem is the definition of *empirical risk*, which simply replaces the expected error $\mathbb{E}[l(h(X), Y)]$ by its empirical counterpart:

Definition 2.9. *The* empirical risk *of a classifier* $h\colon \mathcal{X} \to \mathcal{Y}$ *is given by*

$$\hat{L}_n(h) = \frac{1}{n} \sum_{i=1}^{n} l(h(X_i), Y_i).$$

The empirical risk $\hat{L}_n(h)$ of a classifier h is an *unbiased estimator* of the "true" risk $L(h)$, i.e.,

$$\mathbb{E}[\hat{L}_n(h)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[l(h(X_i), Y_i)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[l(h(X), Y)] = L(h), \tag{2.9}$$

since (X_i, Y_i) are i.i.d. draws from the distribution of (X, Y).

Minimizing the empirical risk over all classifiers is pointless since, for every hypothetical data set, we can always construct a classifier with *no* empirical risk by just mimicking the sampled data and classifying arbitrarily otherwise, e.g., the classifier

$$h_n(x) = \begin{cases} Y_i, & x = X_i \text{ for some } i = 1, \dots, n \\ 0, & x \notin \{X_1, \dots, X_n\} \end{cases}$$

satisfies $\hat{L}_n(h_n) = 0$. This phenomenon is called "overfitting": the classifier perfectly predicts the training data but fails to generalize to previously unseen data. Thus, in order to derive meaningful results from empirical risk minimization, we have to restrict ourselves to classifiers in a certain subset $\mathcal{H}$ of all measurable functions $h\colon \mathcal{X} \to \mathcal{Y}$.

Definition 2.10. *The* empirical risk minimizer *over* $\mathcal{H}$ *is defined as*

$$\hat{h}^{erm} \in \operatorname*{argmin}_{h \in \mathcal{H}} \hat{L}_n(h).$$

As just said, for meaningful results, $\mathcal{H}$ should be much smaller than the set of all classifiers. On the other hand, $\mathcal{H}$ should not be *too* small either, because we want the empirical risk of $\hat{h}^{\text{erm}}$ to be close to the Bayes error. Often, we will be satisfied with a classifier $\hat{h}$ that is reasonably close to the empirical risk minimizer $\hat{h}^{\text{erm}}$ in the sense that

$$\hat{L}_n(\hat{h}) \leq \hat{L}_n(\hat{h}^{\text{erm}}) + \varepsilon, \quad \text{for small } \varepsilon > 0.$$

Given a family $\mathcal{H}$, let $\bar{h}$ be a classifier that minimizes the *true* risk[2], i.e.,

$$\bar{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} L(h).$$

[2]While we cannot construct this classifier from observing finite samples, it does certainly exist nonetheless.

Note that it is impossible to construct such a classifier – often called an *oracle* – from data alone. Nevertheless, we can attempt to find a classifier $\hat{h}$ that mimics the performance of the oracle $\bar{h}$ in the sense that we can hope to prove a bound of the form

$$L(\hat{h}) \leq L(\bar{h}) + \text{something small.} \tag{2.10}$$

Such inequalities are often called *oracle inequalities* for the simple reason that they involve the oracle $\bar{h}$. In (2.10), we once again see the trade-off regarding the size of $\mathcal{H}$. If $\mathcal{H}$ is small, the performance of the oracle $\bar{h}$ is likely to suffer, while it might be possible to approximate $\bar{h}$ quite closely. If, on the other hand, $\mathcal{H}$ is quite large, the oracle $\bar{h}$ will be very powerful, but approximating $\bar{h}$ becomes statistically more difficult.

Ultimately[3], we want to prove tail bounds or bounds in expectation of the form

$$\mathbb{P}(L(\hat{h}) \leq L(\bar{h}) + \Delta_{n,\delta}(\mathcal{H})) \geq 1 - \delta,$$

where $\Delta_{n,\delta}(\mathcal{H})$ is some function of $\mathcal{H}$ depending on the sample size n and the desired level of confidence δ.

Finally, note that we can decompose the excess risk $R(\hat{h})$ of a classifier as follows:

$$R(\hat{h}) = L(\hat{h}) - L^* = \underbrace{L(\hat{h}) - L(\bar{h})}_{\substack{\text{estimation} \\ \text{error}}} + \underbrace{L(\bar{h}) - L^*}_{\substack{\text{approximation} \\ \text{error}}}$$

The second term is the *approximation error* that is unavoidable once we fix a family of classifiers $\mathcal{H}$. Oracle inequalities provide a method to bound the first term, the *estimation error*.

2.4. Hoeffding's Inequality

An important technique for understanding the empirical error and classifiers based on the empirical distribution are so-called *concentration inequalities*, which we will consider in more detail later. These are results that allow us to bound the deviations of a function of random variables from some value (often, its mean). Two very simple inequalities are the following:

- *Markov's inequality*: $\mathbb{P}(X \geq t) \leq \mathbb{E}[X]/t$ for $X \geq 0$ and $t > 0$,
- *Chebyshev's inequality*: $\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \mathrm{Var}(X)/t^2$ for X with $\mathrm{Var}(X) < \infty$ and $t > 0$.

The main result of this section is *Hoeffding's inequality*, which we will subsequently apply in order to bound the estimation error $L(\hat{h}) - L(\bar{h})$ (see Theorem 2.16).

To prove Hoeffding's inequality, we will have to control the moment-generating function[4] of a bounded random variable. Hence, we start with the following lemma:

Lemma 2.11 (Hoeffding's Lemma)**.** *Let X be a random variable satisfying $a \leq X \leq b$ almost surely.*

[3] Remember that $L(\hat{f})$ is a random variable!

[4] Recall that the moment-generating function of a random variable X is defined as $M_X(s) = \mathbb{E}[\mathrm{e}^{sX}]$. Its name reflects the fact that the n-th moment of X (given that it exists) can be obtained by evaluating the n-th derivative of M_X at $s = 0$.

Then, for any $s > 0$, the moment-generating function of $X - \mathbb{E}[X]$ satisfies

$$\mathbb{E}[\mathrm{e}^{s(X-\mathbb{E}[X])}] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right).$$

Proof. Write $\mu = \mathbb{E}[X]$, and let $Z = X - \mu$, so that $\mathbb{E}[Z] = 0$ and $a - \mu \leq Z \leq b - \mu$. It suffices to prove that the *cumulant-generating function* $\psi(s) = \log\big(\mathbb{E}[\mathrm{e}^{sZ}]\big)$ satisfies

$$\psi(s) \leq \frac{s^2(b-a)^2}{8}.$$

We can interchange the order of differentiation and integration since Z is bounded. Doing so yields

$$\psi'(s) = \frac{\mathbb{E}[Z\mathrm{e}^{sZ}]}{\mathbb{E}[\mathrm{e}^{sZ}]}$$

and

$$\psi''(s) = \frac{\mathbb{E}[Z^2\mathrm{e}^{sZ}]\,\mathbb{E}[\mathrm{e}^{sZ}] - \mathbb{E}[Z\mathrm{e}^{sZ}]^2}{\mathbb{E}[\mathrm{e}^{sZ}]^2} = \mathbb{E}\left[Z^2\frac{\mathrm{e}^{sZ}}{\mathbb{E}[\mathrm{e}^{sZ}]}\right] - \left(\mathbb{E}\left[Z\frac{\mathrm{e}^{sZ}}{\mathbb{E}[\mathrm{e}^{sZ}]}\right]\right)^2.$$

Since the term $\frac{\mathrm{e}^{sZ}}{\mathbb{E}[\mathrm{e}^{sZ}]}$ integrates to 1, we can interpret $\psi''(s)$ as the variance of Z with respect to the measure $\mathrm{d}\mathbb{Q} = \frac{\mathrm{e}^{sZ}}{\mathbb{E}[\mathrm{e}^{sZ}]}\,\mathrm{d}\mathbb{P}$, i.e.,

$$\psi''(s) = \mathbb{E}_{\mathbb{Q}}[Z^2] - \mathbb{E}_{\mathbb{Q}}[Z]^2 = \mathrm{Var}_{\mathbb{Q}}(Z) = \mathrm{Var}_{\mathbb{Q}}\left(Z + \mu - \frac{a+b}{2}\right).$$

The last identity in the previous computation holds since the variance of a random variable is not affected by constant shifts. From $Z + \mu \in [a, b]$ almost surely, it follows that $\left|Z + \mu - \frac{a+b}{2}\right| \leq \frac{b-a}{2}$ almost surely, and hence

$$\psi''(s) = \mathrm{Var}_{\mathbb{Q}}\left(Z + \mu - \frac{a+b}{2}\right) \leq \mathbb{E}_{\mathbb{Q}}\left[\left(Z + \mu - \frac{a+b}{2}\right)^2\right] \leq \frac{(b-a)^2}{4}.$$

Finally, the fundamental theorem of calculus yields

$$\psi(s) = \int_0^s \int_0^u \psi''(v)\,\mathrm{d}v\,\mathrm{d}u \leq \frac{(b-a)^2}{4}\int_0^s u\,\mathrm{d}u = \frac{(b-a)^2}{4}\frac{s^2}{2} = \frac{s^2(b-a)^2}{8},$$

concluding the proof. □

Equipped with this result, we can tackle Hoeffding's inequality:

Theorem 2.12 (Hoeffding, 1963)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that, almost surely, $a_i \leq X_i \leq b_i$, and denote the sum of the X_i by $S_n = \sum_{i=1}^n X_i$. Then, for any $t > 0$,*

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

In particular,

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

Proof. It suffices to prove the first inequality

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

as the bound for the lower tail follows analogously. For any $s > 0$, we have

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) = \mathbb{P}(\exp(s(S_n - \mathbb{E}[S_n])) \geq \exp(st)) \leq \exp(-st)\,\mathbb{E}[\exp(s(S_n - \mathbb{E}[S_n]))],$$

where we have applied Markov's inequality[5]. Since the X_i are independent, we have

$$\mathbb{E}[\exp(s(S_n - \mathbb{E}[S_n]))] = \mathbb{E}\left[\exp\left(\sum_{i=1}^n s(X_i - \mathbb{E}[X_i])\right)\right] = \prod_{i=1}^n \mathbb{E}[\exp(s(X_i - \mathbb{E}[X_i]))],$$

and hence, by Hoeffding's lemma,

$$\exp(-st)\prod_{i=1}^n \mathbb{E}[\exp(s(X_i - \mathbb{E}[X_i]))] \leq \exp(-st)\prod_{i=1}^n \exp\left(\frac{s^2(b_i - a_i)^2}{8}\right) = \exp\left(-st + \frac{s^2}{8}\sum_{i=1}^n (b_i - a_i)^2\right).$$

To optimize the bound, we want to minimize the expression inside the exponential function. For ease of notation, let $\lambda = \sum_{i=1}^n (b_i - a_i)^2$. With this notation, differentiating $f(s) = \frac{\lambda}{8}s^2 - ts$ and subsequently setting it equal to 0 gives

$$\frac{\lambda}{4}s - t = 0,$$

which yields the optimal solution $s^* = \frac{4t}{\lambda}$. Hence, the best bound we can find is

$$\exp(f(s^*)) = \exp\left(\frac{\lambda}{8}\left(\frac{4t}{\lambda}\right)^2 - t\frac{4t}{\lambda}\right) = \exp\left(\frac{-2t^2}{\lambda}\right) = \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

which is exactly the upper bound we wanted to prove. □

Remark 2.13. *Let $X_1, \ldots, X_n$ be random variables satisfying the conditions of Theorem 2.12, and write $\bar{X} = S_n/n$ for the arithmetic mean of the X_i. We have*

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

Further, observe that the expression $\bar{X} - \mathbb{E}[\bar{X}]$ can be rewritten as $\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i])$. Hence, Hoeffding's inequality ensures that the (absolute) average deviation of independent and (almost surely) bounded random variables from their respective means decays exponentially fast *in the number of observations n. Finally, if the random variables $X_1, \ldots, X_n$ are i.i.d. random variables such that, almost surely, $a \leq X_i \leq b$, Hoeffding's inequality states that*

$$\mathbb{P}(|\bar{X} - \mathbb{E}[X_1]| \geq t) \leq 2\exp\left(\frac{-2nt^2}{(b-a)^2}\right).$$

For random variables X_i taking values in the unit interval, this reduces to

$$\mathbb{P}(|\bar{X} - \mathbb{E}[X_1]| \geq t) \leq 2\mathrm{e}^{-2nt^2}.$$

[5] This is what is referred to as the *generic Chernoff bound*, which – for a random variable X – is attained by applying Markov's inequality to e^{tX} with $t > 0$. Doing so yields $\mathbb{P}(X \geq a) = \mathbb{P}(\mathrm{e}^{tX} \geq \mathrm{e}^{ta}) \leq \mathrm{e}^{-ta}\,\mathbb{E}(e^{tX}) = \mathrm{e}^{-ta}M_X(t)$.

2.5. Learning with a Finite Dictionary

Recall the setup of Section 2.3: given a family $\mathcal{H}$ of classifiers, we want to bound the estimation error $L(\hat{h})-L(\bar{h})$, where

$$\hat{h} \in \operatorname*{argmin}_{h\in\mathcal{H}} \hat{L}_n(h)$$

is the empirical risk minimizer, which minimizes the *empirical* risk over $\mathcal{H}$, and

$$\bar{h} \in \operatorname*{argmin}_{h\in\mathcal{H}} L(h)$$

is the oracle, which minimizes the *true* risk over the family $\mathcal{H}$. Keep in mind that the empirical risk minimizer $\hat{h}$ is a random variable that depends on the random data $\mathcal{D}_n$. In this section, we will prove a bound on the estimation error $L(\hat{h})-L(\bar{h})$ under the assumption that the family $\mathcal{H} = \{h_1, \dots, h_M\}$ is a *finite* set of classifiers. In particular, we want to better understand the scaling of the estimation error with respect to the number M of classifiers and the number n of observations.

The starting point of this endeavor is Vapnik's lemma, named after Russian mathematician Vladimir Vapnik.

