

Lecture notes: mathematics for artificial intelligence 1

G. Blanchard

October 20, 2025

Contents

0	Some reminders on probability theory	3
0.1	Elements of probability	3
0.2	A few important properties	5
0.3	Conditioning	7
1	Introduction to statistical learning theory (part 1):	
	Decision theory	11
1.1	Mathematical formalization	11
1.2	Optimal risk and prediction function.	12
1.3	Learning from data	16
1.4	Consistency of estimators	18
1.5	Plug-in classification	23
1.6	Negative results: the “no free lunch” theorem	26
2	Linear Discrimination:	
	A brief overview of classical methods	30
2.1	Linear discrimination functions	30
2.2	The naive Bayes classifier	31
2.3	Gaussian generative distribution: LDA and QDA	33
2.4	Classification as regression	35
2.5	Linear logistic regression	37
2.6	Hinge-loss based methods: Perceptron and Support Vector Machine	39
2.7	Regularization	41
3	Introduction to statistical learning theory (part 2): elementary bounds	43
3.1	Controlling the error of a single decision function and Hold-Out principle	43
3.2	A sharp but inconvenient bound in the Bernoulli case: the Clopper-Pearson bound	44
3.3	The Chernov method	45
3.4	Sub-Gaussian random variables and Hoeffding’s inequality	48

3.5	Uniform bounds over a finite class of prediction functions	51
3.6	Uniform bounds over a countable class of prediction functions; regularized ERM	56
4	The Nearest Neighbors method	63
4.1	Basic notation and definitions	63
4.2	Analysis for k fixed	64
4.3	Consistency of the k -nearest-neighbors method	68
5	Reproducing kernel methods	74
5.1	Motivation	74
5.2	Reproducing kernel Hilbert spaces	76
5.3	Construction of spsd kernels	81
5.4	Kernel-based methods	83
5.5	Regularity and approximation properties of functions in a RKHS	87
5.6	Translation invariant kernels and random Fourier features	91
6	Introduction to statistical learning theory (part 3): Rademacher complexities and VC theory	98
6.1	Introduction, reminders	98
6.2	The Azuma-McDiarmid inequality	99
6.3	Rademacher complexity	102
6.4	Properties of the Rademacher complexity	105
6.5	Application to kernel methods	108
6.6	Vapnik-Chervonenkis theory	112
6.7	Application to artificial neural networks	117

Convention used for notation of importance of results (margin stars).

A triple margin star *** indicates a fundamental result, whose proof has to be known. It can be asked to give a proof in the exam, without additional hints or reminders. The proof itself is generally important, and it can also be expected that solutions of some exercises involve variants of the ideas of the proof, so these have to be understood at a deep level.

A double margin star ** indicates a fundamental definition or an important result that has to be known and can be asked to be restated at the exam. If it is a result with a proof, it is recommended to have an idea of how the proof works, but it won't be asked to redo the proof (at least not without reminders or other form of help.)

A single margin star * indicates an important result. Solutions of some exercises can rely on using this result. The proof does not have to be known inside out.

A diamond \diamond indicates a result that is given for illustration; it can be seen as an exercise putting into light some interesting applications, argument techniques or results. It won't be required to state nor prove the result in the exam, but it is of interest to know how the result works for training.

0 Some reminders on probability theory

0.1 Elements of probability

The theoretical approaches to artificial intelligence involve many different fields of mathematics. A central tenet of most approaches to modern mathematical modeling of artificial intelligence methods, which in one form or other receive data and “learn” from it, is that said data should be modeled as random. Thus, while other distinct areas of mathematics play an important role, that of probability theory should be considered central. In these notes, we will chiefly concentrate on probabilistic and statistical aspects – what is usually called “statistical learning theory”.

While we assume the reader to be familiar with mathematical elements of probability theory, we start with recalling a few fundamentals.

Probability spaces. A probability space (Ω, \mathcal{A}, P) consists of a base space Ω , a σ -algebra of subsets of Ω (called measurable subsets, or events), and a probability distribution \mathbb{P} , that is, a mapping $\mathcal{A} \rightarrow [0, 1]$ satisfying the fundamental axioms of probability ($P(\Omega) = 1$, and σ -additivity over disjoint countable unions).

If A_1, A_2, \dots are events, not necessarily disjoint, then we always have as a consequence of the fundamental axioms:

$$\mathbb{P}\left[\bigcup_{i \geq 1} A_i\right] \leq \sum_{i \geq 1} P(A_i), \tag{0.1}$$

which we will call the *union bound* (also known as Boole’s inequality).

A probability distribution is a measure, and as such we can integrate real-valued measurable functions $f : \Omega \rightarrow \mathbb{R}$ with respect to P .

Random variables. A *random variable* (r.v.) over (Ω, \mathcal{A}, P) with values in a measured space $(\mathcal{X}, \mathcal{F})$ is a measurable map from the former to the latter space. It induces an image (“push-forward”) probability measure P_X on the image space (also sometimes noted $X\#P$), defined via

$$\forall F \in \mathcal{F} : P_X(F) = P(X^{-1}(F)) = P(X \in F),$$

called the distribution of X . It will be assumed that all considered random variables in a given context are defined on the same underlying probability space (Ω, \mathcal{A}, P) ; the latter is generally left unspecified, since we will generally only be interested in studying some specific random variables.

If X is a random variable over (Ω, \mathcal{A}, P) and G is a further measurable map $(\mathcal{X}, \mathcal{F}) \rightarrow (\mathcal{Z}, \mathcal{G})$, then $Z = G(X)$ is obviously also a random variable over (Ω, \mathcal{A}, P) (with values in \mathcal{Z}). If $\mathcal{Z} = \mathbb{R}$ and Z is integrable, we have the formula

$$\int_{\mathbb{R}} z P_Z(dz) = \int_{\mathcal{X}} G(x) P_X(dx) =: \mathbb{E}[Z], \quad (0.2)$$

called expectation of Z .

The first equality (“change of variable formula”) can be useful since it can possibly avoid to compute explicitly the distribution of Z to perform the integral. Expectation can be defined for a vector-valued variable ($\mathcal{Z} = \mathbb{R}^d$) in the obvious way (i.e. coordinate-wise).

Densities. If X is a r.v. taking values in \mathcal{X} , μ is some reference measure on $(\mathcal{X}, \mathcal{F})$ and it holds

$$\forall F \in \mathcal{F} : \quad P_X(F) = \int_F f(x) \mu(dx) = \int_{\mathcal{X}} \mathbf{1}\{x \in F\} f(x) \mu(dx)$$

for some measurable function $f : \mathcal{X} \rightarrow \mathbb{R}_+$, then we say that P_X has density f with respect to μ . It is sufficient to check the above equality for events F in a family of sets generating \mathcal{F} and stable by finite intersection (π -system), for it to hold for all $F \in \mathcal{F}$.

For instance, on \mathbb{R} with the standard Borel σ -algebra it is sufficient to check it on (open or closed) intervals; on \mathbb{R}^d , it suffices to consider parallelepipeds.

Marginals. Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{X}', \mathcal{F}')$ denote two measurable spaces. The product $\mathcal{X} \times \mathcal{X}'$ can be endowed by the *product σ -algebra* $\mathcal{F} \otimes \mathcal{F}'$ (generated by products of events in \mathcal{F} and \mathcal{F}').

If P is a probability distribution on a product space $(\mathcal{X} \times \mathcal{X}', \mathcal{F} \otimes \mathcal{F}')$, the first (or \mathcal{X} -)marginal of P is the probability distribution of the random variable given by the projection $(x, x') \mapsto x$, and similarly for the second (or \mathcal{X}' -)marginal.

If Z is a random variable over (Ω, \mathcal{A}) with values in the product space $(\mathcal{X} \times \mathcal{X}', \mathcal{F} \otimes \mathcal{F}')$, then $Z(\omega) = (X(\omega), X'(\omega))$, with X random variable, and P_X is called (first) marginal distribution of Z (it is also the first marginal distribution of P_Z in the above sense).

Independence. The random variables (X, X') on $(\mathcal{X} \times \mathcal{X}', \mathcal{F} \otimes \mathcal{F}')$ are independent (also denoted $X \perp\!\!\!\perp X'$) iff their joint distribution (as a couple) is the product distribution their marginals, or equivalently, if

$$\forall F \in \mathcal{F}, F' \in \mathcal{F}' \quad P_{(X, X')}(F \times F') = P_X(F) P_{X'}(F'),$$

or equivalently

$$\forall F \in \mathcal{F}, F' \in \mathcal{F}' \quad P(X \in F, X' \in F') = P(X \in F) P(X' \in F').$$

Yet equivalently, independence holds when the joint distribution is the product measure of the marginals:

$$P_{(X,X')} = P_X \otimes P_{X'}.$$

As previously, it is sufficient to check the above equalities on a generating π -system to establish independence.

By Fubini's theorem, if X, X' are real-valued independent and integrable, then their product XX' is integrable and $\mathbb{E}[XX'] = \mathbb{E}[X]\mathbb{E}[X']$.

If X, X' are independent variables and F, G are measurable mappings into further measured spaces, then $F(X)$ and $G(X')$ are independent.

This generalizes to finite families of r.v.'s and even to countable families, though we won't really need it, since we will not be interested in a.s. convergence most of the time in the present notes. When X_1, X_2, \dots, X_n are independent and have the same marginal P_X , we say they form an independent identically distributed (i.i.d.) family and denote their joint distribution $P_X^{\otimes n}$.

Exercises

Exercise 0.1. Justify the first equality in (0.2) if the space \mathcal{X} is countable (using formulas of discrete probability, i.e. integrals become sums).

Exercise 0.2. If f is a nonnegative, real-valued function on a measured space $(\mathcal{X}, \mathcal{F}, \mu)$ such that $\int f d\mu = 1$, then $P(F) := \int_F f d\mu$ is a probability distribution on \mathcal{X} (with density f): justify.

Exercise 0.3. If r.v. Z has density $f(x, x')$ with respect to a product measure $\mu \otimes \mu'$ on $\mathcal{X} \times \mathcal{X}'$, then the first marginal distribution P_X of Z has a density with respect to μ : justify why and specify it explicitly.

Exercise 0.4. If P_X has density f wrt. μ and $P_{X'}$ has density f' wrt. μ' , and X, X' are independent, then $P_{X, X'}$ has density $(x, x') \mapsto f(x)f'(x')$ wrt. $\mu \otimes \mu'$. Conversely: if $P_{X, X'}$ has density $f(x, x')$ wrt. $\mu \otimes \mu'$ and it holds that $f(x, x') = g(x)h(x')$ for some functions g, h (not necessarily densities), then X, X' are independent.

0.2 A few important properties

Support. Let $(\mathcal{X}, \mathcal{F})$ be a Borel space, which we recall is a topological space endowed by the σ -algebra generated by its open sets (also called the Borel σ -algebra). If μ is a measure on \mathcal{X} , its support is defined as

$$\text{Supp}(\mu) := \{x \in \mathcal{X} : \text{for any open set } N, N \ni x \Rightarrow \mu(N) > 0\}.$$

The support of a measure is a closed set (exercise). Furthermore,

if \mathcal{X} is a polish space (metrizable, complete, separable) then

$$\mu(\text{Supp}(\mu)^c) = 0.$$

As a consequence, if we establish a certain property holds for all $x \in \text{Supp}(\mu)$, then it holds for μ -almost all $x \in \mathcal{X}$.

Positivity of expectation. If X is a nonnegative real random variable, then $\mathbb{E}[X] = 0$ implies $X = 0$ a.s.

Markov's inequality. If X is a nonnegative real random variable, then for any $t > 0$:

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Jensen's inequality. If X is an integrable real random variable and ϕ is a convex function $\mathbb{R} \rightarrow \mathbb{R}$ such that $\phi(X)$ is integrable, then

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

As a consequence, we have in particular (for $\phi(x) = x^2$) that for a squared integrable random variable:

$$\text{Var}[X] := \mathbb{E}[X^2] - \mathbb{E}[X]^2 \geq 0.$$

However, the latter fact can be also established directly by the variance formula

$$\text{Var}[X] := \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Bias-Variance formula. If X is a real-valued squared integrable random variable, then for any $c \in \mathbb{R}$, the random variable $(X - c)$ is integrable and

$$\mathbb{E}[(X - c)^2] = (\mathbb{E}[X] - c)^2 + \mathbb{E}[(X - \mathbb{E}[X])^2] = (\mathbb{E}[X] - c)^2 + \text{Var}[X].$$

Exercise 0.5. Assume probability \mathbb{P} on a Borel space \mathcal{X} has a continuous density f with respect to a reference measure ν . Prove $\text{Supp}(\mathbb{P}) = \overline{\{x : f(x) > 0\}}$. Is this true if f is not continuous?

Exercise 0.6. Prove the formula: $\text{Var}[X] = \frac{1}{2}\mathbb{E}[(X - X')^2]$, where X, X' are two independent square integrable real variables having the same marginal distribution.

0.3 Conditioning

If A is an event with $P(A) > 0$, the standard definition for the conditional probability of an event B conditioned to A is $P(B|A) := \frac{P(A \cap B)}{P(A)}$. An important property is then that $P(\bullet|A) := (B \mapsto P(B|A))$ is itself a probability distribution (it satisfies the axioms) called conditional probability distribution P conditional to A . If X is a random variable, we can for instance take $A = \{X \in F\}$ provided $P(X \in F) > 0$.

In what follows, we'll very often consider random variables (X, Y) with a joint distribution P_{XY} and would like to consider the conditional distributions conditional to $X = x$ or $Y = y$, but these events unfortunately have null probability in general, so that the above definition does not apply. We need something more general.

Definition 0.1. Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ be two measurable spaces. We call *regular transition probability* or *Markov kernel* from \mathcal{X} to \mathcal{Y} a mapping $\kappa : \mathcal{G} \times \mathcal{X} \rightarrow [0, 1]$ such that:

- (i) For all $x \in \mathcal{X}$, the mapping $\kappa(\bullet, x) : \mathcal{G} \mapsto \kappa(G, x)$ is a probability distribution on $(\mathcal{Y}, \mathcal{G})$;
- (ii) For all $G \in \mathcal{G}$, the mapping $\kappa(G, \bullet) : x \mapsto \kappa(G, x)$ is measurable.

Given a probability distribution P on \mathcal{X} and a regular transition probability κ from \mathcal{X} to \mathcal{Y} , we can define a joint probability on $\mathcal{X} \times \mathcal{Y}$, as

$$\kappa \circ P(F \times G) := \int_F \int_G \kappa(dy, x) P(dx),$$

and more generally for an integrable real-valued function f on $\mathcal{X} \times \mathcal{Y}$:

$$\int_{\mathcal{X} \times \mathcal{Y}} f(x, y) (\kappa \circ P)(dx, dy) := \int_{\mathcal{X}} \int_{\mathcal{Y}} f(x, y) \kappa(dy, x) P(dx). \quad (0.3)$$

A fundamental question is the converse, that is, given a joint probability P_{XY} on $\mathcal{X} \otimes \mathcal{Y}$, and P_X the first marginal distribution, does it exist a transition probability κ from \mathcal{X} to \mathcal{Y} such that $P_{XY} = \kappa \circ P$? The following theorem is fundamental and guarantees the existence of such an object in a sufficiently broad situation.

**

Theorem 0.2 (Disintegration theorem). *Assume $(\mathcal{Y}, \mathcal{G})$ is a “nice probability space” (see below). Let P be a probability distribution on $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \otimes \mathcal{G})$, and P_X its first marginal distribution. Then there exists a transition kernel $P_{\mathcal{Y}|\mathcal{X}}$ from \mathcal{X} to \mathcal{Y} , called a regular conditional probability distribution (rcpd) of P such that $P = P_{\mathcal{Y}|\mathcal{X}} \circ P_X$. Furthermore, this transition kernel is P_X -a.s. unique, in the sense that if two transition kernels κ, κ' satisfy the above properties, then $\kappa(\bullet, x) = \kappa'(\bullet, x)$ for P_X -almost every x .*

In particular any Polish space, that is, a complete separable metrizable space, equipped with its Borel σ -algebra is “nice”. The emphasis here is on *separable*, which somehow limits the “size” of the output space of the kernel.