Lemma 2.14 (Vapnik's Lemma). *For any family of classifiers $\mathcal{H}$, the estimation error $L(\hat{h}) - L(\bar{h})$ is bounded by*

$$L(\hat{h}) - L(\bar{h}) \le 2 \sup_{h\in\mathcal{H}} |\hat{L}_n(h) - L(h)|.$$

Proof. Clearly, the estimation error satisfies

$$L(\hat{h}) - L(\bar{h}) = L(\hat{h}) - \hat{L}_n(\bar{h}) + \hat{L}_n(\bar{h}) - L(\bar{h}) \le L(\hat{h}) - \hat{L}_n(\hat{h}) + \hat{L}_n(\bar{h}) - L(\bar{h}),$$

since the inequality $\hat{L}_n(\hat{h}) \le \hat{L}_n(\bar{h})$ holds by the very definition of the empirical risk minimzer. The result follows from applying a rather crude bound to the RHS of the inequality above:

$$\begin{aligned} L(\hat{h}) - L(\bar{h}) &\le L(\hat{h}) - \hat{L}_n(\hat{h}) + \hat{L}_n(\bar{h}) - L(\bar{h}) \\ &\le |\hat{L}_n(\hat{h}) - L(\hat{h})| + |\hat{L}_n(\bar{h}) - L(\bar{h})| \le 2 \sup_{h\in\mathcal{H}} |\hat{L}_n(h) - L(h)|. \end{aligned}$$

□

Hence, in order to find a bound of the estimation error of a class $\mathcal{H}$, it seems promising to find bounds of the discrepancy between the empirical risk $\hat{L}_n(h)$ and the true risk $L(h)$ of classifiers h in $\mathcal{H}$. A first result is the following:

Theorem 2.15. *(1) For a classifier $h\colon \mathcal{X} \to \{0,1\}$ and a confidence level $\delta > 0$,*

$$\mathbb{P}\left(|\hat{L}_n(h) - L(h)| < \sqrt{\frac{\log(2/\delta)}{2n}} \right) \ge 1 - \delta.$$

(2) If $\mathcal{H} = \{h_1, \dots, h_M\}$ is a finite *family of M classifiers, then*

$$\mathbb{P}\left(\max_{h\in\mathcal{H}} |\hat{L}_n(h) - L(h)| < \sqrt{\frac{\log(2M/\delta)}{2n}} \right) \ge 1 - \delta.$$

Proof. Since (X_i, Y_i) are i.i.d. draws from $\mathbb{P}_{(X,Y)}$, we have

$$\mathbb{P}(h(X) \neq Y) = \mathbb{E}[\mathbf{1}_{\{h(X)\neq Y\}}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbf{1}_{\{h(X)\neq Y\}}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbf{1}_{\{h(X_i)\neq Y_i\}}],$$

and hence,

$$\hat{L}_n(h) - L(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{h(X_i)\neq Y_i\}} - \mathbb{E}[\mathbf{1}_{\{h(X_i)\neq Y_i\}}].$$

Consequently, the first assertion follows from Hoeffding's inequality applied to $\mathbf{1}_{\{h(X_i)\neq Y_i\}}$, with $i = 1, \ldots, n$, and $t = \sqrt{\frac{\log(2/\delta)}{2n}}$, since $\mathbb{P}(|\ldots| < t) = 1 - \mathbb{P}(|\ldots| \geq t) \geq 1 - 2\mathrm{e}^{-2nt^2} = 1 - \delta$.

For the second assertion, observe that, for $t = \sqrt{\frac{\log(2M/\delta)}{2n}}$,

$$\left\{\max_{h\in\mathcal{H}} |\hat{L}_n(h) - L(h)| \geq t\right\} = \bigcup_{j=1}^{M} \{|\hat{L}_n(h_j) - L(h_j)| \geq t\}$$

and hence,

$$\mathbb{P}(\max_{h\in\mathcal{H}} |\hat{L}_n(h) - L(h)| < t) = 1 - \mathbb{P}(\max_{h\in\mathcal{H}} |\hat{L}_n(h) - L(h)| \geq t) \geq 1 - \sum_{j=1}^{M} \mathbb{P}(|\hat{L}_n(h_j) - L(h_j)| \geq t)$$

by applying a simple union bound. As before, Hoeffding's inequality yields

$$\mathbb{P}(|\hat{L}_n(h_j) - L(h_j)| \geq t) \leq 2\mathrm{e}^{-2nt^2} = \frac{\delta}{M}, \quad j = 1, \ldots, M.$$

Putting everything together, we obtain

$$\mathbb{P}(\max_{h\in\mathcal{H}} |\hat{L}_n(h) - L(h)| < t) \geq 1 - \sum_{j=1}^{M} \mathbb{P}(|\hat{L}_n(h_j) - L(h_j)| \geq t) \geq 1 - \sum_{j=1}^{M} \frac{\delta}{M} = 1 - \delta. \qquad \square$$

We can now use this intermediate result to obtain a bound on the estimation error $L(\hat{h}) - L(\bar{h})$. Loosely speaking, while the empirical risk minimizer $\hat{h}$ generally performs worse (in terms of the true risk) than the oracle $\bar{h}$, their difference is not too big, given that the sample size n is large enough compared to the number M of classifiers in $\mathcal{H}$.

Theorem 2.16. *For a finite family of classifiers $\mathcal{H} = \{h_1, \ldots, h_M\}$ and a confidence level $\delta > 0$, the estimation error $L(\hat{h}) - L(\bar{h})$ satisfies*

$$\mathbb{P}\left(L(\hat{h}) - L(\bar{h}) < \sqrt{\frac{2\log(2M/\delta)}{n}}\right) \geq 1 - \delta.$$

Proof. From Vapnik's lemma (Lemma 2.14), we know that

$$L(\hat{h}) - L(\bar{h}) \leq 2\sup_{h\in\mathcal{H}} |\hat{L}_n(h) - L(h)| = 2\max_{1\leq j\leq M} |\hat{L}_n(h_j) - L(h_j)|.$$

Hence, Theorem 2.15 implies

$$\mathbb{P}\left(L(\hat{h}) - L(\bar{h}) < \sqrt{\frac{2\log(2M/\delta)}{n}}\right) \geq \mathbb{P}\left(2 \max_{1\leq j\leq M} |\hat{L}_n(h_j) - L(h_j)| < \sqrt{\frac{2\log(2M/\delta)}{n}}\right)$$
$$= \mathbb{P}\left(\max_{1\leq j\leq M} |\hat{L}_n(h_j) - L(h_j)| < \sqrt{\frac{\log(2M/\delta)}{2n}}\right) \geq 1 - \delta. \qquad \square$$

Let us make a few remarks to close this chapter. First, the parameter $\delta > 0$ allows us to control the confidence we want to have in the bound on the estimation error. Here, we see a typical trade-off: for smaller δ, the bound on the estimation error becomes less restrictive, but at the same time, the probability that the bound actually holds, increases (and vice versa). Further, for fixed δ, the bound becomes sharper as the number n of observations increases and the number M of classifiers decreases.

3. Concentration Inequalities

Concentration inequalities are results that allow us to bound the deviations of a function of random variables from its mean. The first inequality we will consider improves on Hoeffding's inequality by allowing for some dependence between the random variables.

3.1. Azuma-Hoeffding Inequality

To state the main result of this section, we first need to introduce the concept of *martingales*.

Definition 3.1. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space.*

(1) A filtration *$\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ is an increasing sequence of sub-σ-algebras $\mathcal{F}_n$ of $\mathcal{A}$, i.e.,*

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{A}.$$

(2) A stochastic process $\{X_n\}_{n\in\mathbb{N}}$ is called a martingale *if, for every $n \in \mathbb{N}$,*

(a) X_n is $\mathcal{F}_n$-measurable and integrable, i.e., $\mathbb{E}[|X_n|] < \infty$,

(b) $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n$ holds almost surely.

Loosely speaking, martingales are a generalization of sums of zero-mean, independent random variables: Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent random variables with $\mathbb{E}[X_n] = 0$, and let $S_n = \sum_{i=1}^n X_i$. Then, for $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, we have

$$\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[X_{n+1} + S_n \mid \mathcal{F}_n] = \mathbb{E}[X_{n+1}] + \mathbb{E}[S_n \mid \mathcal{F}_n] = S_n,$$

since X_{n+1} is centered and independent of $\mathcal{F}_n$, and S_n is $\mathcal{F}_n$-measurable.

Another definition we will need is that of a *martingale difference sequence*.

Definition 3.2. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space with filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$. A stochastic process $\{X_n\}_{n\in\mathbb{N}}$ is called a* martingale difference sequence (MDS)*, if*

(1) X_n is $\mathcal{F}_n$-measurable and integrable, i.e., $\mathbb{E}[|X_n|] < \infty$,

(2) $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = 0$ holds almost surely.

Whenever we have a martingale $\{X_n\}_{n\in\mathbb{N}}$, we can construct a martingale difference sequence by setting $\Delta_n = X_n - X_{n-1}$, since then

$$\mathbb{E}[\Delta_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[X_{n+1} - X_n \mid \mathcal{F}_n] = \mathbb{E}[X_{n+1} \mid \mathcal{F}_n] - X_n = 0.$$

With these definitions at hand, we can state the main result of this section:

Theorem 3.3 (Azuma-Hoeffding). *Assume that $\{\Delta_i\}_{i\in\mathbb{N}}$ is a martingale difference sequence with respect to a filtration $\{\mathcal{F}_i\}_{i\in\mathbb{N}}$ and let A_i, B_i be $\mathcal{F}_i$-measurable random variables such that*

$$A_i \leq \Delta_i \leq B_i$$

holds almost surely for all $i \in \mathbb{N}$. Then,

$$\mathbb{P}\left(\sum_{i=1}^{n} \Delta_i \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} \|B_i - A_i\|_\infty^2}\right).$$

In comparison to Hoeffding's inequality, the Azuma-Hoeffding inequality allows for non-uniform boundedness and does *not* require independence of the random variables.

Proof. For ease of notation, we write $S_k = \sum_{i=1}^{k} \Delta_i$, with $k = 1, \ldots, n$. As in the proof of Hoeffding's inequality, we first apply the generic Chernoff bound to obtain

$$\begin{aligned}\mathbb{P}(S_n \geq t) \leq \mathrm{e}^{-st}\mathbb{E}[\exp(sS_n)] &= \mathrm{e}^{-st}\,\mathbb{E}[\mathbb{E}[\exp(sS_n) \mid \mathcal{F}_{n-1}]] \\ &= \mathrm{e}^{-st}\,\mathbb{E}[\exp(sS_{n-1})\mathbb{E}[\exp(s\Delta_n) \mid \mathcal{F}_{n-1}]].\end{aligned}$$

Next, applying Hoeffding's lemma (Lemma 2.11) to the inner expectation $\mathbb{E}[\exp(s\Delta_n) \mid \mathcal{F}_{n-1}]$ yields

$$\mathbb{P}(S_n \geq t) \leq \mathrm{e}^{-st}\,\mathbb{E}\left[\exp(sS_{n-1})\exp\left(s^2\frac{(B_n - A_n)^2}{8}\right)\right] \leq \mathrm{e}^{-st}\exp\left(\frac{s^2}{8}\|B_n - A_n\|_\infty^2\right)\mathbb{E}[\exp(sS_{n-1})].$$

Iteratively splitting off the Δ_i one-by-one and then applying Hoeffding's lemma, we eventually get

$$\mathbb{P}(S_n \geq t) \leq \mathrm{e}^{-st}\exp\left(\frac{s^2}{8}\sum_{i=1}^{n}\|B_i - A_i\|_\infty^2\right).$$

Optimizing the RHS over s yields the desired result. □

3.2. Bounded Differences Inequality

Although the Azuma-Hoeffding Inequality is a strong result, it can be challenging to apply to a specific problem, and its complete usefulness is often wasted. Thankfully, there is a natural choice of the filtration $\{\mathcal{F}_i\}_{i\in\mathbb{N}}$ and the MDS $\{\Delta_i\}_{i\in\mathbb{N}}$ that provides an equally potent result that is easier to use. Before we can state said result, we need to introduce yet another definition.

Definition 3.4. *A function $g\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ satisfies the* bounded differences condition *for constants $c_1, \ldots, c_n \in \mathbb{R}$, if the inequality*

$$\sup_{\tilde{x}_i \in \mathcal{X}_i} |g(x_1, \ldots, x_n) - g(x_1, \ldots, \tilde{x}_i, \ldots, x_n)| \leq c_i$$

holds for all $i = 1, \ldots, n$ and all $x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n$.

Intuitively, g meets the bounded differences condition if changing only one coordinate of g at a time cannot cause the value of g to deviate too far. For example, the empirical risk $\hat{L}_n(h) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{h(X_i)\neq Y_i\}}$ of a classifier h satisfies the bounded differences condition for $c_i = \frac{1}{n}$. Similarly, the empirical mean $\frac{1}{n}\sum_{i=1}^n X_i$ of bounded random variables $X_i \in (a_i, b_i)$ satisfies the bounded differences inequality for $c_i = b_i - a_i$. It is not too surprising that these types of functions concentrate somewhat strongly around their average, and this intuition is made precise by the following result.

Theorem 3.5 (Bounded Differences Inequality [McDiarmid, 1989])**.** *Let $X_1, \dots, X_n$ be independent random variables and let $g\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be a function satisfying the bounded differences condition for constants $c_1, \dots, c_n \in \mathbb{R}$. Then,*

$$\mathbb{P}(g(X_1, \dots, X_n) - \mathbb{E}[g(X_1, \dots, X_n)] \geq t) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Before we proceed to prove the bounded differences inequality, we observe the following:

Lemma 3.6. *Every function $g\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ satisfying the bounded differences condition is bounded.*

Proof. Assume that g satisfies the bounded differences inequality for constants $c_1, \dots, c_n \in \mathbb{R}$. Further, let $x_i, y_i \in \mathcal{X}_i$, for $i = 1, \dots, n$. Then, by the triangle inequality

$$\begin{aligned}|g(x_1, \dots, x_n) - g(y_1, \dots, y_n)| &\leq |(g(x_1, \dots, x_n) - g(x_1, y_2, \dots, y_n)| + |g(x_1, y_2, \dots, y_n) - g(y_1, \dots, y_n)| \\ &\leq |(g(x_1, \dots, x_n) - g(x_1, y_2, \dots, y_n)| + c_1.\end{aligned}$$

Iteratively applying the triangle inequality in this manner yields $|g(x_1, \dots, x_n) - g(y_1, \dots, y_n)| \leq \sum_{i=1}^n c_i$, showing that g is indeed bounded, since the RHS is independent of the chosen x_i and y_i. □

Proof of Theorem 3.5. Since g is bounded by Lemma 3.6, the random variable $g(X_1, \dots, X_n)$ is integrable. Hence, we can construct the *Doob martingale*[1]

$$M_0 = \mathbb{E}[g(X_1, \dots, X_n)], \qquad M_i = \mathbb{E}[g(X_1, \dots, X_n) \mid \mathcal{F}_i], \quad i = 1, \dots, n,$$

where $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$ is the σ-algebra generated by the X_i. From this martingale, we can construct a MDS $\{\Delta_i\}_i$ by setting $\Delta_i = M_i - M_{i-1}$, $i = 1, \dots, n$. Next, for each $i = 1, \dots, n$, we define

$$\begin{aligned}L_i &= \inf_{x \in \mathcal{X}_i} \mathbb{E}[g(X_1, \dots, X_{i-1}, x, X_{i+1}, \dots, X_n) \mid \mathcal{F}_i] - \mathbb{E}[g(X_1, \dots, X_n) \mid \mathcal{F}_{i-1}], \\ &= \inf_{x \in \mathcal{X}_i} \mathbb{E}[g(X_1, \dots, X_{i-1}, x, X_{i+1}, \dots, X_n) - g(X_1, \dots, X_n) \mid \mathcal{F}_{i-1}],\end{aligned}$$

and similarly

$$\begin{aligned}U_i &= \sup_{x \in \mathcal{X}_i} \mathbb{E}[g(X_1, \dots, X_{i-1}, x, X_{i+1}, \dots, X_n) \mid \mathcal{F}_i] - \mathbb{E}[g(X_1, \dots, X_n) \mid \mathcal{F}_{i-1}], \\ &= \sup_{x \in \mathcal{X}_i} \mathbb{E}[g(X_1, \dots, X_{i-1}, x, X_{i+1}, \dots, X_n) - g(X_1, \dots, X_n) \mid \mathcal{F}_{i-1}].\end{aligned}$$

[1] See here. The original paper by J. L. Doob can be found here.