The disintegration theorem applies to product spaces without requiring random variables, but in probabilistic terms, it is more convenient to think of the joint distribution P_{XY} of the two random variables $(X, Y) : \Omega \rightarrow \mathcal{X} \times \mathcal{Y}$, and to call the transition kernel *regular conditional probability of Y given $X = x$* , denoted $P_{Y|X}(\cdot|x)$. Thus to reiterate (0.3), the property characterizing an rcpd is that it is a transition probability satisfying:

**

$$\text{For any integrable } f : \int_{\mathcal{X} \times \mathcal{Y}} f(x, y) P_{XY}(dx, dy) = \int_{\mathcal{X}} \int_{\mathcal{Y}} f(x, y) P_{Y|X}(dy|x) P_X(dx), \quad (0.4)$$

and since it is sufficient to check it for indicators of a product of events, equivalently an rcpd is characterized by

$$\text{For any events } A \text{ on } \mathcal{X} \text{ and } B \text{ on } \mathcal{Y} : \quad P_{XY}(A \times B) = \int_A P_{Y|X}(B|x) P_X(dx). \quad (0.5)$$

Remark 0.3. In the definition above, it is important to notice that there is no unicity: in fact, given any two regular transition probabilities κ, κ' from \mathcal{X} to \mathcal{Y} , and $P_{\mathcal{X}}$ a distribution on \mathcal{X} , we check from (0.3) that $\kappa \circ P_{\mathcal{X}} = \kappa' \circ P_{\mathcal{X}}$ as soon as $\kappa(\bullet, x)$ and $\kappa'(\bullet, x)$ coincide $P_{\mathcal{X}}$ -a.s. Similarly, if $P_{\mathcal{Y}|\mathcal{X}}$ is a rcpd of P , modifying $x \mapsto P_{\mathcal{Y}|\mathcal{X}}(\bullet, x)$ on $P_{\mathcal{X}}$ -null set gives another rcpd of P . In this sense a regular conditional distribution is only defined a.s. with respect to the marginal of the conditioning variable. This means that there is no good, absolute definition of “the conditional distribution of Y given $X = x_0$ ” (if $P(X = x_0) = 0$) but rather a family of such conditional distributions (indexed by x_0).

But this family is unique up to a.s. equivalence; in particular, in the case where X, Y are independent, we will always implicitly consider the “canonical” choice $P_{Y|X}(\cdot|x) = P_Y(\cdot)$ for all x .

If X is a random variable $(\Omega, \mathcal{A}, P) \rightarrow (\mathcal{X}, \mathcal{A})$, and (Ω, \mathcal{A}) is a nice space, then we may apply the previous theorem to the product space $\Omega \times \mathcal{X}$ equipped with the distribution of the random variable $Z(\omega) = (\omega, X(\omega))$ (this is the distribution $\delta_X \circ P$, where δ_X is the transition kernel $\kappa(\cdot, \omega) = \delta_{X(\omega)}$). We thus obtain a regular conditional probability on Ω given X . This can be used to give (P_X -a.s.) a sense to the expression $P(A|X = x)$, where A is any event of Ω , which we will be using often; in particular the event $A \subset \Omega$ may be defined depending on further random variables Y_1, Y_2, \dots , but we don’t have to explicitly

apply the disintegration theorem on a complicated product space given by the value space of these variables, but may directly use the notation $P(\cdot|X = x_0)$ without further ado.

For the remainder of these notes, we will assume without repeating that (Ω, \mathcal{A}) is a nice space, so that the previous argument applies and all regular conditional probabilities exist.

**

Definition 0.4 (Conditional expectation). Let h be a real-valued integrable function on $\mathcal{X} \times \mathcal{Y}$ wrt. the (joint) probability distribution P . Assume that the conditions of Theorem 0.2 are met, so that an rcpd $P_{Y|X}$ exists. For any $x_0 \in \mathcal{X}$, the conditional expectation of h given $X = x_0$ is

$$\mathbb{E}[h(X, Y)|X = x_0] := \int h(x_0, y)P_{Y|X}(dy|x_0).$$

If Y is real-valued, denoting $F(x) = \mathbb{E}[Y|X = x]$, the random variable $F(X) : \omega \mapsto F(X(\omega))$ is the conditional expectation of Y given X , denoted $\mathbb{E}[Y|X]$.

Remark 0.5. Since the rcpd $P_{Y|X}(\bullet, x)$ is only unique up to P_X -a.s. equivalence, conditional expectations are also only defined uniquely up to P_X -a.s. equivalence over x_0 above. However, we will always implicitly assume that we have chosen a specific representative rcpd $P_{Y|X}(\bullet, x)$ “once and for all” and define all conditional expectations as above with respect to this particular choice. The reason why we spell this out is that we want *all* conditional expectations for *all* integrable functions to be defined with using the same common representative rcpd, which allows us to forget about the a.s. equivalence issue and write properties such as the next proposition.

**

Proposition 0.6 (Properties of conditional expectation). *The considered variables are real-valued and integrable as necessary for the statements below to make sense. A common representative rcpd $P_{Y|X}$ is implicitly chosen to define all conditional expectations below.*

(i) $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$

(ii) $\mathbb{E}[h(X, Y)|X = x_0] = \mathbb{E}[h(x_0, Y)|X = x_0]$, for P_X -almost all x_0 .

(iii) $\mathbb{E}[h(X)Y|X = x_0] = h(x_0)\mathbb{E}[Y|X = x_0]$, for P_X -almost all x_0 .

(iv) $\mathbb{E}[h(X)Y|X] = h(X)\mathbb{E}[Y|X]$, P_X -a.s.

Remark 0.7. In classical probability courses, it is common to define the conditional expectation first, in a different and more general way (conditional expectation with respect to a σ -algebra) that requires less formalism and is sometimes easier to handle. The advantage of considering rcpd's $P_{Y|X}$ is that they *are* probability distributions for each fixed value of the conditioning variable x , and we can apply without restriction all theorems of integration when integrating over the rcpd for any fixed x . Things get sometimes a little bit more awkward when starting with conditional expectations. In particular the “obvious” property (ii) above is not granted with the “usual” way of defining conditional expectations – in fact in that “usual” framework it does not even formally make sense since “usual” conditional expectations are only defined under a.s. equivalence, *separately* for each function $h_{x_0}(\cdot) := h(x_0, \cdot)$; it does not make sense to say that two functions both defined a.s. agree at a particular point. As we are using rcpd's here, the a.s. equivalence is “factored in” the choice of the common representative rcpd.

Proposition 0.8 (Conditioning and densities). *Assume (X, Y) is a couple of random variables on $\mathcal{X} \times \mathcal{Y}$ with joint probability distribution P_{XY} having density $f_{XY}(x, y)$ with respect to a reference product measure $\mu \otimes \nu$. Then the rcpd $P_{Y|X}(\cdot|x)$ coincides P_X -a.s. with the probability distribution on \mathcal{Y} having the following density wrt. ν :*

*

$$f_{Y|X}(y|x) := \begin{cases} \frac{f_{XY}(x,y)}{\int_{\mathcal{Y}} f_{XY}(x,y)\nu(dy)} & \text{if } \int_{\mathcal{Y}} f_{XY}(x,y)\nu(dy) = f_X(x) > 0; \\ f_0 & \text{if } f_X(x) = 0, \end{cases}$$

where f_0 is any fixed a priori density with respect to ν .

Observe that the second case deals with points outside of the support of P_X , and almost surely does not happen by definition; this is why the choice of f_0 does not matter.

Exercises

Exercise 0.7. Check that P is the first marginal distribution of $\kappa \circ P$, defined in (0.3).

Exercise 0.8. Prove the properties of Proposition 0.6 directly from the definition.

Exercise 0.9. Prove Proposition 0.8, forgetting about a.s. unicity. Just check that the transition kernel $\tilde{P}(\cdot, x)$ having the proposed density satisfies the fundamental properties of a rcpd.

1 Introduction to statistical learning theory (part 1): Decision theory

1.1 Mathematical formalization

In a nutshell, a learning task is formalized as a *prediction* of a certain target variable Y from the observation of an object X . The variability in these quantities is modeled via a joint probability distribution P of these objects. What needs to be formalized is:

- what is a prediction of Y from X ?
- how is the goodness of a prediction assessed?
- what would be a theoretically optimal prediction?
- how is a prediction function “learnt” from data?

In what follows $(\mathcal{X}, \mathfrak{X})$ is a measurable space (for instance \mathbb{R}^d or a subset of \mathbb{R}^d) called the *input space* and $(\mathcal{Y}, \mathfrak{Y})$ another measurable space called *label space*, most often $\mathcal{Y} = \mathbb{R}$, $\mathcal{Y} = [a, b]$, $\mathcal{Y} = \{1, \dots, K\}$, or $\mathcal{Y} = \mathbb{R}^k$. When \mathcal{Y} is a real interval the setting is typically called that of regression, and if \mathcal{Y} is a finite set, classification.

It is assumed that the variables (X, Y) have a joint distribution P over the product space, called *generative distribution*. In a prediction setting, we assume a random realization (x, y) of P ; the value of y (the *label*) is unknown to us but we would like to predict it as well as possible from the knowledge of x (the *input*, *predictor* or *covariate*). To allow for some flexibility in the sequel the prediction can take values in a space $(\tilde{\mathcal{Y}}, \tilde{\mathfrak{Y}})$ distinct from \mathcal{Y} .

**

Definition 1.1 (Prediction (or decision) function). A prediction (or decision) function is a measurable function $f : \mathcal{X} \mapsto \tilde{\mathcal{Y}}$.

Definition 1.2 (Loss function).

- A loss (or cost) function is a (measurable) function $\ell : \tilde{\mathcal{Y}} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ (loss functions taking negative values are possible but we will mostly consider the nonnegative loss case).
- The pointwise loss of a prediction function f on a realization (x, y) is $\ell(f(x), y)$.
- The *risk* or *generalization error* of f is the expected loss under a given generative distribution P :

$$\mathcal{E}_\ell(f, P) = \mathbb{E}_{(X, Y) \sim P}[\ell(f(X), Y)].$$

This is abbreviated as $\mathcal{E}(f)$ if both the loss function and the generative distribution are unambiguous. We allow possibly $\mathcal{E}(f) = \infty$.

In the sequel we will call the tuple $(\mathcal{X}, \mathcal{Y}, \tilde{\mathcal{Y}}, \ell)$ (we omit the σ -algebras for each space but they are implicit) a “prediction setting” and often assume without specifying it that we consider some given prediction setting.

** **Standard examples.**

Example 1.3 (Regression with least squares loss). Take $\mathcal{Y} = \tilde{\mathcal{Y}} = \mathbb{R}$, and $\ell(y', y) = (y - y')^2$; so $\mathcal{E}(f) = \mathbb{E}[(f(X) - Y)^2]$.

Example 1.4 (Classification). Take $\mathcal{Y} = \tilde{\mathcal{Y}} = \{1, \dots, K\}$ ($K \geq 2$). Then the *misclassification loss* is $\ell(y', y) = \mathbf{1}\{y \neq y'\}$, and $\mathcal{E}(f) = \mathbb{P}[f(X) \neq Y]$ is simply the probability of an incorrect prediction of the class.

A generalized case is *weighted classification*, where $\ell(y, y') = M_{yy'}$, M is the loss matrix with diagonal 0. This represents situations where one type of error is considered more costly than another.

A widely considered particular case is *binary classification* ($K = 2$; by contrast $K > 2$ is called *multiclass classification*). Depending on the context, it might be more convenient to encode the two classes as $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{-1, 1\}$. In the binary classification setting, often the prediction function f takes real values ($\tilde{\mathcal{Y}} = \mathbb{R}$). We can consider several losses in this case, for now we mention

- the *0 – 1 loss* or *hard loss*: $\mathcal{Y} = \{-1, 1\}$, $\tilde{\mathcal{Y}} = \mathbb{R}$ and $\ell(y', y) = \mathbf{1}\{yy' \leq 0\}$: this is equivalent to using the misclassification loss and interpreting the class prediction as $\text{sign}(y')$ ($y' = 0$ is interpreted as a classification error no matter what).
- the *quadratic loss for classification*: $\mathcal{Y} = \{0, 1\}$, $\tilde{\mathcal{Y}} = \mathbb{R}$ and $\ell(y', y) = (y - y')^2$. If the prediction y' is actually in $\{0, 1\}$, this is identical to the misclassification loss.

It is also possible to use a quadratic loss for multi-class classification, in this case we take $\tilde{\mathcal{Y}} = \mathbb{R}^K$, and $\ell(y', y) = \|y' - e_y\|^2$, where e_i is the i -th canonical basis vector. Note that the mapping $y \in \{1, \dots, K\} \mapsto e_y \in \mathbb{R}^K$ is called “one-hot” encoding for multiclass.

1.2 Optimal risk and prediction function.

** **Definition 1.5.** Consider a prediction setting $(\mathcal{X}, \mathcal{Y}, \tilde{\mathcal{Y}}, \ell)$.

The *optimal error* (also called *Bayes error*) for generating distribution P is

$$\mathcal{E}_\ell^*(P) = \inf_{f \in \mathcal{F}(\mathcal{X}, \tilde{\mathcal{Y}})} \mathcal{E}_\ell(f, P),$$

where $\mathcal{F}(\mathcal{X}, \tilde{\mathcal{Y}})$ denotes the set of (measurable) functions from \mathcal{X} to $\tilde{\mathcal{Y}}$. This is often simply abbreviated as $\mathcal{E}^*(P)$ or simply \mathcal{E}^* .

If this infimum is a minimum, an optimal prediction function f_P^* is one achieving the minimum:

$$\mathcal{E}(f_P^*) = \mathcal{E}^*(P), \text{ or } f_P^* \in \underset{f \in \mathcal{F}(\mathcal{X}, \tilde{\mathcal{Y}})}{\text{Arg Min}} \mathcal{E}_\ell(f, P).$$

It is possible to restrict the search for a prediction function to a subset (sometimes called *model* or *hypothesis class*) $\mathcal{G} \subset \mathcal{F}(\mathcal{X}, \tilde{\mathcal{Y}})$, in which case one defines correspondingly $\mathcal{E}_\mathcal{G}^*$ and (if it exists) $f_\mathcal{G}^*$ by restricting the inf or min to \mathcal{G} :

$$\mathcal{E}_{\mathcal{G}, \ell}^*(P) = \mathcal{E}_\mathcal{G}^* = \inf_{f \in \mathcal{G}} \mathcal{E}_\ell(f, P).$$

The optimal prediction function f^* (or possibly $f_\mathcal{G}^*$), generally implicitly assumed to exist, will be regarded as the target of the learning procedure.

Since the risk measures the goodness of prediction, it will be of interest to analyze how far a given prediction function f is from the optimal: thus we will often study the *excess risk* $\mathcal{E}(f) - \mathcal{E}^*$ (or possibly $\mathcal{E}(f) - \mathcal{E}_\mathcal{G}^*$, the excess risk with respect to prediction function class \mathcal{G}).

The following proposition is helpful to determine an (unrestricted) optimal prediction function.

Proposition 1.6.

For a given prediction setting $(\mathcal{X}, \mathcal{Y}, \tilde{\mathcal{Y}}, \ell)$, and a given joint generating probability P_{XY} , if a prediction function f^* satisfies

$$f^*(x) \in \underset{c \in \tilde{\mathcal{Y}}}{\text{Arg Min}} \mathbb{E}[\ell(c, Y) | X = x], \quad P_X - \text{ almost surely }, \quad (1.1)$$

then f^* is an optimal prediction function, i.e. $\mathcal{E}(f^*, P) = \mathcal{E}^*(P)$.

Proof. Assume f^* satisfies (1.1). Let $f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$, be any other decision function. and any fixed $x \in \mathcal{X}$, such that (1.1) is satisfied, let us denote $c^* = f^*(x)$ and $c = f(x)$. We have, using property (ii) of Prop. 0.6:

$$\begin{aligned} \mathbb{E}[\ell(f(X), Y) | X = x] &= \mathbb{E}[\ell(f(x), Y) | X = x] \\ &= \mathbb{E}[\ell(c, Y) | X = x] \\ &\geq \mathbb{E}[\ell(c^*, Y) | X = x] \\ &= \mathbb{E}[\ell(f^*(x), Y) | X = x] \\ &= \mathbb{E}[\ell(f^*(X), Y) | X = x], \end{aligned}$$

where the inequality is because of (1.1). Now taking expectation over x , i.e. integrating with the distribution P_X , the inequality remains true after integration since it is true for P_X -almost all x ; and using point (i) of Prop. 0.6, we obtain

$$\begin{aligned}
\mathcal{E}(f^*, P_{XY}) &= \mathbb{E}[\ell(f^*(X), Y)] \\
&= \mathbb{E}[\mathbb{E}[\ell(f^*(X), Y)|X]] \\
&= \int_{\mathcal{X}} \mathbb{E}[\ell(f^*(X), Y)|X = x]P_X(dx) \\
&\leq \int_{\mathcal{X}} \mathbb{E}[\ell(f(X), Y)|X = x]P_X(dx) \\
&= \mathbb{E}[\mathbb{E}[\ell(f(X), Y)|X]] \\
&= \mathcal{E}(f, P_{XY}).
\end{aligned}$$

□

The interest of the above result is that it allows to reduce the problem of finding an optimal prediction *function* to the case of a *constant* prediction for an arbitrary probability distribution on \mathcal{Y} . If we understand how to choose an optimal constant prediction c^* for the “simpler” problem where we look at the prediction error $\ell(c^*, Y)$, averaged over some fixed but arbitrary marginal distribution P_Y , and how c^* depends on P_Y , then we deduce how to construct an optimal decision function $f^*(x)$ for the full problem, namely for any given $x \in \mathcal{X}$ we take the optimal constant decision for the conditional distribution $P_{Y|X=x}$ on \mathcal{Y} .