We have $L_i \leq \Delta_i \leq U_i$ by construction, and further,

$$\begin{aligned} U_i - L_i &= \sup_{u,l \in \mathcal{X}_i} \mathbb{E}[g(X_1, \ldots, X_{i-1}, u, X_{i+1}, \ldots, X_n) - g(X_1, \ldots, X_{i-1}, l, X_{i+1}, \ldots, X_n) \mid \mathcal{F}_{i-1}] \\ &\leq \mathbb{E}[c_i \mid \mathcal{F}_i] = c_i, \end{aligned}$$

since g satisfies the bounded differences condition for constants $c_1, \ldots, c_n \in \mathbb{R}$. Finally, observe that

$$\sum_{i=1}^{n} \Delta_i = g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)].$$

Hence, we can apply the Azuma-Hoeffding inequality (Theorem 3.3) to obtain

$$\mathbb{P}(g(X_1, \ldots, X_n) - \mathbb{E}[g(X_1, \ldots, X_n)] \geq t) = \mathbb{P}\left(\sum_{i=1}^{n} \Delta_i \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right). \qquad \square$$

Observe that, assuming $X_1, \ldots, X_n$ are independent random variables such that $a_i \leq X_i \leq b_i$ (almost surely), we can recover Hoeffding's inequality (Theorem 2.12) by applying Theorem 3.5 to $g(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$, since, in this case, g satisfies the bounded differences condition for constants $c_i = b_i - a_i$.

Another drawback of Hoeffding's inequality is that it completely ignores the random variables' variances[2]. When the random variables' variances are known, an ideal concentration inequality should capture the idea that the variance of a random variable is a measure of concentration to some degree, and thus should include it in the inequality. This is exactly what Bernstein's inequality does. Most importantly, when the variance of the random variables involved is small, Bernstein's inequality provides a sharper bound than Hoeffding's inequality.

Theorem 3.7 (Bernstein's Inequality). *Let $X_1, \ldots, X_n$ be independent, centered random variables such that $|X_i| \leq c$ for $i = 1, \ldots, n$. Then,*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(\frac{-\frac{1}{2}t^2}{\sum_{i=1}^{n} \mathbb{E}[X_i^2] + \frac{1}{3}ct}\right).$$

We aim to make use of the fact that we know the random variables' variances to obtain an improved bound on their moment generating functions. Once we have this, we can apply the generic Chernoff bound as in the proof of Hoeffding's inequality to obtain the desired result.

Lemma 3.8. *Let X be a centered random variable satisfying $|X| \leq c$. For any $s > 0$, the moment generating function of X satisfies*

$$\mathbb{E}[\mathrm{e}^{sX}] \leq \exp\left(s^2 \, \mathbb{E}[X^2]\left(\frac{\mathrm{e}^{sc} - 1 - sc}{(sc)^2}\right)\right).$$

Proof. First, observe that

$$\mathbb{E}[X^k] \leq \mathbb{E}[X^2 |X|^{k-2}] \leq \mathbb{E}[X^2] c^{k-2}, \quad k \geq 2.$$

[2] However, this also makes Hoeffding's inequality so powerful, because it assumes so little about the random variables that are involved!

Hence, we have

$$\sum_{k=2}^{\infty} \frac{s^{k-2}\,\mathbb{E}[X^k]}{\mathbb{E}[X^2]k!} \leq \sum_{k=2}^{\infty} \frac{s^{k-2}\,\mathbb{E}[X^2]c^{k-2}}{\mathbb{E}[X^2]k!} = \frac{1}{(sc)^2} \sum_{k=2}^{\infty} \frac{(sc)^k}{k!} = \frac{\mathrm{e}^{sc} - 1 - sc}{(sc)^2},$$

and thus

$$\mathbb{E}[\mathrm{e}^{sX}] = \mathbb{E}\left[1 + sX + \sum_{k=2}^{\infty} \frac{s^k X^k}{k!}\right] = 1 + s^2\,\mathbb{E}[X^2] \sum_{k=2}^{\infty} \frac{s^{k-2}\,\mathbb{E}[X^k]}{\mathbb{E}[X^2]k!} \leq 1 + s^2\,\mathbb{E}[X^2]\,\frac{\mathrm{e}^{sc} - 1 - sc}{(sc)^2},$$

since $\mathbb{E}[sX] = s\,\mathbb{E}[X] = 0$. Finally, by applying the inequality $1 + x \leq \mathrm{e}^x$, we obtain

$$\mathbb{E}[\mathrm{e}^{sX}] \leq 1 + s^2\,\mathbb{E}[X^2]\frac{\mathrm{e}^{sc} - 1 - sc}{(sc)^2} \leq \exp\left(s^2\,\mathbb{E}[X^2]\left(\frac{\mathrm{e}^{sc} - 1 - sc}{(sc)^2}\right)\right). \qquad \square$$

Equipped with this result, we can prove Bernstein's inequality.

Proof of Theorem 3.7. As in the proof of Hoeffding's inequality, we have

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \mathrm{e}^{-st} \prod_{i=1}^{n} \mathbb{E}[e^{sX_i}] \leq \mathrm{e}^{-st} \exp\left(s^2 \sum_{i=1}^{n} \mathbb{E}[X_i^2]\left(\frac{\mathrm{e}^{sc} - 1 - sc}{(sc)^2}\right)\right), \quad s > 0,$$

where the second inequality is a consequence of Lemma 3.8. The RHS can be written as

$$\exp\left(s^2\left(\sum_{i=1}^{n} \mathbb{E}[X_i^2]\right) g(s) - st\right), \quad g(s) = \frac{\mathrm{e}^{sc} - 1 - sc}{(sc)^2},$$

and optimizing this expression over $s > 0$ yields the desired result. $\square$

4. Vapnik-Chervonenkis (VC) Theory

Once again, recall the general setup of Section 2.3: given a family $\mathcal{H}$ of classifiers, we can express the excess risk of the empirical risk minimizer $\hat{h}$ as

$$R(\hat{h}) = \underbrace{L(\hat{h}) - L(\bar{h})}_{\substack{\text{estimation} \\ \text{error}}} + \underbrace{L(\bar{h}) - L^*}_{\substack{\text{approximation} \\ \text{error}}}$$

where $\bar{h} \in \mathcal{H}$ is the oracle, i.e., the classifier in $\mathcal{H}$ that minimizes the true risk over $\mathcal{H}$. The second term on the RHS is the approximation error, which remains fixed once we have settled on a family $\mathcal{H}$. Hence, we try to find bounds of the first term of the RHS, the estimation error. In Theorem 2.16, we had seen that the inequality

$$L(\hat{h}) - L(\bar{h}) < \sqrt{\frac{2\log(2M/\delta)}{n}}$$

holds with probability at least $1 - \delta$, if $\mathcal{H}$ is a *finite* family of M classifiers. Note that this upper bound for finite $\mathcal{H}$ cannot simply be extended to the case that $\mathcal{H}$ is *infinite*. Essentially, to extend our previous results to the infinite case, we would have to show that only finitely many elements in a possibly infinite dictionary $\mathcal{H}$ "really matter". This exactly is the goal of the theory developed by Russian mathematicians Vladimir Vapnik and Alexey Chervonenkis from 1960 to 1990.

We start with the unrealistic case that our family $\mathcal{H}$ of classifiers is finite and that the oracle $\bar{h} \in \mathcal{H}$ has zero error probability. In that case, the empirical risk minimizer $\hat{h}$ satisfies $\hat{L}_n(\hat{h}) = 0$ almost surely, which can be seen as follows: Since $\hat{h}$ is the classifier in $\mathcal{H}$ minimizing the empirical risk, we have

$$\mathbb{E}[\hat{L}_n(\hat{h})] = \mathbb{E}[\min_{h\in\mathcal{H}} \hat{L}_n(h)] \leq \min_{h\in\mathcal{H}} \mathbb{E}[\hat{L}_n(h)] = \min_{h\in\mathcal{H}} L(h) = L(\bar{h}) = 0.$$

Because $\hat{L}_n(\hat{h})$ is non-negative, this implies that $\mathbb{P}(\hat{L}_n(\hat{h}) = 0) = 1$. We can also bound the *true* error of the ERM as follows:

Theorem 4.1 (Vapnik and Chervonenkis, 1974). *Let $\mathcal{H}$ be finite and assume $L(\bar{h}) = \min_{h\in\mathcal{H}} L(h) = 0$. Then, for every $n \in \mathbb{N}$, the empirical risk minimzer $\hat{h}$ satisfies*

$$\mathbb{P}(L(\hat{h}) > \varepsilon) \leq |\mathcal{H}|\mathrm{e}^{-n\varepsilon}, \quad \varepsilon > 0$$

and

$$\mathbb{E}[L(\hat{h})] \leq \frac{1 + \log(|\mathcal{H}|)}{n}.$$

Proof. Since $\hat{L}_n(\hat{h}) = 0$ almost surely, we have $\hat{h} \in \mathcal{H}_0 = \{h \in \mathcal{H} \mid \hat{L}_n(h) = 0 \text{ a.s.}\}$, and hence

$$\begin{aligned}\mathbb{P}(L(\hat{h}) > \varepsilon) &\leq \mathbb{P}\left(\max_{h \in \mathcal{H}_0} L(h) > \varepsilon\right) = \mathbb{E}\left[\mathbf{1}_{\{\max_{h \in \mathcal{H}_0} L(h) > \varepsilon\}}\right] \\ &= \mathbb{E}\left[\max_{h \in \mathcal{H}} \mathbf{1}_{\{\hat{L}_n(h)=0\}} \mathbf{1}_{\{L(h)>\varepsilon\}}\right] \leq \sum_{h \in \mathcal{H},\, L(h)>\varepsilon} \mathbb{P}(\hat{L}_n(h) = 0).\end{aligned}$$

Since the probability $\mathbb{P}(\hat{L}_n(h) = 0)$ that no (X_i, Y_i) falls in the set $\{(x, y) \mid h(x) \neq y\}$ is at most $(1 - \varepsilon)^n$ if the probability of this set[1] is larger than ε, we conclude

$$\mathbb{P}(L(\hat{h}_n) > \varepsilon) \leq \sum_{h \in \mathcal{H},\, L(h)>\varepsilon} \mathbb{P}(\hat{L}_n(h) = 0) \leq |\mathcal{H}|(1 - \varepsilon)^n \leq |\mathcal{H}| \mathrm{e}^{-n\varepsilon}.$$

To bound the expected error probability of $\hat{h}$ for $n \in \mathbb{N}$, first recall that the expected value of a non-negative random variable $Z \geq 0$ can be computed as follows:

$$\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(X > t)\, \mathrm{d}t.$$

Hence, we have

$$\begin{aligned}\mathbb{E}[L(\hat{h})] &= \int_0^\infty \mathbb{P}(L(\hat{h}) > t)\, \mathrm{d}t \\ &\leq u + \int_u^\infty \mathbb{P}(L(\hat{h}) > t)\, \mathrm{d}t \leq u + |\mathcal{H}| \int_u^\infty \mathrm{e}^{-nt}\, \mathrm{d}t = u + |\mathcal{H}| \frac{\mathrm{e}^{-nu}}{n}.\end{aligned}$$

Since $u > 0$ was arbitrary, we can minimize the RHS to obtain the optimal solution $u^* = \log(|\mathcal{H}|)/n$, giving the desired result. □

4.1. Empirical Measure

We want to derive bounds of the estimation error similar to those established in Section 2.5, but for *infinite* dictionaries $\mathcal{H}$. Recall from Vapnik's lemma (Lemma 2.14) that the key quantity we need to control in order to do so is

$$\sup_{h \in \mathcal{H}} |\hat{L}_n(h) - L(h)|.$$

The goal of bounding this quantity leads to the study of uniform deviations of relative frequencies from their theoretical probabilities. By definition, we have

$$\hat{L}_n(h) - L(h) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{h(X_i) \neq Y_i\}} - \mathbb{P}(h(X) \neq Y).$$

Write $Z = (X, Y)$ and similarly $Z_i = (X_i, Y_i)$ for $i = 1, \ldots, n$. We define two measures μ and μ_n on the space $\mathcal{X} \times \{0, 1\}$ by

$$\mu(A) = \mathbb{P}(Z \in A) \qquad \text{and} \qquad \mu_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{Z_i \in A\}}.$$

[1] This is precisely the true loss $L(h) = \mathbb{P}(h(X) \neq Y)$, which is larger than ε by assumption.

Further, for a classifier $h \in \mathcal{H}$, let A_h be defined by[2]

$$A_h = \{(x, y) \in \mathcal{X} \times \{0, 1\} \mid h(x) \neq y\}. \tag{4.1}$$

By definition of A_h, we have

$$\{Z \in A_h\} = \{h(X) \neq Y\} \qquad \text{and} \qquad \{Z_i \in A_h\} = \{h(X_i) \neq Y_i\},$$

resulting in

$$\mu(A_h) = \mathbb{P}(Z \in A_h) = \mathbb{P}(h(X) \neq Y) = L(h)$$

and

$$\mu_n(A_h) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Z_i \in A_h\}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{h(X_i) \neq Y_i\}} = \hat{L}_n(h).$$

Setting

$$\mathcal{A} = \{A_h \mid h \in \mathcal{H}\}, \tag{4.2}$$

we see that

$$\sup_{h \in \mathcal{H}} |\hat{L}_n(h) - L(h)| = \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|.$$

The law of large numbers tells us that, for a measurable set $A \in \mathcal{A}$,

$$\mu_n(A) \xrightarrow{\text{a.s.}} \mu(A),$$

given that the Z_i are i.i.d. draws from the distribution of Z. Further, as we can write

$$\mu_n(A) - \mu(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Z_i \in A\}} - \mathbb{P}(Z \in A) = \frac{1}{n} \sum_{i=1}^{n} \big(\mathbf{1}_{\{Z_i \in A\}} - \mathbb{E}[\mathbf{1}_{\{Z_i \in A\}}]\big),$$

Hoeffding's inequality (Theorem 2.12) bounds the probability that the two measures differ on a measurable set $A \in \mathcal{A}$ by at least $t > 0$ as follows (cf. Theorem 2.15):

$$\mathbb{P}(|\mu_n(A) - \mu(A)| \geq t) \leq 2\mathrm{e}^{-2nt^2}, \quad t > 0.$$

Additionally, for finite $\mathcal{A}$, we have

$$\mathbb{P}(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \geq t) \leq 2|\mathcal{A}|\mathrm{e}^{-2nt^2}, \quad t > 0$$

by applying a simple union bound (again, cf. Theorem 2.15). However, for infinite $\mathcal{A}$ (i.e., infinite dictionaries $\mathcal{H}$), the most powerful tools to attack these problems are distribution-free large deviation type inequalities proved by Vapnik and Chervonenkis (1971).