(Important) examples.

Proposition 1.7. *Under the setting of regression with quadratic loss, provided Y is square integrable under P_Y , the optimal prediction function is*

$$f^*(x) = \mathbb{E}[Y|X = x],$$

And the corresponding risk

$$\mathcal{E}^* = \text{Var}[Y|X] = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2],$$

and for the excess risk of an arbitrary prediction function f it holds

$$\mathcal{E}(f) - \mathcal{E}^* = \mathbb{E}[(f(X) - f^*(X))^2] = \|f - f^*\|_{2, P_X}^2.$$

Proof. Following Proposition 1.6, and the remark following it, we analyze the simple case of a given probability distribution P_Y on \mathbb{R} (for which Y is square integrable) and of a

constant prediction c . We have

$$\mathbb{E}[(Y - c)^2] = \mathbb{E}[(Y - \mathbb{E}[Y])^2] + (\mathbb{E}[Y] - c)^2,$$

so that $c^* = \mathbb{E}_{Y \sim P_Y}[Y]$ is an optimal constant prediction for distribution P_Y .

Now for a joint distribution P_{XY} such that $\mathbb{E}[Y^2] < \infty$, since $\mathbb{E}[Y^2] = \mathbb{E}[\mathbb{E}[Y^2|X]]$, it must be the case that $\mathbb{E}[Y^2|X]$ is a.s. finite, in other words, P_X -almost surely $P_{Y|X}(\cdot|x)$ has a second moment. For any fixed x we can therefore apply the previous argument to $P_{Y|X=x}$ and conclude that $f^*(x) = \mathbb{E}[Y|X = x]$ satisfies (1.1). The formula for the risk follows. As for the excess risk, we have

$$\begin{aligned} \mathcal{E}(f) &= \mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y)^2] \\ &= \mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] \quad (1.2) \\ &= \mathbb{E}[(f(X) - f^*(X))^2] + \mathcal{E}^*, \end{aligned}$$

rearranging gives the last claim of the proposition. \square

Exercise 1.1. Justify equality (1.2).

Proposition 1.8. *Consider the classification setting with K classes and the misclassification loss. Then a prediction function f^* with*

$$f^*(x) \in \underset{y \in \{1, \dots, K\}}{\text{Arg Max}} P(Y = y|X = x)$$

is an optimal classification rule. This is known as a Bayes classifier and the resulting \mathcal{E}^ as the Bayes error rate.*

Proof. Using Proposition 1.6, we analyze the simple case of a given probability distribution P_Y on $\{1, \dots, K\}$ and of a constant prediction c . Then

$$\mathbb{E}_{Y \sim P_Y}[\mathbf{1}\{Y \neq c\}] = 1 - P_Y(\{c\}).$$

Obviously this is minimized for $c \in \underset{y \in \{1, \dots, K\}}{\text{Arg Max}} P_Y(\{y\})$. Now applying this argument to $P_{Y|X=x}$ for any x , and using Proposition 1.6, this gives the claim. \square

The following two examples, presented as exercises, are fundamental and should not be skipped.

**

Exercise 1.2. Consider the classification setting with the 0-1 loss. Assume the *class-conditional distributions* $\mathbb{P}_{X|Y}[\bullet|Y = i]$, $i = 1, \dots, K$, have respective densities f_i with respect to a common reference measure μ on \mathcal{X} . Denote also $\pi_i := \mathbb{P}_Y[Y = i]$. Prove that a prediction function g^* such that

$$g^*(x) \in \underset{i \in \{1, \dots, K\}}{\text{Arg Max}} (\pi_i f_i(x))$$

is a Bayes classification function.

Hint: Prove that the distribution of Y given $X = x$ is given by the discrete distribution

$$P(Y = i|X = x) = \frac{\pi_i f_i(x)}{\sum_{j=1}^K \pi_j f_j(x)}, \quad j = 1, \dots, K.$$

To establish this, you can use Prop. 0.8 on the product space $\mathcal{X} \times \mathcal{Y}$, and determine what is the density of P_{XY} with respect to $\mu \otimes \nu$, where ν is the counting measure on Y (i.e. $\nu(\{i\}) = 1, i = 1, \dots, K$).

** *Exercise 1.3.* What is the optimal prediction function for the setting of classification using real-valued prediction and the quadratic loss (in the binary and in the multiclass-classification case)?

Hint: The answer is $f^*(x) = \mathbb{P}[Y = 1|X = x]$ in the binary classification case with $\mathcal{Y} = \{0, 1\}$ and $\tilde{\mathcal{Y}} = \mathbb{R}$; and $f^*(x) = (\mathbb{P}[Y = i|X = x])_{i=1, \dots, K}$ in the multiclass-classification case, where we recall that $\tilde{Y} = \mathbb{R}^K$, see Example 1.4 for the definition of the settings.

1.3 Learning from data

We now consider a method for learning a decision function from data; this is also called “estimator” in statistical terms.

In what follows, available observed data will be referred to as a *training sample* S_n , which is a n -uple $(x_1, y_1), \dots, (x_n, y_n)$.

** **Definition 1.9.** Given an integer $n \in \mathbb{N}_{>0}$, an estimator (for the decision function) acting on training samples of size n is a mapping

$$\begin{aligned} (\mathcal{X} \times \mathcal{Y})^n &\rightarrow \mathcal{F}(\mathcal{X}, \tilde{\mathcal{Y}}) \\ S_n &\mapsto \hat{f}_{S_n} \end{aligned}$$

such that $(S_n, x) \mapsto \hat{f}_{S_n}(x)$ is (jointly) measurable.

Note: it is usual practice in statistics that quantities with a hat-notation are estimators, i.e. are implicitly functions of the data (sample); the explicit dependence is often dropped from the notation.

The generalization error or risk of an estimator is $\mathcal{E}(\hat{f}_{S_n}) = \mathbb{E}_{X,Y}[\ell(\hat{f}_{S_n}(X), Y)]$; observe that in this notation the sample S_n is considered as fixed and we take the expectation with respect to a “new” point (X, Y) , hence the name “generalization error”. If the sample data is modeled as random, this notation means that we are implicitly considering a conditional expectation conditional to the sample, and that the new point (X, Y) is independent of the sample S_n . In the latter case (random sample), the expected generalization error/risk

of estimator \hat{f} is $\mathbb{E}_{S_n \sim P^{\otimes n}} [\mathcal{E}(\hat{f}_{S_n})]$, i.e. a double expectation of the loss over the training sample and the new independent point (X, Y) . (We will always assume in these notes that the training sample is i.i.d. with the marginal distribution P).

By contrast, the *training* or *empirical error* of an estimator \hat{f} is

**

$$\hat{\mathcal{E}}(\hat{f}) = \hat{\mathcal{E}}(\hat{f}_{S_n}, S_n) := \frac{1}{n} \sum_{i=1}^n \ell(\hat{f}_{S_n}(X_i), Y_i) = \mathcal{E}(\hat{f}_{S_n}, \hat{P}_n),$$

where $\hat{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i, Y_i}$ is the *empirical distribution* associated to the sample S_n .

Observe the crucial fact that the empirical error “doubly” depends on the sample S_n : first because the sample determines the estimator \hat{f}_{S_n} , second because the error is evaluated on the *same* sample.

In particular, in general, $\mathbb{E}_{S_n \sim P^{\otimes n}} [\hat{\mathcal{E}}(\hat{f}_{S_n})] \neq \mathbb{E}_{S_n \sim P^{\otimes n}} [\mathcal{E}(\hat{f}_{S_n})]$. This is to be contrasted with the fact that, for a *fixed* (non data-dependent) prediction function f , we have $\mathbb{E}_{S_n \sim P^{\otimes n}} [\hat{\mathcal{E}}(f)] = \mathcal{E}(f)$ by simple linearity of the expectation.

(Counter)Example: overfitting. Consider the classification setting and assume $X \sim \text{Unif}[0, 1]$ and $P(Y = 1) = P(Y = 0) = \frac{1}{2}$, with Y independent of X . Then, it is obvious that $\mathcal{E}(f) = \mathbb{P}[f(X) \neq Y] = \frac{1}{2}$ for *any* prediction function f . On the other hand, if $(X_i, Y_i)_{i=1, \dots, n}$ is i.i.d. from this generating distribution, almost surely the points X_i are distinct, and we can a.s. pick a function \hat{f} such that $\hat{f}(X_i) = Y_i$ for all i . Then $\hat{\mathcal{E}}(\hat{f}) = 0$, yet $\mathcal{E}(\hat{f}) = \frac{1}{2}$ almost surely. This is an extreme case of what is known as the *overfitting* phenomenon. In general, some overfitting will happen (i.e. $\hat{\mathcal{E}}(\hat{f})$ will, generally, be in expectation smaller than $\mathbb{E}[\mathcal{E}(f)]$; i.e. will have a negative bias, in statistical terms). One of the goals of the course is to control the amount of overfitting from a theoretic point of view. The above example suggests that we should limit ourselves in the possible choice of prediction function by considering appropriate models $\mathcal{G} \subsetneq \mathcal{F}(\mathcal{X}, \tilde{\mathcal{Y}})$.

Empirical risk minimization (ERM). With the previous caveat in mind, a general approach to construct an estimator \hat{f}_{ERM} is to minimize the empirical risk given a given model $\mathcal{G} \subsetneq \mathcal{F}(\mathcal{X}, \tilde{\mathcal{Y}})$ (the model \mathcal{G} is assumed to be given in the context and omitted from the notation):

**

$$\hat{f}_{ERM} \in \underset{f \in \mathcal{G}}{\text{Arg Min}} \hat{\mathcal{E}}(f).$$

An example of such an estimator from classical statistics is the *maximum likelihood estimator*. In this setting there is no actual “prediction”, but we can formally make it

enter the considered framework by taking $\mathcal{Y} = \{0\}$, $\tilde{\mathcal{Y}} = \mathbb{R}$, the possible “prediction” functions are densities $f : \mathcal{X} \rightarrow \mathbb{R}_+$ with respect to some reference measure μ on \mathcal{X} , and models are of the form $\{f_\theta, \theta \in \Theta\}$ for some index space Θ (often a subset of \mathbb{R}^k : we then speak of a parametric model). The loss function is $\ell(y') = -\log(y')$ (allowed in this case to take negative values), and we have

$$\hat{f}_{ML} \in \text{Arg Min}_{\theta \in \Theta} \sum_{i=1}^n -\log f_\theta(x_i).$$

Another example from classical statistics is that of *ordinary least squares linear regression (OLS)*: in the regression ($\mathcal{X} = \mathbb{R}^d$) with the squared loss setting, consider the model $\mathcal{G} = \{f_\beta(x) = \langle \beta, x \rangle | \beta \in \mathbb{R}^d\}$, the ERM over that model is $f_{\hat{\beta}_{OLS}}$ with

*

$$\hat{\beta}_{OLS} \in \text{Arg Min}_{\beta \in \mathbb{R}^d} \sum_{i=1}^n (y_i - \langle \beta, x_i \rangle)^2. \quad (1.3)$$

The “population” analogue (i.e. the optimal theoretical predictor over the same model \mathcal{G}) is

$$\beta_{OLS}^* \in \text{Arg Min}_{\beta \in \mathbb{R}^d} \mathbb{E}[(Y - \langle \beta, X \rangle)^2],$$

in other words we have $f_{\mathcal{G}}^* = f_{\beta_{OLS}^*}$ (see Definition 1.5). Since

$$\mathbb{E}[(Y - \langle \beta, X \rangle)^2] = \mathbb{E}[Y^2] + \beta^t \Sigma \beta - 2\beta^t \gamma,$$

with $\Sigma := \mathbb{E}[XX^t]$ (the *second moment matrix*) and $\gamma := \mathbb{E}[XY]$, the solution is $\beta_{OLS}^* := \Sigma^{-1}\gamma$ by classical formulas for the minimum of a quadratic function (assuming Σ invertible). To compute $\hat{\beta}_{OLS}$ we replace the generating distribution P_{XY} by the empirical distribution \hat{P}_n from the observed sample, therefore

*

$$\hat{\beta}_{OLS} = \hat{\Sigma}^{-1}\hat{\gamma}, \text{ where } \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^n x_i x_i^T, \text{ and } \hat{\gamma} := \frac{1}{n} \sum_{i=1}^n x_i y_i; \quad (1.4)$$

(here we assume $\hat{\Sigma}$ invertible) a classical equivalent form is $\hat{\beta}_{OLS} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$, where \mathbf{X} is the (n, d) matrix whose rows are x_1^t, \dots, x_n^t , and $\mathbf{Y} = (y_1, \dots, y_n)^t$.

Exercise 1.4. Justify the validity of the last form for $\hat{\beta}_{OLS}$.

1.4 Consistency of estimators

A primary goal of the analysis of statistical learning estimators is to understand their behavior as the number n of data grows to infinity. A fundamental property is to ensure

asymptotical convergence of their risk to the optimal risk \mathcal{E}^* (possibly the optimal risk \mathcal{E}^* over a model \mathcal{G}). This is called *consistency* and is made more formal in the following definition.

**

Definition 1.10 (Consistency). Let $\hat{f}^{(n)}$, $n \geq 1$ be a sequence of estimators (where $\hat{f}^{(n)}$ learns from a sample S_n of size n) for a prediction problem with loss function ℓ .

Let \mathcal{P} be a set of joint generating distributions on $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{G} \subseteq \mathcal{F}(\mathcal{X}, \tilde{\mathcal{Y}})$ a subset of prediction functions.

Then the sequence $\hat{f}^{(n)}$, $n \geq 1$ is *consistent in expectation* on the distribution model \mathcal{P} and the prediction model \mathcal{G} , if

$$\forall P \in \mathcal{P} : \quad \limsup_{n \rightarrow \infty} \mathbb{E}_{S_n \sim P^{\otimes n}} \left[\mathcal{E}(\hat{f}_{S_n}^{(n)}, P) \right] \leq \mathcal{E}_{\mathcal{G}}^*(P).$$

It is *consistent in probability* if

$$\forall P \in \mathcal{P}, \forall \varepsilon > 0 : \quad \mathbb{P}_{S_n \sim P^{\otimes n}} \left[\mathcal{E}(\hat{f}_{S_n}^{(n)}, P) \geq \mathcal{E}_{\mathcal{G}}^*(P) + \varepsilon \right] \rightarrow 0, \text{ as } n \rightarrow \infty.$$

Finally it is *consistent almost surely* if for any $P \in \mathcal{P}$, an i.i.d. sequence $(X_i, Y_i)_{i \geq 1}$ from distribution P and $S_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ it holds

$$\limsup_{n \rightarrow \infty} \mathcal{E}(\hat{f}_{S_n}^{(n)}, P) \leq \mathcal{E}_{\mathcal{G}}^*(P), \text{ almost surely.}$$

The sequence of estimators is called *universally consistent* if the above holds for $\mathcal{P} =$ all joint distributions on $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{G} =$ all prediction functions.

In the case where we analyze consistency over a specific class $\mathcal{G} \subsetneq \mathcal{F}(\mathcal{X}, \tilde{\mathcal{Y}})$ of prediction functions, it is generally also assumed that the estimator \hat{f} outputs predictor functions belonging to \mathcal{G} , though it does not have to be the case.

Simple example. Consider the regression with quadratic loss setting, where \mathcal{X} is reduced to a singleton $\{0\}$, so that decision functions are given by a constant $\theta \in \mathbb{R}$, and the optimal prediction θ^* is just the (marginal) expectation $\mathbb{E}[Y]$. Then the ERM is the empirical average $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n y_i$, which is consistent by the law of large numbers.

Consistency of ERM over a finite class.

Proposition 1.11. Let $\mathcal{G} \subsetneq \mathcal{F}(\mathcal{X}, \tilde{\mathcal{Y}})$ be a finite set of prediction functions and $\hat{f}_{ERM}^{(n)}$ be the ERM estimator on the set \mathcal{G} using a sample of size n . Then $(\hat{f}_{ERM}^{(n)})_{n \geq 1}$ is almost surely consistent over \mathcal{G} and for all probability distributions on $\mathcal{X} \times \mathcal{Y}$.