Finally, observe that, for every collection $\mathcal{A}$ of measurable sets $A \subset \mathcal{X} \times \{0, 1\}$, the function

$$(z_1, \ldots, z_n) \mapsto \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$$

[2] Note that A_h coincides with $(h^{-1}(1) \times \{0\}) \cup (h^{-1}(0) \times \{1\})$, which is clearly measurable, assuming that $h \colon \mathcal{X} \to \{0, 1\}$ is measurable.

satisfies the bounded differences condition for constants $c_i = \frac{1}{n}$. Hence, by the bounded differences inequality (Theorem 3.5), we have

$$\mathbb{P}(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| - \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] < t) \geq 1 - \mathrm{e}^{-2nt^2}.$$

Evaluating this inequality at $t = \sqrt{\log(\delta^{-1})/2n}$, we see that, with probability at least $1 - \delta$,

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| < \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] + \sqrt{\frac{\log(\delta^{-1})}{2n}}. \tag{4.3}$$

In the next section, we will focus on bounding the first term on the RHS. To do so, we will introduce a technique called symmetrization.

Remark 4.2. *The results discussed in this section are* not *restricted to our particular setup of binary classification with $\mathcal{Z} = \mathcal{X} \times \{0,1\}$ and $Z = (X,Y)$ a feature-label pair, with $X\colon \Omega \to \mathcal{X}$ and $Y\colon \Omega \to \{0,1\}$. Instead, for* any *random variable $W\colon \Omega \to \mathcal{Z}$ defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, we can let μ be the pushforward measure of $\mathbb{P}$ with respect to W, i.e.,*

$$\mu(A) = \mathbb{P}_W(A) = \mathbb{P}(W \in A),$$

and let μ_n be the empirical measure associated with μ, i.e.,

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{W_i\in A\}},$$

where $W_1, \ldots, W_n$ are i.i.d. draws from the distribution of W. We can then apply the results discussed above to any family of measurable sets $A \subset \mathcal{Z}$.

4.2. Symmetrization and Rademacher Complexity

Symmetrization is a commonly used technique in machine learning. Let $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$ be our sample set. We take another independent copy of the sample set, which we denote $\mathcal{D}'_n = \{Z'_1, \ldots, Z'_n\}$. In machine learning, this sample is sometimes referred to as a "ghost sample". We have

$$\mu(A) = \mathbb{P}(Z \in A) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Z'_i\in A\}}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Z'_i\in A\}} \,\middle|\, \mathcal{D}_n\right] = \mathbb{E}[\mu'_n(A) \mid \mathcal{D}_n],$$

where $\mu'_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Z'_i\in A\}}$, since $Z'_1, \ldots, Z'_n$ are identically distributed as Z, and are independent of $\mathcal{D}_n$. Hence,

$$\begin{aligned}\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] &= \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mathbb{E}[\mu'_n(A) \mid \mathcal{D}_n]|] \\ &= \mathbb{E}[\sup_{A\in\mathcal{A}} |\mathbb{E}[\mu_n(A) - \mu'_n(A) \mid \mathcal{D}_n]|] \leq \mathbb{E}[\sup_{A\in\mathcal{A}} \mathbb{E}[|\mu_n(A) - \mu'_n(A)| \mid \mathcal{D}_n]],\end{aligned}$$

where we have used the fact that $\mu_n(A)$ is $\sigma(Z_1, \dots, Z_n)$-measurable, and then applied Jensen's inequality to the conditional expectation. Clearly, $|\mu_n(A) - \mu'_n(A)| \leq \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu'_n(A)|$ for all $A \in \mathcal{A}$, and thus,

$$\mathbb{E}[|\mu_n(A) - \mu'_n(A)| \mid \mathcal{D}_n] \leq \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu'_n(A)| \mid \mathcal{D}_n]$$

by monotonicity of conditional expectation. Taking the supremum over all $A \in \mathcal{A}$ and then applying monotonicity of expectation, we get

$$\mathbb{E}[\sup_{A\in\mathcal{A}} \mathbb{E}[|\mu_n(A) - \mu'_n(A)| \mid \mathcal{D}_n]] \leq \mathbb{E}[\sup_{A\in\mathcal{A}} \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu'_n(A)| \mid \mathcal{D}_n]].$$

The expression highlighted in blue no longer depends on $A \in \mathcal{A}$, so that we can drop the outer supremum to arrive at

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \leq \mathbb{E}[\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu'_n(A)| \mid \mathcal{D}_n]] = \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu'_n(A)|]$$

by the law of total expectation. By definition,

$$\mu_n(A) - \mu'_n(A) = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{1}_{\{Z_i\in A\}} - \mathbf{1}_{\{Z'_i\in A\}}).$$

Since our ghost sample $\mathcal{D}'_n$ has the same distribution as $\mathcal{D}_n$, by symmetry, $\mathbf{1}_{\{Z_i\in A\}} - \mathbf{1}_{\{Z'_i\in A\}}$ has the same distribution as $\sigma_i(\mathbf{1}_{\{Z_i\in A\}} - \mathbf{1}_{\{Z'_i\in A\}})$, where $\sigma_1, \dots, \sigma_n$ are i.i.d. $\mathrm{Rad}(1/2)$ random variables[3] that are independent of both samples $\mathcal{D}_n$ and $\mathcal{D}'_n$. Thus,

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \leq \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu'_n(A)|] = \mathbb{E}\left[\sup_{A\in\mathcal{A}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i(\mathbf{1}_{\{Z_i\in A\}} - \mathbf{1}_{\{Z'_i\in A\}})\right|\right].$$

By the triangle inequality,

$$\left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i(\mathbf{1}_{\{Z_i\in A\}} - \mathbf{1}_{\{Z'_i\in A\}})\right| \leq \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbf{1}_{\{Z_i\in A\}}\right| + \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbf{1}_{\{Z'_i\in A\}}\right|,$$

so that finally

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \leq 2\,\mathbb{E}\left[\sup_{A\in\mathcal{A}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbf{1}_{\{Z_i\in A\}}\right|\right], \tag{4.4}$$

where we have again used the fact that $\mathcal{D}_n$ and $\mathcal{D}'_n$ are identically distributed. Altogether, we have managed to bound $\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|]$ by a much nicer quantity. Nonetheless, we still want the upper bound to rely solely on the *structure* of $\mathcal{A}$ and *not* on the particular random sample $\{Z_1, \dots, Z_n\}$. To achieve this, we simply take the supremum over all possible observations $z_i \in \mathcal{X} \times \{0, 1\}$.

Definition 4.3. *The* Rademacher complexity of a family of sets $\mathcal{A}$ *in a space* $\mathcal{Z}$ *is the quantity*

$$\mathfrak{R}_n(\mathcal{A}) = \sup_{z\in\wp(\mathcal{Z})} \mathbb{E}\left[\sup_{A\in\mathcal{A}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbf{1}_{\{z_i\in A\}}\right|\right],$$

[3] i.e., $\mathbb{P}(\sigma_i = \pm 1) = 1/2$

where the supremum ranges over all subsets $z = \{z_1, \dots, z_n\} \in \wp(\mathcal{Z})$*, and* $\sigma_1, \dots, \sigma_n$ *are i.i.d.* $\mathrm{Rad}(1/2)$ *random variables. The* Rademacher complexity of a set $B \subset \mathbb{R}^n$ *is defined as*

$$\mathfrak{R}_n(B) = \mathbb{E}\left[\sup_{b \in B} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i b_i \right| \right],$$

where $\sigma_1, \dots, \sigma_n$ *are defined as above.*

The Rademacher complexity measures the complexity of a set $B \subset \mathbb{R}^n$ in the following sense: The quantity $\left|\frac{1}{n}\sum_{i=1}^n \sigma_i b_i\right|$ measures how well a vector $b \in B$ correlates with a random sign pattern $(\sigma_1, \dots, \sigma_n)$. The more complex B is, the better some vector $b \in B$ can replicate a given sign pattern. For example, if $B = [-1, 1]^n$, then $\mathfrak{R}_n(B) = 1$. If, instead, $B \subset [-1, 1]^n$ consists only of k-sparse vectors, then $\mathfrak{R}_n(B) = k/n$.

The definition of Rademacher complexity allows us to restate inequality (4.4) as follows:

$$\mathbb{E}[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|] \leq 2\mathfrak{R}_n(\mathcal{A}). \tag{4.5}$$

4.3. Shattering

The sets of vectors that interest us in the definition of Rademacher complexity of $\mathcal{A}$ are sets of the form

$$T(z) = \{(\mathbf{1}_{\{z_1 \in A\}}, \dots, \mathbf{1}_{\{z_n \in A\}})^\top \mid A \in \mathcal{A}\}, \tag{4.6}$$

for $z = \{z_1, \dots, z_n\} \in \wp(\mathcal{Z})$, where[4] $z_i \in \mathcal{X} \times \{0, 1\}$. Particularly, the cardinality of $T(z)$, i.e., the number of binary patterns a set of points $\{z_1, \dots, z_n\}$ can replicate as A ranges over the family of sets $\mathcal{A}$, will be extremely important, as it will arise when trying to control the Rademacher complexity of $\mathcal{A}$. A first hint at this is the following result, which allows us to link the Rademacher complexity of $\mathcal{A}$ to the Rademacher complexity of the sets $T(z)$:

Lemma 4.4. *The Rademacher complexity of a family of sets* $\mathcal{A}$ *in a space* $\mathcal{Z}$ *satisfies*

$$\mathfrak{R}_n(\mathcal{A}) = \sup_{z \in \wp(\mathcal{Z})} \mathfrak{R}_n(T(z)),$$

where the supremum ranges over all subsets $z = \{z_1, \dots, z_n\} \in \wp(\mathcal{Z})$.

Proof. This is straightforward: by the definition of the set $T(z)$ in (4.6), we have

$$\sup_{z \in \wp(\mathcal{Z})} \mathfrak{R}_n(T(z)) = \sup_{z \in \wp(\mathcal{Z})} \mathbb{E}\left[\sup_{b \in T(z)} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i b_i \right| \right] = \sup_{z \in \wp(\mathcal{Z})} \mathbb{E}\left[\sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \mathbf{1}_{\{z_i \in A\}} \right| \right] = \mathfrak{R}_n(\mathcal{A}). \qquad \square$$

Let us briefly investigate the set $T(z)$ and its cardinality to get a better understanding. First, even though the cardinality of $\mathcal{A}$ can be infinite, the cardinality of $T(z)$ is at most 2^n. To further illustrate the concept, let us look at an example: for the set $z = \{z_1, z_2, z_3, z_4\}$, the binary pattern $(1, 1, 0, 1)$ is contained in $T(z)$ if and

[4] Again, this definition makes sense for arbitrary spaces $\mathcal{Z}$, i.e., we can take a family of sets $\mathcal{A}$ and a set $z = \{z_1, \dots, z_n\} \in \wp(\mathcal{Z})$ and then define $T(z)$ as above.

only if there exists some set $A \in \mathcal{A}$ such that $z \cap A = \{z_1, z_2, z_4\}$, i.e., the set of points $\{z_1, z_2, z_4\}$ can be separated from the remaining points in z by a set A in $\mathcal{A}$. In this case, the set of points $\{z_1, z_2, z_4\}$ is said to be *picked up* by $\mathcal{A}$. Hence, the cardinality of $T(z)$ equals the number of elements of the power set $\wp(z)$ that can be picked up by $\mathcal{A}$, i.e.,

$$|T(z)| = |z \cap \mathcal{A}|,$$

where

$$z \cap \mathcal{A} = \{z \cap A \mid A \in \mathcal{A}\}.$$

This leads to the following definition:

Definition 4.5. *A family of sets* $\mathcal{A}$ shatters *the set of points* $z = \{z_1, \ldots, z_n\}$*, if*

$$|T(z)| = |z \cap \mathcal{A}| = 2^n.$$

Clearly, a family of sets $\mathcal{A}$ shatters the set of points $\{z_1, \ldots, z_n\}$ if and only if *every* subset of z can be picked up by $\mathcal{A}$. Recall that the set of points we are interested in are i.i.d. realizations $Z_1, \ldots, Z_n$ of $Z = (X, Y)$. As the distribution of Z is unknown, these realizations may theoretically take on any value over the sample space $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$. Hence, we define the *shatter coefficients* of a family of sets $\mathcal{A}$ as follows:

Definition 4.6. *The* n-th shatter coefficient $\mathcal{S}_{\mathcal{A}}(n)$ *of a family of sets* $\mathcal{A}$ *is given by*

$$\mathcal{S}_{\mathcal{A}}(n) = \sup_{z \in \wp(\mathcal{Z})} |T(z)| = \sup_{z \in \wp(\mathcal{Z})} |z \cap \mathcal{A}|,$$

where the suprema range over all subsets $z = \{z_1, \ldots, z_n\} \in \wp(\mathcal{Z})$*.*

By definition, $\mathcal{S}_{\mathcal{A}}(n) = 2^n$ if $\mathcal{A}$ shatters a set $\{z_1, \ldots, z_n\} \in \wp(\mathcal{Z})$, i.e., if there exists a set $z = \{z_1, \ldots, z_n\}$ such that every subset of z can be picked up by $\mathcal{A}$. Analogously, we have $\mathcal{S}_{\mathcal{A}}(n) < 2^n$ if *no* set of points $\{z_1, \ldots, z_n\}$ can be shattered by $\mathcal{A}$. The largest integer k for which there still exists a set that can be shattered by $\mathcal{A}$ is precisely the *Vapnik-Chervonenkis dimension* of $\mathcal{A}$:

Definition 4.7. *The* Vapnik-Chervonenkis dimension $\mathrm{VC}(\mathcal{A})$ *of a family of sets* $\mathcal{A}$ *is the largest integer* k *such that* $\mathcal{S}_{\mathcal{A}}(k) = 2^k$*. If* $\mathcal{S}_{\mathcal{A}}(n) = 2^n$ *holds for all* n*, we set* $\mathrm{VC}(\mathcal{A}) = \infty$*.*

Note that this definition only makes sense if $\mathcal{S}_{\mathcal{A}}(n) < 2^n$ implies $\mathcal{S}_{\mathcal{A}}(m) < 2^m$ for all $m > n$, which indeed is true. Hence, $\mathrm{VC}(\mathcal{A}) = d$ implies that for every integer $n = 1, \ldots, d$, there exists *at least one* set $\{z_1, \ldots, z_n\}$ that $\mathcal{A}$ shatters, but there exist *no* sets of cardinality $m > d$ that $\mathcal{A}$ shatters. We will see that the VC dimension will play a role similar to the cardinality of sets, but on an exponential scale.