To establish this property, we first state the following elementary lemma:

Lemma 1.12. *Let $(Z_n^{(k)})_{k \in \{1, \dots, K\}, n \geq 1}$ be a finite family of sequences of real random variables such that $\forall k \in \{1, \dots, K\} : Z_n^{(k)} \rightarrow 0$ almost surely as $n \rightarrow \infty$. Then $U_n = \sup_{k \in \{1, \dots, K\}} |Z_n^{(k)}| \rightarrow 0$ almost surely as $n \rightarrow \infty$.*

Proof. Let A_k denote the event $\left\{ \lim_{n \rightarrow \infty} Z_n^{(k)} = 0 \right\}$. By assumption, $\mathbb{P}[A_k] = 1$, so $\mathbb{P}[A_k^c] = 0$, and therefore by the union bound

$$1 \geq \mathbb{P}\left[\bigcap_{k=1}^K A_k\right] = 1 - \mathbb{P}\left[\bigcup_{k=1}^K A_k^c\right] \geq 1 - \sum_{k=1}^K \mathbb{P}[A_k^c] = 1,$$

so $\mathbb{P}\left[\bigcap_{k=1}^K A_k\right] = 1$. Furthermore, for any ω in the event $\bigcap_{k=1}^K A_k$, by definition it holds that $\forall k \in \{1, \dots, K\} : Z_n^{(k)}(\omega) \rightarrow 0$ (in the usual deterministic sense of sequence convergence). By standard arguments on (deterministic) sequence convergence, this implies $\sup_{k \in \{1, \dots, K\}} |Z_n^{(k)}| \rightarrow 0$ as $n \rightarrow \infty$, which therefore also happens with probability 1. \square

Note: this lemma is not true as soon as one considers a countably infinite family of sequences of random variables. Can you point where the argument in the proof fails – is it for the “probabilistic” part or for the “deterministic” (sequence convergence) part?

Proof of Proposition 1.11. We write $\mathcal{G} = \{f_1, \dots, f_k\}$. We will only spell out the proof in the case where $\mathcal{E}(f_k) < \infty$ for all $k \in \{1, \dots, K\}$. In this situation, for any $k \in \{1, \dots, K\}$ we have $\widehat{\mathcal{E}}(f_k, S_n) \rightarrow \mathcal{E}(f_k)$ almost surely, by the law of large numbers, since $\widehat{\mathcal{E}}(f_k, S_n) = \frac{1}{n} \sum_{i=1}^n W_i^{(k)}$, with $W_i^{(k)} = \ell(f^{(k)}(X_i), Y_i)$, $i \in \mathbb{N}$, which are i.i.d. random variables with expectation $\mathcal{E}(f_k)$. By Lemma 1.12, it follows that almost surely,

$$\sup_{k \in \{1, \dots, K\}} |\widehat{\mathcal{E}}(f_k) - \mathcal{E}(f_k)| \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Now, remember that we *cannot* apply the law of large numbers to $\widehat{\mathcal{E}}(\widehat{f}_{ERM}^{(n)}, S_n)$ since it is an average of variables which are *not* i.i.d. because on the “double” dependence on the sample S_n .

However, by definition of the ERM estimator: (a) $\widehat{f}_{ERM}^{(n)} \in \mathcal{G}$ and (b) $\widehat{\mathcal{E}}(\widehat{f}_{ERM}^{(n)}) \leq \widehat{\mathcal{E}}(f_k)$, for any $k \in \{1, \dots, K\}$. Furthermore, let us denote k^* an index such that $\mathcal{E}(f_{k^*}) = \mathcal{E}_{\mathcal{G}}^*$. Using these properties, it holds

$$\begin{aligned} 0 \leq \mathcal{E}(\widehat{f}_{ERM}^{(n)}) - \mathcal{E}_{\mathcal{G}}^* &= \left(\mathcal{E}(\widehat{f}_{ERM}^{(n)}) - \widehat{\mathcal{E}}(\widehat{f}_{ERM}^{(n)}) \right) + \underbrace{\left(\widehat{\mathcal{E}}(\widehat{f}_{ERM}^{(n)}) - \widehat{\mathcal{E}}(f_{k^*}) \right)}_{\leq 0} + \left(\widehat{\mathcal{E}}(f_{k^*}) - \mathcal{E}(f_{k^*}) \right) \\ &\leq 2 \sup_{f \in \mathcal{G}} |\widehat{\mathcal{E}}(f) - \mathcal{E}(f)|, \end{aligned} \tag{1.5}$$

and we have seen that the latter quantity converges to 0 almost surely, hence the conclusion.

In the case where all prediction functions $f \in \mathcal{G}$ have infinite risk, $\mathcal{E}_{\mathcal{G}}^* = \infty$ and there is nothing to prove. In the case where *some* prediction functions $f \in \mathcal{G}$ have infinite risk, note that if $\mathcal{E}(f) = \infty$, for a fixed prediction function f , then the law of large numbers implies $\lim_{n \rightarrow \infty} \widehat{\mathcal{E}}(f, S_n) = \infty$ almost surely. The above arguments still work in this case with minor adaptations which are left to the reader.

Please observe that inequality (1.5) is fundamental and will be used again later for more detailed analysis of the ERM.

◇ **More elaborate example of consistency analysis: regressogram.** We will analyze consistency in a fully nonparametric regression model with the generating distribution (X, Y) such that

$$Y = f^*(X) + \varepsilon, \quad \mathbb{E}[\varepsilon] = 0, \quad \mathbb{E}[\varepsilon^2] = \sigma^2, \quad \varepsilon \perp\!\!\!\perp X, \quad (1.6)$$

where we will consider $\mathcal{X} = [0, 1]$; note that the marginal distribution P_X is left totally arbitrary by the above model. We will make the assumption that f^* is a L -Lipschitz function. Recall that in the case of regression with quadratic loss, $f^*(x) = \mathbb{E}[Y|X = x]$ is the optimal predictor.

The estimator \widehat{f} we will consider based on the observed sample $S_n = (X_i, Y_i)_{1 \leq i \leq n}$ will be a piecewise constant function on the equal width intervals $I_j := \left[\frac{j-1}{K}, \frac{j}{K}\right)$, $j = 1, \dots, K-1$, $I_K = \left[\frac{K-1}{K}, 1\right]$; where the choice of the number K of intervals will be determined later. Putting $N_j := \{1 \leq i \leq n : X_i \in I_j\}$, we define the regressogram estimator with K bins as

$$\widehat{f}(x) := \sum_{j=1}^K \mathbf{1}\{x \in I_j\} \widehat{a}_j; \quad \widehat{a}_j := \frac{1}{|N_j| + 1} \sum_{i \in N_j} Y_i, j = 1, \dots, K. \quad (1.7)$$

We will prove the following property of the above estimator.

Proposition 1.13. *Under the regression model (1.6) for $\mathcal{X} = [0, 1]$ with f^* Lipschitz, the regressogram estimator \widehat{f}_n given by (1.7) with $K(n)$ bins from a sample of size n :*

- *is consistent in average risk, i.e. satisfies $\mathbb{E}_{S_n} [\mathcal{E}(\widehat{f}_n) - \mathcal{E}^*] \rightarrow 0$, as $n \rightarrow \infty$, provided $K(n) \rightarrow \infty$ but $K(n) = o(n)$ as $n \rightarrow \infty$;*
- *satisfies $\mathbb{E}_{S_n} [\mathcal{E}(\widehat{f}_n) - \mathcal{E}^*] = \mathcal{O}(n^{-2/3})$ if $K(n) \sim n^{1/3}$.*

Proof. To analyze the averaged quadratic risk of \widehat{f} we write

$$\mathbb{E}_{S_n} [\mathcal{E}(\widehat{f}) - \mathcal{E}^*] = \mathbb{E}_{S_n} \mathbb{E}_X [(\widehat{f}(X) - f^*(X))^2] = \sum_{j=1}^K \mathbb{E}_{S_n} \mathbb{E}_X [\underbrace{(\widehat{a}_j - f^*(X))^2}_{=: A_j} \mathbf{1}\{X \in I_j\}].$$

Note: we write $\mathbb{E}_{S_n} \mathbb{E}_X[\dots]$ to emphasize that the expectation is over a new independent point X and the random sample $S_n \sim P^{\otimes n}$; it would be more appropriate to write the

inner expectation as a conditional expectation conditional to S_n , but because X, S_n are independent, we use the above notation that can be understood as an iterated integral. Furthermore, according to the generative model we can decompose the expectation over S_n as an expectation over $(X_i)_{1 \leq i \leq n}$ and $(\varepsilon_i)_{1 \leq i \leq n}$; finally since all these variables are independent we can perform expectations in the order we want by Fubini's theorem.