Let us consider an example to get a better understanding of all of the terms introduced in this section. Take, for example, $\mathcal{A} = \{(-\infty, a] \mid a \in \mathbb{R}\} \cup \{[a, \infty) \mid a \in \mathbb{R}\}$ to be the set of half-lines, and let our sample space be the real line $\mathbb{R}$. Clearly, $|\mathcal{A}| = \infty$. As it turns out, we can shatter every set of two points $\{z_1, z_2\}$: WLOG, assume $z_1 < z_2$ and let $\varepsilon = (z_2 - z_1)/2 > 0$. Then,

- $A = (-\infty, z_1 - 1]$ picks up the empty set,
- $A = (-\infty, z_1 + \varepsilon]$ picks up the set $\{z_1\}$,

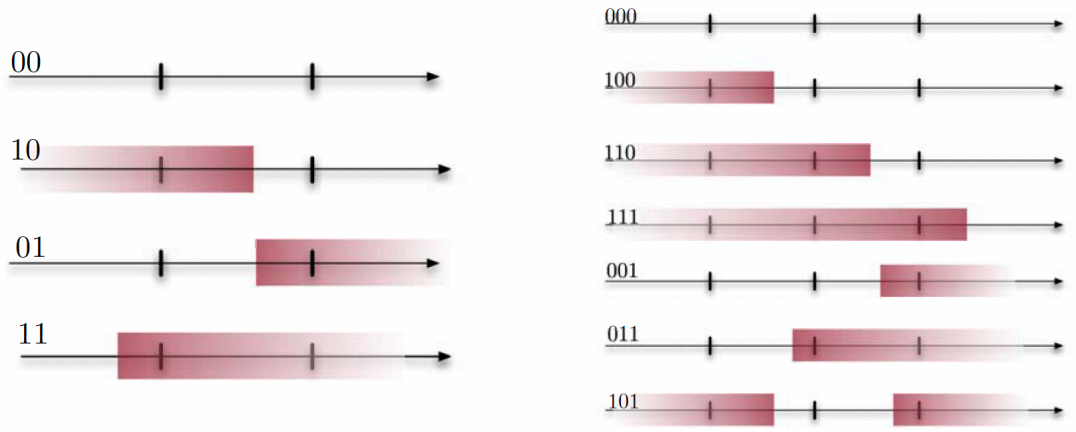

Figure 4.1. The family $\mathcal{A} = \{(-\infty, a] \mid a \in \mathbb{R}\} \cup \{[a, \infty) \mid a \in \mathbb{R}\}$ of half-lines shatters every set of cardinality 2, but *no* set of three points. In particular, $\mathrm{VC}(\mathcal{A}) = 2$. Note that the binary pattern 101 (bottom right) can only be created if one extends the family $\mathcal{A}$ to also include unions of half-lines.

Note. From "Mathematics of Machine Learning (18.657)," by P. Rigollet, Fall 2015, *Massachusetts Institute of Technology: MIT OpenCouseWare*, p. 29 (https://ocw.mit.edu/). CC BY-NC-SA 4.0.

- $A = [z_1 + \varepsilon, \infty)$ picks up the set $\{z_2\}$,
- $A = [z_1 - 1, \infty)$ picks up the set $\{z_1, z_2\}$.

In particular, $\mathcal{A}$ shatters the set $\{z_1, z_2\}$, and we can conclude that[5] $\mathcal{S}_{\mathcal{A}}(2) = 2^2$, and hence $\mathrm{VC}(\mathcal{A}) \geq 2$. However, $\mathcal{A}$ shatters *no* set of three points $\{z_1, z_2, z_3\}$: WLOG, we can assume $z_1 < z_2 < z_3$. In this case, the set $\{z_2\}$ cannot be picked up by $\mathcal{A}$! This is easy to see: let $A \in \mathcal{A}$ be a set such that $z_2 \in A$. This implies that $A = (-\infty, z_2 + \delta]$ or $A = [z_2 - \delta, \infty)$ with $\delta \geq 0$. However, we have $z_1 \in (-\infty, z_2 + \delta]$ and $z_3 \in [z_2 - \delta, \infty)$, i.e., it is impossible to pick up the singleton $\{z_2\}$ with half-lines! More generally, one can show that half-lines can only generate binary patterns with ones followed by zeros or zeros followed by ones but they cannot generate alternating patterns like 010 or 101. In particular, $\mathcal{S}_{\mathcal{A}}(n) = 2n < n^2$ for $n > 2$.

Let us discuss another example. Let $\mathcal{A}$ be the set of hyperrectangles in $\mathbb{R}^d$, i.e., $\mathcal{A} = \{[\mathbf{a}, \mathbf{b}] \mid \mathbf{a}, \mathbf{b} \in \mathbb{R}^d\}$, where

$$[\mathbf{a}, \mathbf{b}] = \prod_{i=1}^{d} [a^{(i)}, b^{(i)}] = \{(x^{(1)}, \ldots, x^{(d)}) \in \mathbb{R}^d \mid x^{(i)} \in [a^{(i)}, b^{(i)}] \text{ for } i = 1, \ldots, d\}$$

with $\mathbf{a} = (a^{(1)}, \ldots, a^{(d)})^\top$ and $\mathbf{b} = (b^{(1)}, \ldots, b^{(d)})^\top$ such that $a^{(i)} \leq b^{(i)}$ for all $i = 1, \ldots, d$. For $d = 2$, the family $\mathcal{A}$ shatters the set $\{(1, 0)^\top, (0, 1)^\top, (-1, 0)^\top, (0, -1)^\top\}$ of four points in $\mathbb{R}^2$, and, more generally, one can show that $\mathcal{A}$ always shatters a set of $2d$ points for arbitrary dimensions d. This tells us that $\mathcal{S}_{\mathcal{A}}(2d) = 2^{2d}$ and $\mathrm{VC}(\mathcal{A}) \geq 2d$. However, the family $\mathcal{A}$ can never shatter a set of $2d + 1$ points. The argument is similar in spirit to the one made earlier for half-lines. Let $B = \{x_1, \ldots, x_{2d+1}\}$ be a set of $2d + 1$ points and define points $y_1, \ldots, y_d$ and $z_1, \ldots, z_d$ as follows:

$$y_i = \operatorname*{argmin}_{x_j \in B} x_j^{(i)}, \qquad z_i = \operatorname*{argmax}_{x_j \in B} x_j^{(i)}, \qquad i = 1, \ldots, d.$$

With these definitions, any set that contains $\{y_1, \ldots, y_d, z_1, \ldots, z_d\}$, will inevitably contain the entire set B. Hence, it is impossible to pick up the set $\{y_1, \ldots, y_d, z_1, \ldots, z_d\}$ and $\mathcal{S}_{\mathcal{A}}(2d + 1) < 2^{2d+1}$. Consequently, $\mathrm{VC}(\mathcal{A}) = 2d$.

[5] Note that, by definition, $\mathcal{A}$ need only shatter a single set of two points for the identity $\mathcal{S}_{\mathcal{A}}(2) = 2^2$ to hold!

4.4. The VC Inequality

We have now introduced all the concepts required to formulate the main result of this chapter: the VC inequality.

Theorem 4.8 (VC Inequality). *Every family of sets $\mathcal{A}$ with VC dimension $\mathrm{VC}(\mathcal{A}) = D$ satisfies*

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \leq 2\sqrt{\frac{2D\log(2en/D)}{n}}.$$

We split the proof of this result into three steps:

(1) Using symmetrization, we prove that

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \leq 2\mathfrak{R}_n(\mathcal{A}).$$

Note that this is precisely the bound (4.5) obtained in Section 4.2.

(2) We bound the Rademacher complexity of $\mathcal{A}$ using shatter coefficients:

$$\mathfrak{R}_n(\mathcal{A}) \leq \sqrt{\frac{2\log(2\mathcal{S}_{\mathcal{A}}(n))}{n}}.$$

(3) Finally, we bound the shatter coefficients $\mathcal{S}_{\mathcal{A}}(n)$ by the VC dimension $\mathrm{VC}(\mathcal{A})$:

$$\mathcal{S}_{\mathcal{A}}(n) \leq \left(\frac{en}{D}\right)^D, \quad D = \mathrm{VC}(\mathcal{A}).$$

This is known as the *Sauer-Shelah* lemma.

Since the first step is already done, let's tackle step number 2. As an intermediate result, we prove a bound of the Rademacher complexity of finite subsets in $\mathbb{R}^n$. This will turn out to be immensely useful, as we can express the Rademacher complexity of a family of sets $\mathcal{A}$ in terms of the Rademacher complexity of finite subsets in $\mathbb{R}^n$, i.e., $\mathfrak{R}_n(\mathcal{A}) = \sup_{z\in\wp(\mathcal{Z})} \mathfrak{R}_n(T(z))$, where $T(z)$ is defined as in (4.6).

Lemma 4.9. *For a* finite *set $B \subset \mathbb{R}^n$, it holds*

$$\mathfrak{R}_n(B) \leq \max_{b\in B} \|b\|_2 \frac{\sqrt{2\log(2|B|)}}{n},$$

where $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbb{R}^d$.

Proof. By definition,

$$n\mathfrak{R}_n(B) = \mathbb{E}[\max_{b\in B} |Z_b|],$$

where $Z_b = \sum_{i=1}^n \sigma_i b_i$. Since $-|b_i| \leq \sigma_i b_i \leq |b_i|$ holds almost surely, Hoeffding's lemma (Lemma 2.11) bounds the moment generating function of Z_b by

$$\mathbb{E}[\exp(sZ_b)] = \prod_{i=1}^n \mathbb{E}[\exp(s\sigma_i b_i)] \leq \prod_{i=1}^n \exp\big(s^2 b_i^2/2\big) = \exp(s^2\|b\|_2^2/2). \tag{4.7}$$

To bound the quantity $\mathbb{E}[\max_{b\in B}|Z_b|]$ we are interested in, let $\overline{B} = B \cup -B$, where $-B = \{-b \mid b \in B\}$. For $s > 0$,

$$\mathbb{E}[\max_{b\in B}|Z_b|] = \mathbb{E}[\max_{b\in \overline{B}} Z_b] = s^{-1}\log(\exp(s\,\mathbb{E}[\max_{b\in\overline{B}} Z_b])) \le s^{-1}\log(\mathbb{E}[\exp(s\max_{b\in\overline{B}} Z_b)]),$$

where the last inequality follows from Jensen's inequality applied to the function $x \mapsto \exp(sx)$. Since the logarithm is increasing, we can bound $\mathbb{E}[\exp(s\max_{b\in\overline{B}} Z_b)] \le \sum_{b\in\overline{B}}\mathbb{E}[\exp(sZ_b)]$ to obtain

$$\begin{aligned}\mathbb{E}[\max_{b\in B}|Z_b|] &\le s^{-1}\log\left(\sum_{b\in\overline{B}}\mathbb{E}[\exp(sZ_b)]\right) \le s^{-1}\log\left(\sum_{b\in\overline{B}}\exp(s^2\|b\|_2^2/2)\right)\\ &\le \frac{\log(2|B|)}{s} + \frac{s}{2}\max_{b\in B}\|b\|_2^2 = \frac{2\log(2|B|) + s^2\max_{b\in B}\|b\|_2^2}{2s},\end{aligned}$$

where the second inequality follows from (4.7), and the third inequality follows from the observation that $\max_{b\in\overline{B}}\|b\|_2^2 = \max_{b\in B}\|b\|_2^2$ and $|\overline{B}| \le 2|B|$. Optimizing the RHS over s yields the optimal solution

$$s^* = \sqrt{\frac{2\log(2|B|)}{\max_{b\in B}\|b\|_2^2}}.$$

Plugging this back in yields the desired result, i.e.,

$$n\mathfrak{R}_n(B) = \mathbb{E}[\max_{b\in B}|Z_b|] \le \max_{b\in B}\|b\|_2\sqrt{2\log(2|B|)}.$$ □

By applying the previous result to the finite set $T(z)$, we obtain

$$\mathfrak{R}_n(T(z)) \le \max_{b\in T(z)}\|b\|_2\frac{\sqrt{2\log(2|T(z)|)}}{n}.$$

Since each entry of a vector in $T(z)$ takes values in the set $\{0,1\}$, we know that the Euclidean norm $\|b\|_2$ of any vector $b \in T(z)$ is at most $\sqrt{n}$. Further, we know that the shatter coefficients $\mathcal{S}_\mathcal{A}(n)$ of $\mathcal{A}$ depend directly on the cardinality of the sets $T(z)$. Hence, the next result should come at no surprise.

Proposition 4.10. *For a family of sets $\mathcal{A}$, it holds*

$$\mathfrak{R}_n(\mathcal{A}) \le \sqrt{\frac{2\log(2\mathcal{S}_\mathcal{A}(n))}{n}}.$$

Proof. Observe that $\mathfrak{R}_n(\mathcal{A}) = \sup_z \mathfrak{R}_n(T(z))$, by Lemma 4.4, where $T(z)$ is defined as in (4.6). Since $T(z) \subset \{0,1\}^n$, we have $\|b\|_2 \le \sqrt{n}$ for all $b \in T(z)$ and all subsets $z = \{z_1,\dots,z_n\} \in \wp(\mathcal{Z})$. Hence, by Lemma 4.9, we have

$$\mathfrak{R}_n(\mathcal{A}) \le \sup_{z\in\wp(\mathcal{Z})}\sqrt{\frac{2\log(2|T(z)|)}{n}}.$$

Finally, by the definition of the shatter coefficients of $\mathcal{A}$, we have $|T(z)| \le \sup_z|T(z)| = \mathcal{S}_\mathcal{A}(n)$, and hence

$$\mathfrak{R}_n(\mathcal{A}) \le \sqrt{\frac{2\log(2\mathcal{S}_\mathcal{A}(n))}{n}}.$$ □

Up to this point, we have shown that

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \leq 2\mathfrak{R}_n(\mathcal{A}) \leq 2\sqrt{\frac{2\log(2\mathcal{S}_{\mathcal{A}}(n))}{n}}. \tag{4.8}$$

Note that this bound would not be informative (in the sense that it does not imply convergence of the uniform deviations to zero as the sample size n goes to infinity) if the shatter coefficients $\mathcal{S}_{\mathcal{A}}(n)$ were exponential in n. For example, if $\mathcal{S}_{\mathcal{A}}(n) = 2^n$ for $n \leq D$ and $\mathcal{S}_{\mathcal{A}}(n) = 2^n - 1$ for all $n > D$, the RHS of (4.8) would be greater than 2 for all n. The VC inequality suggests that this cannot be the case, and indeed, if the VC dimension of a family of sets is *finite*, the shatter coefficients of $\mathcal{A}$ can be at most *polynomial* in n. This result is known as the Sauer-Shelah lemma, which we now turn to.

Lemma 4.11 (Sauer-Shelah). *For a family of sets $\mathcal{A}$ with finite VC dimension $\mathrm{VC}(\mathcal{A}) = D$, the shatter coefficients satisfy*

$$\mathcal{S}_{\mathcal{A}}(n) \leq \sum_{k=0}^{D} \binom{n}{k} \leq \left(\frac{en}{D}\right)^D, \quad n \in \mathbb{N}.$$

For a proof of this result, we refer the reader to Theorem 13.2 of Section 13.1 in [DGL96, p. 216]. Replacing $\mathcal{S}_{\mathcal{A}}(n)$ with $(en/D)^D$ in (4.8) clearly yields the VC inequality

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \leq 2\sqrt{\frac{2D\log(2en/D)}{n}}.$$

Recall that, using the bounded differences inequality (Theorem 3.5), we had shown that the bound

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| < \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] + \sqrt{\frac{\log(\delta^{-1})}{2n}}$$

holds with probability at least $1 - \delta$. Hence, we conclude:

Corollary 4.12 (VC Inequality). *For a family of sets $\mathcal{A}$ with finite VC dimension $\mathrm{VC}(\mathcal{A}) = D$, the upper bound*

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| < 2\sqrt{\frac{2D\log(2en/D)}{n}} + \sqrt{\frac{\log(\delta^{-1})}{2n}}$$

holds with probability at least $1 - \delta$.

Note that, while this bound can be improved, it is always of rate $\sqrt{\log(n)/n}$, which is a "slow rate".

4.5. Application to the ERM

Recall our goal that we started with at the beginning of this chapter: given a family $\mathcal{H}$ of classifiers, we wanted to bound the excess risk $R(\hat{h})$ of the empirical risk minimizer $\hat{h}$, which can be decomposed into estimation error and approximation error. Since the approximation error is fixed for a given family of classifiers $\mathcal{H}$, we have focused on bounding the estimation error $L(\hat{h}) - L(\bar{h})$. In Lemma 2.14, we had already observed that the estimation error is bounded by

$$L(\hat{h}) - L(\bar{h}) \leq 2\sup_{h\in\mathcal{H}} |\hat{L}_n(h) - L(h)|.$$

Defining A_h and $\mathcal{A}$ as in (4.1) and (4.2), respectively, and letting $\mu_n(A) = n^{-1}\sum_{i=1}^n \mathbf{1}_{\{Z_i \in A\}}$ and $\mu(A) = \mathbb{P}(Z \in A)$, we had observed that $\sup_{h\in\mathcal{H}} |\hat{L}_n(h) - L(h)| = \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|$. Hence, the VC inequality tells us that

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} |\hat{L}_n(h) - L(h)| < 2\sqrt{\frac{2D\log(2en/D)}{n}} + \sqrt{\frac{\log(\delta^{-1})}{2n}}\right) \geq 1-\delta, \tag{4.9}$$

where D denotes the VC dimension of $\mathcal{A} = \{A_h \mid h \in \mathcal{H}\}$. However, the VC dimension of this class $\mathcal{A}$ is not very natural, and in many cases it is more convenient to consider the class

$$\bar{\mathcal{A}} = \{\bar{A}_h \mid h \in \mathcal{H}\}, \quad \bar{A}_h = \{x \in \mathcal{X} \mid h(x) = 1\}. \tag{4.10}$$

A priori it is not clear, how the VC dimension of the class $\mathcal{A}$ and the VC dimension of the class $\bar{\mathcal{A}}$ are related, if at all. However, as the next result shows, these two quantities actually coincide.

Theorem 4.13. *Let $\mathcal{H}$ be a family of classifiers and let $\mathcal{A}$ and $\bar{\mathcal{A}}$ be defined as in (4.2) and (4.10), respectively. Then,*

$$\mathcal{S}_{\mathcal{A}}(n) = \mathcal{S}_{\bar{\mathcal{A}}}(n), \quad n \geq 1,$$

which implies that $\mathrm{VC}(\mathcal{A}) = \mathrm{VC}(\bar{\mathcal{A}})$.

Proof. Recall that the n-th shatter coefficent $\mathcal{S}_{\mathcal{A}}(n)$ is defined by $\mathcal{S}_{\mathcal{A}}(n) = \sup_{z\in\wp(\mathcal{Z})} |T(z)|$, where $T(z)$ is the set of binary patterns that can be generated by the set $z = \{z_1, \ldots, z_n\} \in \wp(\mathcal{Z})$, i.e.,

$$T(z) = \{(\mathbf{1}_{\{z_1\in A\}}, \ldots, \mathbf{1}_{\{z_n\in A\}})^\top \mid A \in \mathcal{A}\}, \quad z_i = (x_i, y_i) \in \mathcal{X}\times\{0,1\}.$$

By definition of the collection of sets $\mathcal{A}$ in (4.2), we can rewrite this as

$$T(z) = \{(\mathbf{1}_{\{h(x_1)\neq y_1\}}, \ldots, \mathbf{1}_{\{h(x_n)\neq y_n\}})^\top \mid h \in \mathcal{H}\}.$$

Similarly, let $\bar{T}(z)$ denote the set of binary patterns generated by the set $z \in \wp(\mathcal{Z})$ for the collection of sets $\bar{\mathcal{A}}$ defined in (4.10), i.e.,

$$\bar{T}(z) = \{(\mathbf{1}_{\{h(x_1)=1\}}, \ldots, \mathbf{1}_{\{h(x_n)=1\}})^\top \mid h \in \mathcal{H}\}.$$

Next, we fix $v \in \{0,1\}$ and let $u \oplus v$ denote the logical XOR operation applied to $u \in \{0,1\}$, i.e.,

$$\oplus\colon \{0,1\} \to \{0,1\}, \quad u \mapsto u \oplus v = \mathbf{1}_{\{u\neq v\}}.$$

The XOR operation is an involution, i.e., it satisfies $(u \oplus v) \oplus v = u$. In particular, the XOR operation is bijective. By applying the XOR operation to each entry of a vector, we have

$$(\mathbf{1}_{\{h(x_1)\neq y_1\}}, \ldots, \mathbf{1}_{\{h(x_n)\neq y_n\}})^\top = (\mathbf{1}_{\{h(x_1)=1\}}, \ldots, \mathbf{1}_{\{h(x_n)=1\}})^\top \oplus (y_1, \ldots, y_n)^\top.$$

Since the XOR operation is bijective, this tells us that the cardinalities of $T(z)$ and $\bar{T}(z)$ coincide. Consequently, so do the shatter coefficients and the VC dimension of the collections $\mathcal{A}$ and $\bar{\mathcal{A}}$. □

Based on our discussion of the emprical risk minimzer at the beginning of this section, we conclude the

following:

Corollary 4.14. *Let $\mathcal{H}$ be a family of classifiers such that the family $\bar{\mathcal{A}}$ defined in (4.10) has VC dimension D. Then, the empirical risk minimizer $\hat{h}$ of $\mathcal{H}$ satisfies*

$$L(\hat{h}) < L(\bar{h}) + 4\sqrt{\frac{2D\log(2en/D)}{n}} + 2\sqrt{\frac{\log(\delta^{-1})}{2n}}$$

with probability at least $1-\delta$.

4.6. A "Fast Rate" VC Inequality

While the VC inequality stated in Section 4.4 was of rate $\sqrt{\log(n)/n}$ (i.e., "slow"), there also exists an upper bound that decreases exponentially fast.

Theorem 4.15 (Vapnik-Chervonenkis, 1971)**.** *For a probability measure μ and a collection of measurable sets $\mathcal{A}$, it holds*

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| > \varepsilon\right) \leq 8\mathcal{S}_{\mathcal{A}}(n)\mathrm{e}^{-n\varepsilon^2/32}$$

for all $n \geq 1$ and $\varepsilon > 0$.

The proof of this statement (which we omit) is similar to the proof of an important result by Glivenko-Cantelli (see below), both of which can be found in Chapter 12 of [DGL96].

Theorem 4.16 (Fundamental Theorem of Statistics [Glivenko-Cantelli, 1933])**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables with distribution function $F(t)$ and denote the corresponding empirical distribution function by $F_n(t)$. Then,*

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| > \varepsilon\right) \leq 8(n+1)\mathrm{e}^{-n\varepsilon^2/32}.$$

In particular, by the Borel-Cantelli lemma,

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow{\text{a.s.}} 0.$$

Remark 4.17. *While the proofs of the two statements are quite similar, the fast rate VC inequality (Theorem 4.15) also directly implies the result by Glivenko-Cantelli (Theorem 4.16): for $t \in \mathbb{R}$, define $A_t = (-\infty, t]$ and let $\mathcal{A} = \{A_t \mid t \in \mathbb{R}\}$. Clearly, $\mathcal{A}$ is a collection of measurable sets. Further, it is easy to see that $\mathcal{A}$ can only generate binary patterns of 1s followed by 0s. Hence, the n-th shatter coefficient $\mathcal{S}_{\mathcal{A}}(n)$ of $\mathcal{A}$ is given by $n+1$. Finally, we define a measure μ by $\mu(A) = \mathbb{P}(X \in A)$ for any measurable set A (not necessarily in the collection $\mathcal{A}$), where $X \sim F$. In that case, we have $\mu(A_t) = \mathbb{P}(X \in A_t) = \mathbb{P}(X \leq t) = F(t)$ and $\mu_n(A_t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i \in A_t\}} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i \leq t\}} = F_n(t)$. Altogether,*

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| > \varepsilon\right) = \mathbb{P}\left(\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| > \varepsilon\right) \leq 8\mathcal{S}_{\mathcal{A}}(n)\mathrm{e}^{-n\varepsilon^2/32} = 8(n+1)\mathrm{e}^{-n\varepsilon^2/32}.$$

5. Learning with a General Loss Function

So far, in chapters 2 and 4, we have solely focused on the problem of binary classification, i.e., we have studied classifiers $h\colon \mathcal{X} \to \{0,1\}$ and their performance. To gauge the performance of these classifiers we introduced the binary loss function $\mathbf{1}_{\{h(X)\neq Y\}}$ and measured the risk of a classifier h by its misclassification probability

$$L(h) = \mathbb{E}[\mathbf{1}_{\{h(X)\neq Y\}}] = \mathbb{P}(h(X) \neq Y).$$

We then split the excess risk of a classifier h into estimation error and approximation error as follows:

$$R(h) = \underbrace{L(h) - L(\bar{h})}_{\substack{\text{estimation}\\ \text{error}}} + \underbrace{L(\bar{h}) - L^*}_{\substack{\text{approximation}\\ \text{error}}}.$$

Finally, we used concentration inequalities and VC theory to find bounds for the estimation error $L(\hat{h}) - L(\bar{h})$ of the empirical risk minimizer $\hat{h}$. Let's look at some of the limitations that these techniques carry with them:

- *Hoeffding's inequality*: Only useful for finite families $\mathcal{H}$ of classifiers, requires boundedness of the loss function.
- *Bounded differences inequality*: Suitable for infinite families $\mathcal{H}$ of classifiers, also requires boundedness of the loss function.
- *VC theory*: Requires binary nature of the loss function.

In this chapter, we replace the binary output variable Y with a *continuous* output Y that takes values in the interval $[-1,1]$, and we replace the binary loss function $\mathbf{1}_{\{h(X)\neq Y\}}$ with a smooth (and symmetric) loss function $l(y_1, y_2)$ that we assume to be *bounded*, i.e. $0 \leq l(y_1, y_2) \leq 1$. Some examples are:

- $l(y_1, y_2) = |y_1 - y_2|$ (absolute loss)
- $l(y_1, y_2) = (y_1 - y_2)^2$ (squared loss)
- $l(y_1, y_2) = |y_1 - y_2|^p$ (L^p-loss)

Let's review the notation we will stick to during this chapter (which will largely remain unchanged compared to chapters 2 and 4): From now on, we will denote the *regression function* by $f\colon \mathcal{X} \to [-1,1]$ whose *expected error* is given by

$$L(f) = \mathbb{E}[l(f(X), Y)].$$

As always, our *dataset* will be denoted $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. The *empirical error* of f is defined as

$$\hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} l(f(X_i), Y_i),$$

which is a random variable! As before, we denote the *empirical risk minimizer* of a class of regression functions $\mathcal{F}$ by $\hat{f}_n$, where n denotes the number of observations (X_i, Y_i). For convenience, we often drop the subscript n and simply write $\hat{f}$. Finally, the *oracle* of the class $\mathcal{F}$ (i.e., the regression function minimizing the true error $L(f)$) will be denoted by $\bar{f}$.

5.1. Empirical Risk Minimization

As in chapters 2 and 4, we want to control the estimation error of the empirical risk minimizer $\hat{f}$. Using the same arguments as before, by Lemma 2.14, we can bound this error somewhat crudely as follows:

$$L(\hat{f}) - L(\bar{f}) \leq 2 \sup_{f \in \mathcal{F}} |\hat{L}_n(f) - L(f)|.$$

The function

$$(z_1, \dots, z_n) \mapsto \sup_{f \in \mathcal{F}} |\hat{L}_n(f) - L(f)| = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i) - L(f) \right|$$

satisfies the bounded differences condition for constants $c_i = \frac{1}{n}$, since the loss function l is assumed to be bounded, i.e., $0 \leq l \leq 1$. Thus, by the bounded differences inequality (Theorem 3.5), we obtain

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\hat{L}_n(f) - L(f)| - \mathbb{E}[\sup_{f \in \mathcal{F}} |\hat{L}_n(f) - L(f)|] \geq t \right) \leq \mathrm{e}^{-2nt^2}.$$

In particular,

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\hat{L}_n(f) - L(f)| - \mathbb{E}[\sup_{f \in \mathcal{F}} |\hat{L}_n(f) - L(f)|] < t \right) \geq 1 - \mathrm{e}^{-2nt^2},$$

and by evaluating this inequality at $t = \sqrt{\log(\delta^{-1})/2n}$, we see that the inequality

$$\sup_{f \in \mathcal{F}} |\hat{L}_n(f) - L(f)| < \mathbb{E}[\sup_{f \in \mathcal{F}} |\hat{L}_n(f) - L(f)|] + \sqrt{\frac{\log(\delta^{-1})}{2n}}$$

holds with probability at least $1 - \delta$. Therefore, as in Chapter 4, it remains to control the first term on the RHS. Ideally, we can do so independently of the (unknown) distribution of (X, Y).

5.2. Symmetrization and Rademacher Complexity

We have already introduced the technique of symmetrization and the concept of Rademacher complexity in Section 4.2. The line of argument we use here is naturally very similar to the one employed before. We introduce a so-called "ghost sample" $\mathcal{D}'_n = \{(X'_1, Y'_1), \dots, (X'_n, Y'_n)\}$, which is a set of random variables identically distributed as and independent of the random variables (X_i, Y_i) that make up our dataset $\mathcal{D}_n$. Using this ghost sample, we rewrite the error of f as

$$L(f) = \mathbb{E}[l(f(X), Y)] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} l(f(X'_i), Y'_i) \right] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} l(f(X'_i), Y'_i) \,\middle|\, \mathcal{D}_n \right] = \mathbb{E}[\hat{L}'_n(f) \mid \mathcal{D}_n],$$

where $\hat{L}'_n(f) = \frac{1}{n}\sum_{i=1}^n l(f(X'_i), Y'_i)$, since $(X'_1, Y'_1), \ldots, (X'_n, Y'_n)$ are identically distributed as (X, Y), and are independent of $\mathcal{D}_n$. Since the empirical error $\hat{L}_n(f)$ is $\sigma(Z_1, \ldots, Z_n)$-measurable, where $Z_i = (X_i, Y_i)$, we have

$$\begin{aligned}\mathbb{E}[\sup_{f\in\mathcal{F}} |\hat{L}_n(f) - L(f)|] &= \mathbb{E}[\sup_{f\in\mathcal{F}} |\hat{L}_n(f) - \mathbb{E}[\hat{L}'_n(f) \mid \mathcal{D}_n]|] \\ &= \mathbb{E}[\sup_{f\in\mathcal{F}} |\mathbb{E}[\hat{L}_n(f) - \hat{L}'_n(f) \mid \mathcal{D}_n]|] \leq \mathbb{E}[\sup_{f\in\mathcal{F}} \mathbb{E}[|\hat{L}_n(f) - \hat{L}'_n(f)| \mid \mathcal{D}_n]]\end{aligned}$$

by Jensen's inequality. For the same reasons as in Section 4.2, we can pull the supremum over $f \in \mathcal{F}$ inside the conditional expectation, and then apply the law of total expectation to arrive at

$$\mathbb{E}[\sup_{f\in\mathcal{F}} |\hat{L}_n(f) - L(f)|] \leq \mathbb{E}[\sup_{f\in\mathcal{F}} |\hat{L}_n(f) - \hat{L}'_n(f)|] = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n (l(f(X_i), Y_i) - l(f(X'_i), Y'_i))\right|\right].$$

Using the same symmetry arguments, we can introduce i.i.d. $\mathrm{Rad}(1/2)$ variables $\sigma_1, \ldots, \sigma_n$, independent of both samples $\mathcal{D}_n$ and $\mathcal{D}'_n$, to obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n (l(f(X_i), Y_i) - l(f(X'_i), Y'_i))\right|\right] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i (l(f(X_i), Y_i) - l(f(X'_i), Y'_i))\right|\right] \\ &\leq 2\,\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i l(f(X_i), Y_i)\right|\right],\end{aligned}$$

i.e., we are essentially in the same situation as in Section 4.2, the only difference being that the binary loss function $\mathbf{1}_{\{f(X)\neq Y\}}$ has been replaced by a more general loss function $l(f(X), Y)$ taking values in the unit interval $[0, 1]$. To get rid of the dependence of the bound on the sampled data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we generalize the concept of Rademacher complexity of a family of sets in order to apply it to the current context. As before, we write $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$.

Definition 5.1. *Let $\mathcal{F}$ be a class of regression functions $f\colon \mathcal{X} \to \mathcal{Y}$, and let $l\colon \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ be a loss function. The* Rademacher complexity *of $\mathcal{F}$ given l is defined as*

$$\mathfrak{R}_n(l \circ \mathcal{F}) = \sup_{z\in\wp(\mathcal{Z})} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i l(f(x_i), y_i)\right|\right],$$

where the supremum ranges over all subsets $z = \{z_1, \ldots, z_n\} \in \wp(\mathcal{Z})$, and $\sigma_1, \ldots, \sigma_n$ are i.i.d. $\mathrm{Rad}(1/2)$ random variables.

Taking into account our result obtained by symmetrization, we conclude that

$$\mathbb{E}[\sup_{f\in\mathcal{F}} |\hat{L}_n(f) - L(f)|] \leq 2\mathfrak{R}_n(l \circ \mathcal{F}).$$

In Section 4.3, we introduced the following notation: for a set of points $z = \{z_1, \ldots, z_n\} \subset \mathcal{X} \times \mathcal{Y}$, we let $T(z)$ be the set of all binary patterns $(\mathbf{1}_{\{z_1\in A\}}, \ldots, \mathbf{1}_{\{z_n\in A\}})$ created by sets $A \in \mathcal{A}$. In the context of minimizing the empirical risk, we defined a family of sets $\mathcal{A}_\mathcal{F} = \{A_f \mid f \in \mathcal{F}\}$, where $A_f = \{(x, y) \in \mathcal{X} \times \mathcal{Y} \mid f(x) \neq y\}$. In that setting, the set $T(z)$ thus consisted of elements

$$(\mathbf{1}_{\{f(x_1)\neq y_1\}}, \ldots, \mathbf{1}_{\{f(x_n)\neq y_n\}})^\top \in \mathbb{R}^n, \quad f \in \mathcal{F},$$

i.e., all possible vectors in $\mathbb{R}^n$ that could be generated by evaluating the binary loss function $\mathbf{1}_{\{f(x)\neq y\}}$ on a given set of points $\{z_1,\dots,z_n\}$ for all functions $f\in\mathcal{F}$ in the class of functions under consideration. We had also observed (see Lemma 4.4) that the Rademacher complexity of $\mathcal{A}_\mathcal{F}$ could be expressed in terms of the Rademacher complexity of the sets $T(z)$, i.e.,

$$\mathfrak{R}_n(\mathcal{A}_\mathcal{F}) = \sup_{z\in\wp(\mathcal{Z})} \mathfrak{R}_n(T(z)),$$

where the supremum is taken over all sets $z=\{z_1,\dots,z_n\}\in\wp(\mathcal{Z})$. We can easily adopt this line of thought to the current situation by replacing the binary loss function in the definition of $T(z)$ with a general loss function $l\colon \mathcal{Y}\times\mathcal{Y}\to[0,1]$:

$$T_l(z) = \{(l(f(x_1),y_1),\dots,l(f(x_n),y_n))^\top \mid f\in\mathcal{F}\} \tag{5.1}$$

Exactly as in Lemma 4.4, we have

$$\mathfrak{R}_n(l\circ\mathcal{F}) = \sup_{z\in\wp(\mathcal{Z})} \mathfrak{R}_n(T_l(z)), \tag{5.2}$$

where $\mathfrak{R}_n(T_l(z))$ is the Rademacher complexity of the subset $T_l(z)$ of $\mathbb{R}^n$. Using this identity, we can bound the Rademacher complexity of a *finite* class of regression functions $\mathcal{F}$ as follows:

Proposition 5.2. *Let $\mathcal{F}$ be a* finite *set of regression functions $f\colon\mathcal{X}\to\mathcal{Y}$ and let $l\colon\mathcal{Y}\times\mathcal{Y}\to[0,1]$ be a loss function taking values in the unit interval. Then,*

$$\mathfrak{R}_n(l\circ\mathcal{F}) \le \sqrt{\frac{2\log(2|\mathcal{F}|)}{n}}.$$

Proof. By (5.2) and Lemma 4.9 (applied to the set $T_l(z)\subset\mathbb{R}^n$), we have

$$\mathfrak{R}_n(l\circ\mathcal{F}) = \sup_{z\in\wp(\mathcal{Z})} \mathfrak{R}_n(T_l(z)) \le \sup_{z\in\wp(\mathcal{Z})} \max_{f\in\mathcal{F}} \|l_f(z)\|_2 \frac{\sqrt{2\log(2|T_l(z)|)}}{n},$$

where $l_f(z) = (l(f(x_1),y_1),\dots,l(f(x_n),y_n))^\top\in T_l(z)$, and z ranges over all sets $z=\{z_1,\dots,z_n\}\in\wp(\mathcal{Z})$. By definition of $T_l(z)$, we have $|T_l(z)|\le|\mathcal{F}|$. Further, since the loss function l takes values in $[0,1]$, we know that $\|l_f(z)\|_2\le\sqrt{n}$. Plugging everything back in, we see that

$$\mathfrak{R}_n(l\circ\mathcal{F}) \le \sup_{z\in\wp(\mathcal{Z})} \max_{f\in\mathcal{F}} \|l_f(z)\|_2 \frac{\sqrt{2\log(2|T_l(z)|)}}{n} \le \sqrt{\frac{2\log(2|\mathcal{F}|)}{n}}.$$ □

Remember that we had proven a very similar result in the setting of binary classification, namely Proposition 4.10. There, we showed that, for a family of sets $\mathcal{A}$, we have

$$\mathfrak{R}_n(\mathcal{A}) \le \sqrt{\frac{2\log(2\mathcal{S}_\mathcal{A}(n))}{n}}.$$

The proof of that result was nearly identical to the proof of Proposition 5.2, the only difference being that (to prove Proposition 4.10) we bounded the cardinality of $T(z)$ by the shatter coefficients $\mathcal{S}_\mathcal{A}(n)$ of $\mathcal{A}$, which turned out to be at most polynomial in n for a family $\mathcal{A}$ with finite VC dimension (Sauer-Shelah lemma). To generalize Proposition 5.2 to general (i.e., not necessarily finite) families $\mathcal{F}$ of regression functions, we will need

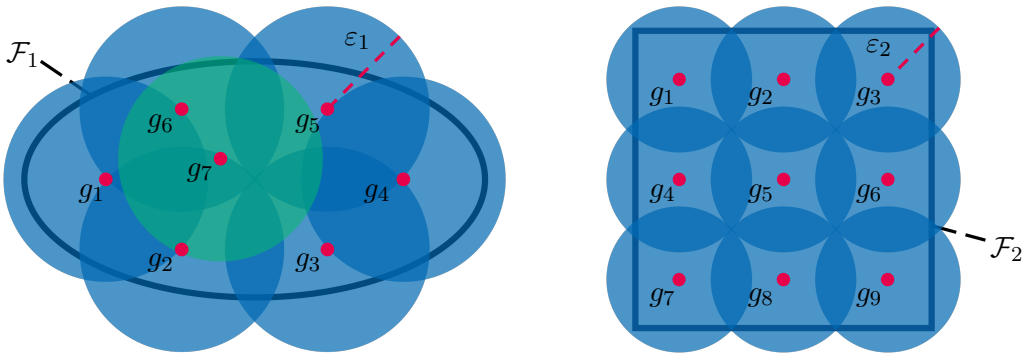

Figure 5.1. Coverings of two classes $\mathcal{F}_1$ and $\mathcal{F}_2$. **Left**, a class $\mathcal{F}_1$ covered by an ε_1-covering $V = \{g_1, \dots, g_7\}$. Note, however, that $\mathcal{N}(\mathcal{F}, d, \varepsilon_1) < 7$ since the ε_1-ball centered at g_7 and depicted in green is redundant (i.e., the remaining 6 ε_1-balls still cover all of $\mathcal{F}_1$). **Right**, a class $\mathcal{F}_2$ covered by a *minimal* ε_2-covering $V = \{g_1, \dots, g_9\}$.

to introduce a suitable alternative to the notion of the VC dimension of a family of sets. This is what we will do in the next section, where we discuss covering numbers.

5.3. Covering Numbers

As pointed out in the previous section, we need a suitable generalization of VC theory to find a bound on the Rademacher complexity $\mathfrak{R}_n(l \circ \mathcal{F})$ for arbitrary (i.e., possibly infinite) function classes $\mathcal{F}$. We have seen that, as in the case of binary classification, the cardinality of the set $T_l(z) = \{(l(f(x_1), y_1), \dots, l(f(x_n), y_n))^\top \mid f \in \mathcal{F}\}$ plays a significant role in bounding the Rademacher complexity of $l \circ \mathcal{F}$. When $\mathcal{F}$ is infinite, the set $T_l(z)$ will most likely be infinite as well. Thus, we need to find a way that lets us treat points in this set that are close to each other as if they were identical. The concept of ε-*coverings* will help us do just that.

Definition 5.3. *Let (X, d) be a pseudometric space, let K be a subset of X, and let $\varepsilon > 0$. An* external ε-covering *of K is a set $V \subset X$ such that*

$$K \subset \bigcup_{x \in V} B_{d,\varepsilon}(x),$$

where $B_{d,\varepsilon}(x) = \{y \in X \mid d(x, y) \leq \varepsilon\}$ is the closed ball of radius ε centered at $x \in X$. V is called an internal ε-covering *of K, if $V \subset K$.*
The external ε-covering number $\mathcal{N}^{\text{ext}}(K, d, \varepsilon)$ *of K is the minimum number of elements needed to form an external ε-covering of K, i.e.,*

$$\mathcal{N}^{\text{ext}}(K, d, \varepsilon) = \inf\{|V| \mid V \text{ is an external } \varepsilon\text{-covering of } K\}.$$

An external ε-covering V of K is called minimal, *if $|V| = \mathcal{N}^{\text{ext}}(K, d, \varepsilon)$. Finally,* internal ε-covering numbers $\mathcal{N}^{\text{int}}(K, d, \varepsilon)$ *and* minimal *internal ε-coverings are defined analogously.*

If V is an ε-covering (external or internal) of a set K, then every $y \in K$ is within distance of at most ε to some $x \in V$, i.e., for every $y \in K$ there exists $x \in V$ such that $d(x, y) \leq \varepsilon$. For a class of functions $\mathcal{F}$, we will assume ε-coverings to be *internal* coverings, unless explicitly stated otherwise. To ease notation, we simply write $\mathcal{N}(\mathcal{F}, d, \varepsilon)$ for the internal ε-covering numbers.

Next, we introduce the *conditional Rademacher average* of a class[1] of functions $\mathcal{F}$ given a set of points $\{z_1, \dots, z_n\}$.

Definition 5.4. *The* conditional Rademacher average *of a class $\mathcal{F}$ of functions $f\colon \mathcal{Z} \to \mathbb{R}$ given a set of n points $z = \{z_1, \dots, z_n\} \subset \mathcal{Z}$ is defined as*

$$\hat{\mathfrak{R}}_n^z(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right|\right],$$

where $\sigma_1, \dots, \sigma_n$ are i.i.d. $\mathrm{Rad}(1/2)$ random variables.

Notice that, if $\mathcal{F}$ is a class of functions $f\colon \mathcal{Z} \to \mathbb{R}$, and we define the Rademacher complexity of $\mathcal{F}$ as[2]

$$\mathfrak{R}_n(\mathcal{F}) = \sup_{z\in\wp(\mathcal{Z})} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right|\right],$$

where the supremum ranges over all subsets $z = \{z_1, \dots, z_n\} \in \wp(\mathcal{Z})$, then clearly

$$\mathfrak{R}_n(\mathcal{F}) = \sup_{z\in\wp(\mathcal{Z})} \hat{\mathfrak{R}}_n^z(\mathcal{F}). \tag{5.3}$$

There is one more term we have to introduce before we state the first result of this section.

Definition 5.5. *Given a set of points $z = \{z_1, \dots, z_n\} \subset \mathcal{Z}$ and a class $\mathcal{F}$ of functions $f\colon \mathcal{Z} \to \mathbb{R}$, the* empirical L^1-distance *between $f, g \in \mathcal{F}$ is given by*

$$d_1^z(f,g) = \frac{1}{n}\sum_{i=1}^n |f(z_i) - g(z_i)|.$$

With these definitions in place, we will prove an upper bound of the conditional Rademacher average of a class $\mathcal{F}$ that (besides the number of observations n) depends only on the ε-covering numbers of $\mathcal{F}$ with respect to the empirical L^1-distance d_1^z. In order for these (covering numbers) to be well-defined, we would have to show that the empirical L^1-distance defines a pseudometric on a given class $\mathcal{F}$ (since this is assumed to be the case in the definition of covering numbers presented earlier). This is indeed true, and it follows directly from the properties of the absolute value.

Theorem 5.6. *Let $\mathcal{F}$ be a class of functions $f\colon \mathcal{Z} \to [-1, 1]$, and let $z = \{z_1, \dots, z_n\} \subset \mathcal{Z}$ be a set of n points. Then,*

$$\hat{\mathfrak{R}}_n^z(\mathcal{F}) \le \inf_{\varepsilon>0} \varepsilon + \sqrt{\frac{2\log(2\mathcal{N}(\mathcal{F}, d_1^z, \varepsilon))}{n}}.$$

Proof. We can assume $\mathcal{N}(\mathcal{F}, d_1^z, \varepsilon) < \infty$, since the inequality is trivially true otherwise. Given $\varepsilon > 0$, we let V_ε be a minimal ε-covering of $\mathcal{F}$, i.e., $|V_\varepsilon| = \mathcal{N}(\mathcal{F}, d_1^z, \varepsilon) < \infty$. For every function $f \in \mathcal{F}$, there exists $f^\circ \in V_\varepsilon$

[1] In what follows, we use a general class of functions $\mathcal{F}$ and a set of points $\{z_1, \dots, z_n\}$. In the setting of empirical risk minimization, we would substitute $\mathcal{F}$ for $l \circ \mathcal{F}$ and the set under consideration would be the observations $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

[2] This is simply the generalization of Definition 5.1 to an arbitrary class of functions $\mathcal{F}$.

such that $d_1^z(f, f^\circ) \leq \varepsilon$. By the triangle inequality, we have

$$\hat{\mathfrak{R}}_n^z(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i f(z_i)\right|\right] \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i (f(z_i) - f^\circ(z_i))\right|\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i f^\circ(z_i)\right|\right].$$

The triangle inequality also implies

$$\frac{1}{n}\left|\sum_{i=1}^n \sigma_i (f(z_i) - f^\circ(z_i))\right| \leq \frac{1}{n}\sum_{i=1}^n |f(z_i) - f^\circ(z_i)| = d_1^z(f, f^\circ) \leq \varepsilon,$$

since $|\sigma_i| = 1$ almost surely. Hence,

$$\hat{\mathfrak{R}}_n^z(\mathcal{F}) \leq \varepsilon + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i f^\circ(z_i)\right|\right].$$

As $f^\circ \in V_\varepsilon$, we obtain

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i f^\circ(z_i)\right|\right] = \mathbb{E}\left[\max_{g\in V_\varepsilon} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i g(z_i)\right|\right] = \mathfrak{R}_n(B),$$

where $\mathfrak{R}_n(B)$ denotes the Rademacher complexity of the *finite* set $B = \{(g(z_1), \ldots, g(z_n))^\top \mid g \in V_\varepsilon\} \subset \mathbb{R}^n$. Since every $g \in V_\varepsilon \subset \mathcal{F}$ takes values in $[-1, 1]$, we have $\max_{b\in B} \|b\|_2 \leq \sqrt{n}$, and Lemma 4.9 hence tells us that

$$\mathfrak{R}_n(B) \leq \sqrt{\frac{2\log(2|B|)}{n}} \leq \sqrt{\frac{2\log(2|V_\varepsilon|)}{n}} = \sqrt{\frac{2\log(2\mathcal{N}(\mathcal{F}, d_1^z, \varepsilon))}{n}},$$

since $|B| \leq |V_\varepsilon| = \mathcal{N}(\mathcal{F}, d_1^z, \varepsilon)$. Altogether,

$$\hat{\mathfrak{R}}_n^z(\mathcal{F}) \leq \varepsilon + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i f^\circ(z_i)\right|\right] \leq \varepsilon + \sqrt{\frac{2\log(2\mathcal{N}(\mathcal{F}, d_1^z, \varepsilon))}{n}}.$$

Since this is true for every $\varepsilon > 0$, we can take the infimum over ε to arrive at the desired result. □

Observe that there is a trade-off in the upper bound of the previous result, since the ε-covering numbers $\mathcal{N}(\mathcal{F}, d_1^z, \varepsilon)$ of $\mathcal{F}$ increase as ε decreases, and vice versa. Next, we introduce another set of empirical distances.

Definition 5.7. *Given a set of points $z = \{z_1, \ldots, z_n\} \subset \mathcal{Z}$ and a class $\mathcal{F}$ of functions $f\colon \mathcal{Z} \to \mathbb{R}$, the* empirical L^p-distance *between $f, g \in \mathcal{F}$ is given by*

$$d_p^z(f, g) = \left(\frac{1}{n}\sum_{i=1}^n |f(z_i) - g(z_i)|^p\right)^{1/p}, \quad p \geq 1.$$

For $p = \infty$, we set

$$d_\infty^z(f, g) = \max_{1\leq i\leq n} |f(z_i) - g(z_i)|.$$

Before we proceed to state the next result on ε-covering numbers, we recall a special case of Hölder's inequality.

Proposition 5.8 (Hölder, 1889). *For any $r, s \geq 0$, and $x, y \in \mathbb{R}^n$, it holds*

$$\left(\sum_{i=1}^{n} |x_i|^r |y_i|^s\right)^{r+s} \leq \left(\sum_{i=1}^{n} |x_i|^{r+s}\right)^{r} \left(\sum_{i=1}^{n} |y_i|^{r+s}\right)^{s}.$$

It is clear that the ε-covering numbers of a class $\mathcal{F}$ decrease as ε increases. Essentially, this is due to the fact that the open balls $B_{d,\varepsilon}(f)$ increase in size[3] for larger ε, i.e.,

$$\varepsilon \uparrow \quad \Rightarrow \quad \text{size of } B_{d,\varepsilon}(f) \uparrow \quad \Rightarrow \quad \mathcal{N}(\mathcal{F}, d, \varepsilon) \downarrow .$$

We will now show that something similar in spirit is true for the family of empirical distances d_p^z we have defined above, namely,

$$p \uparrow \quad \Rightarrow \quad \text{size of } B_{d_p^z,\varepsilon}(f) \downarrow \quad \Rightarrow \quad \mathcal{N}(\mathcal{F}, d_p^z, \varepsilon) \uparrow .$$

Proposition 5.9. *Let $\mathcal{F}$ be a class of functions $f\colon \mathcal{Z} \to \mathbb{R}$ and fix a set of points $z = \{z_1, \ldots, z_n\}$. For $1 \leq p \leq q < \infty$ and $\varepsilon > 0$, it holds*

$$\mathcal{N}(\mathcal{F}, d_p^z, \varepsilon) \leq \mathcal{N}(\mathcal{F}, d_q^z, \varepsilon) \leq \mathcal{N}(\mathcal{F}, d_\infty^z, \varepsilon).$$

Proof. Throughout this proof, let us write $u_i = f(z_i) - g(z_i)$ and $u = \max_{1\leq i\leq n} |u_i|$. We first tackle the inequality $\mathcal{N}(\mathcal{F}, d_q^z, \varepsilon) \leq \mathcal{N}(\mathcal{F}, d_\infty^z, \varepsilon)$. Since the functions $x \mapsto x^q$ and $x \mapsto x^{1/q}$ are increasing on $[0, \infty)$, we have

$$d_q^z(f,g) = \left(\frac{1}{n}\sum_{i=1}^{n} |u_i|^q\right)^{1/q} \leq \left(\frac{1}{n}\sum_{i=1}^{n} u^q\right)^{1/q} = u = \max_{1\leq i\leq n} |u_i| = d_\infty^z(f,g).$$

If $g \in B_{d_\infty^z,\varepsilon}(f)$, so that $d_\infty^z(f,g) < \varepsilon$, then the above inequality implies $d_q^z(f,g) \leq d_\infty^z(f,g) < \varepsilon$. Hence, $B_{d_\infty^z,\varepsilon}(f) \subset B_{d_q^z,\varepsilon}(f)$. This, in turn, implies that every ε-covering of $\mathcal{F}$ w.r.t. d_∞^z defines an ε-covering of $\mathcal{F}$ w.r.t. d_q^z, and thus, $\mathcal{N}(\mathcal{F}, d_q^z, \varepsilon) \leq \mathcal{N}(\mathcal{F}, d_\infty^z, \varepsilon)$.

To prove the inequality $\mathcal{N}(\mathcal{F}, d_p^z, \varepsilon) \leq \mathcal{N}(\mathcal{F}, d_q^z, \varepsilon)$, we will make use of the version of Hölder's inequality stated in Proposition 5.8. With $r = \frac{1}{p} - \frac{1}{q} \geq 0$ and $s = \frac{1}{q} \geq 0$, we have $r + s = \frac{1}{p}$, and hence

$$\begin{aligned} d_p^z(f,g) = \left(\frac{1}{n}\sum_{i=1}^{n} |u_i|^p\right)^{1/p} &= n^{-1/p}\left(\sum_{i=1}^{n} |1|^r |u_i^{pq}|^s\right)^{r+s} \\ &\leq n^{-1/p}\left(\sum_{i=1}^{n} |1|^{r+s}\right)^{r}\left(\sum_{i=1}^{n} |u_i^{pq}|^{r+s}\right)^{s} = n^{-1/q}\left(\sum_{i=1}^{n} |u_i|^q\right)^{1/q} = d_q^z(f,g). \end{aligned}$$

By the same logic as before, the inequality $d_p^z(f,g) \leq d_q^z(f,g)$ implies $\mathcal{N}(\mathcal{F}, d_p^z, \varepsilon) \leq \mathcal{N}(\mathcal{F}, d_q^z, \varepsilon)$. □

[3] More precisely, $\varepsilon > \varphi$ implies $B_{d,\varepsilon}(f) \supset B_{d,\varphi}(f)$.

Part II

Neural Networks

6. Binary Classification with a Perceptron

6.1. Single-Layer Perceptron

6.2. Mutli-Layer Perceptron

7. Statistical Learning Theory for Neural Networks

7.1. Approximation by Neural Networks

7.2. The VC Dimension of Neural Networks

8. Features and Architectures of Neural Networks

8.1. Multi-Class Classification

8.2. Convolutional Neural Networks

8.3. Recurrent Neural Networks

8.4. Autoencoders

9. Training Neural Networks

9.1. Forward and Backward Propagation

9.2. A First Look at (Stochastic) Gradient Descent

Part III

Exercises & Solutions

10. Exercises

10.1. Exercise Set 1

Exercise 1. *In the setting of Proposition 1.1, prove*

$$\mathbb{E}[(f_i(X) - \mathbb{E}[Y_i \mid X])(\mathbb{E}[Y_i \mid X] - Y_i)] = 0$$

for all $i = 1, \dots, d$, where $f_i \colon \mathcal{X} \to \mathbb{R}$ are measurable functions.

Exercise 2. *Let H be a random variable that almost surely takes values in the unit interval $[0, 1]$. Show*

$$\mathbb{E}[\min(1 - H, H)] = 1/2 - 1/2\,\mathbb{E}[|2H - 1|].$$

Exercise 3. *Let h^* be the Bayes classifier and let L^* be the Bayes error. For $Y \in \{0, 1\}$, prove*

$$\min_{h \colon \mathbb{R}^d \to \mathbb{R}} \mathbb{E}[|h(X) - Y|] = \mathbb{E}[|h^*(X) - Y|] = L^*.$$

Exercise 4. *Assume that X has a density f with respect to the Lebesgue measure, i.e.,*

$$\mathbb{P}(X \in A) = \int_A f(x)\,\mathrm{d}x,$$

and assume further that the class-conditional densities f_i of X given $Y = i$ exist for $i = 1, 2$, i.e.,

$$\mathbb{P}(X \in A \mid Y = i) = \int_A f_i(x)\,\mathrm{d}x, \quad i = 1, 2.$$

Finally, denote the class probabilities by $p = \mathbb{P}(Y = 1) = 1 - \mathbb{P}(Y = 0)$. Demonstrate that

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{f_1(x)p}{f_1(x)p + f_0(x)(1 - p)}.$$

10.2. Exercise Set 2

Exercise 5. *Let $X_1, \dots, X_n$ be i.i.d. random variables with $X_i \in [0, 1]$, representing the sizes of packages that ought to be shipped. Each shipping container has a capacity of 1, i.e., a single container can fit k packages $X_{i_1}, \dots, X_{i_k}$ if these packages satisfy*

$$\sum_{l=1}^{k} X_{i_l} \leq 1.$$

Let B_n denote the minimum *number of containers needed to ship all n packages. Show that*

$$\mathbb{P}(|B_n - \mathbb{E}[B_n]| \geq t) \leq 2 \exp\left(\frac{-2t^2}{n}\right).$$

Exercise 6. *Let $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^d$ be i.i.d. random vectors with $\mathbb{E}[\mathbf{X}_i] = \mathbf{0}$ and $\|\mathbf{X}_i\|_2 \leq 1$, where $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbb{R}^d$, and let $\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$. Show that there exist constants C_a and C_b such that*

(1) $\mathbb{P}(\|\overline{\mathbf{X}}\|_2 - \mathbb{E}[\|\overline{\mathbf{X}}\|_2] \geq t) \leq \exp\bigl(-C_a n t^2\bigr)$

(2) $\mathbb{E}[\|\overline{\mathbf{X}}\|_2] \leq C_b/\sqrt{n}$

Exercise 7. *Let $X_1, \dots, X_n$ be i.i.d. random variables with distribution function $F(t)$ and let*

$$F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i \leq t\}}, \quad t \in \mathbb{R}$$

denote the empirical distribution function*.*

(1) Compute the mean and variance of $F_n(t)$.

(2) Show that $\mathbb{P}(|F_n(t) - F(t)| > \varepsilon) \leq 2\exp\bigl(-2n\varepsilon^2\bigr)$.

(3) Conclude that $F_n(t) \xrightarrow{\text{a.s.}} F(t)$.

10.3. Exercise Set 3

Exercise 8.

Exercise 9.

Exercise 10.

10.4. Exercise Set 4

Exercise 11.

Exercise 12.

Exercise 13.

10.5. Exercise Set 5

Exercise 14.

Exercise 15. *Let $\mathcal{F}$ be a class of functions $f\colon \mathcal{X} \to \mathbb{R}$, and let $X_1, \dots, X_n$ be i.i.d. random variables with values in $\mathcal{X}$. Further, let $\sigma_1, \dots, \sigma_n$ be i.i.d. $\mathrm{Rad}(1/2)$ random variables that are independent of $X_1, \dots, X_n$. Prove the* desymmetrization inequality*:*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i(f(X_i) - \mathbb{E}[f(X_i)])\right|\right] \leq 2\,\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)])\right|\right].$$

Exercise 16.

11. Solutions

As you work through the exercises provided in the previous chapter, you may notice the absence of solutions. Allow me to explain this deliberate choice. In mathematics, true understanding is not achieved by passively reading over material nor by copying solutions someone else has written up but by actively grappling with problems on one's own. The exercises in the previous chapter allow you to do just that. Should you encounter difficulties, I encourage you to persist. Struggle is an inevitable part of the learning process and often leads to profound insights. In case you have dedicated considerable effort to a given problem[1], and still do not seem to make any progress, feel free to reach out to me via mrvnthss@mail.de and I will be happy to assist.

[1] i.e., work on it for an extended period of time, ideally put the problem aside for a couple of days and then come back to it

Bibliography

[DGL96] Luc Devroye, Laszlo Györfi, and Gabor Lugosi. *A probabilistic theory of pattern recognition*. Springer New York, NY, 1996. DOI: 10.1007/978-1-4612-0711-5.

[HTF17] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. Springer Series in Statistics. Springer New York, NY, 2017. DOI: 10.1007/b94608.

[Rig15] Philippe Rigollet. *18.657 Mathematics for machine learning*. 2015. URL: https://ocw.mit.edu/courses/18-657-mathematics-of-machine-learning-fall-2015/pages/lecture-notes/ (visited on 10/25/2023).

[VC71] Vladimir N. Vapnik and Alexey Y. Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities". In: *Theory of Probability and Its Applications* 16.2 (1971), pp. 264–280. DOI: 10.1137/1116025.

[VC74] Vladimir N. Vapnik and Alexey Y. Chervonenkis. *Theory of pattern recognition* [in Russian]. (German translation: W. N. Vapnik and A. Ja. Chervonenkis. *Theorie der Zeichenerkennung*. Berlin, Germany: Akademie-Verlag, 1979.) Moscow, Russia: Nauka, 1974.