Concentrating on one term, we have

$$\begin{aligned} A_j &= (\hat{a}_j - f^*(X))^2 \mathbf{1}\{X \in I_j\} \\ &= \left(\frac{1}{|N_j| + 1} \left(\sum_{i \in N_j} ((f^*(X_i) + \varepsilon_i) - f^*(X)) + f^*(X) \right) \right)^2 \mathbf{1}\{X \in I_j\} \\ &= \frac{1}{(|N_j| + 1)^2} \left(\underbrace{\sum_{i \in N_j} (f^*(X_i) - f^*(X))}_{=: W_j} + f^*(X) + \underbrace{\sum_{i \in N_j} \varepsilon_i}_{=: Z_j} \right)^2 \mathbf{1}\{X \in I_j\} \end{aligned}$$

Taking expectations over $(\varepsilon_i)_{1 \leq i \leq n}$ first, we notice $\mathbb{E}_{(\varepsilon_i)}[(W_j + Z_j)^2] = W_j^2 + \mathbb{E}[Z_j^2] = W_j^2 + |N_j|\sigma^2$. For $X_i, X \in I_j$ we have $|f^*(X_j) - f^*(X)| \leq L/K$ by the Lipschitz assumption on f^* , and we can also write $|f^*(X)| \leq M$ for some M , since a Lipschitz function must be bounded on a compact. All in all we have thus $W_j^2 \mathbf{1}\{X \in I_j\} \leq \left(L \frac{|N_j|}{K} + M\right)^2 \mathbf{1}\{X \in I_j\}$. We thus finally get, putting $p_j := \mathbb{P}[X \in I_j]$, and using $(a + b)^2 \leq 2(a^2 + b^2)$:

$$\begin{aligned} \mathbb{E}[A_j] &\leq \mathbb{E} \left[\frac{1}{(|N_j| + 1)^2} \left(\left(L \frac{|N_j|}{K} + M \right)^2 + |N_j|\sigma^2 \right) \mathbf{1}\{X \in I_j\} \right] \\ &\leq p_j \left(\frac{2L^2}{K^2} + (\sigma^2 + 2M^2) \mathbb{E} \left[\frac{1}{|N_j| + 1} \right] \right). \end{aligned}$$

Since $|N_j| = \sum_{i=1}^n \mathbf{1}\{X_i \in I_j\}$, it has a Binom(n, p_j) distribution and therefore satisfies $\mathbb{E}[(|N_j| + 1)^{-1}] \leq (p_j(n+1))^{-1}$ (left as an exercise). Summing over j and using $\sum_{j=1}^K p_j = 1$, we finally get

$$\mathbb{E}_{S_n} \left[\mathcal{E}(\hat{f}) - \mathcal{E}^* \right] \leq 2 \frac{L^2}{K^2} + K \frac{(\sigma^2 + 2M^2)}{n+1}. \quad (1.8)$$

The conclusion of the proposition follow from the above inequality, by plugging in $K = K(n)$ according to the assumptions of the proposition, letting n go to infinity and some standard estimates. \square

A few remarks are in order:

- The result of Proposition 1.13 holds *whatever is the marginal distribution of X* . Results in statistical learning theory generally try to avoid strong assumptions on the generative distribution; this is called a “distribution-free” approach.
- The strong assumption that was made is the Lipschitz character of the regression function f^* .

- The estimator considered is very similar to (but not quite equal to) the ERM estimator over the model \mathcal{G}_K of piecewise constant predictions on K equally sized intervals (in fact, the conclusion of the proposition would be essentially the same for the ERM estimator on \mathcal{G}_K). Thus, this is an example where it is of advantage of selecting a prediction model \mathcal{G}_K of limited “complexity” (even though we know the optimal prediction does not lie in this model), and letting the complexity grow appropriately with the sample size n .
- The two terms appearing in the bound (1.8) can be interpreted as an *approximation* term (decreasing with K and independent of n ; it comes from the fact that we can better approximate a Lipschitz function by a piecewise constant function as K grows, regardless of the sample) and an *estimation* term (growing with K but decreasing with n ; it comes from the fact that we can average out the noise variance σ^2 provided we have enough points per interval, but the average number of points per interval decreases with K). This “balancing” phenomenon is very common in statistical learning theory.

Exercise 1.5. Prove the following implications between the different types of consistency:

- Almost sure consistency implies consistency in probability.
- If the estimators $\widehat{f}^{(n)}$ take their values in the set of decision functions \mathcal{G} , consistency in expectation on \mathcal{G} implies consistency in probability on \mathcal{G} .
- If the loss function is bounded, almost sure consistency implies consistency in expectation.

Hint: consider the sequence of random variables $Z_n := \mathcal{E}(\widehat{f}_{S_n}^{(n)})$, and use fundamental relations and implications of probabilistic convergence from a probability lecture. For (b), establish and then use the fact that $Z_n - \mathcal{E}_{\mathcal{G}}^* \geq 0$. For (c), use Fatou’s lemma.

Exercise 1.6. How do the conclusions of Proposition 1.13 change if one assumes instead that f^* is α -Hölder ($\alpha \in (0, 1]$)?

Exercise 1.7. Write explicitly an ERM estimator for the model \mathcal{G}_K defined above. How does it differ from the estimator defined in (1.7)? (More technical:) how can Proposition 1.13 be adapted for the ERM estimator?

Exercise 1.8. Prove that $\widehat{f} := f_{\widehat{\beta}_{OLS}}$ is almost surely consistent in the model \mathcal{G} of linear functions. *Hint:* recall the formulae (1.4) for $\widehat{\beta}_{OLS}$. Compare to the formula for β_{OLS}^* (assume Σ invertible). Use the LLN.

1.5 Plug-in classification

In this section we analyze a simple example of a “plug-in” rule for classification, which “transfers” a decision problem to another one. Here, we transfer a regression estimate to

a classification estimate. Let us recall the classification setting:

**

Consider the binary classification case, $\mathcal{Y} = \{0, 1\}$. In this case, if we denote $\eta(x) := \mathbb{P}[Y = 1|X = x]$, the (or more precisely “a”) Bayes classifier takes the form $f^*(x) = \mathbf{1}\{\eta(x) \geq \frac{1}{2}\}$, and $\mathcal{E}^* = \mathbb{E}[\min(\eta(X), 1 - \eta(X))]$.

For various reasons it can be more convenient to estimate the probability function $\eta(x)$ rather than the classifier $f^*(x)$, for instance by using a regression method aiming at having a small risk for the quadratic loss (see Exercise 1.3 below). Let $\hat{\eta}$ be such an estimate (we assume it is given beforehand and do not specify how it was obtained). A natural way to transform this estimate to an actual classifier is to “plug in” the estimate in place of η in the formula for f^* , that is define

$$\hat{f}(x) = \mathbf{1}\left\{\hat{\eta}(x) \geq \frac{1}{2}\right\}.$$

A natural question is whether \hat{f} is a good estimate of f^* whenever $\hat{\eta}$ is a good estimate of η . Here, the “goodness” of an estimate will be measured via the *excess risk to the optimal*, for the classification risk and the squared loss risk, respectively.

**

Proposition 1.14 (Excess risk inequality for plug-in classification). *Let ℓ denote the misclassification loss and h the quadratic loss. For any function $\hat{\eta}$ estimating η , define the plug-in classifier \hat{f} as above. Then*

$$0 \leq \mathcal{E}_\ell(\hat{f}) - \mathcal{E}_\ell^* \leq 2\mathbb{E}[|\hat{\eta}(X) - \eta(X)|] \leq 2\mathbb{E}[(\hat{\eta}(X) - \eta(X))^2]^{\frac{1}{2}} = 2(\mathcal{E}_h(\hat{\eta}) - \mathcal{E}_h^*)^{\frac{1}{2}}.$$

Proof. Note that the first inequality of the claim is from definition of \mathcal{E}_ℓ^* , and the third is Jensen’s inequality. As in the proof of Proposition 1.6, it is possible to consider an argument conditionally to $X = x$ for any fixed x , only considering expectations with respect to a (conditional) probability distribution over \mathcal{Y} , then integrate the obtained pointwise inequalities over X at the end. We therefore consider x as fixed, omit the dependence in x of $\hat{f}, f^*, \hat{\eta}, \eta$, and treat them as constants.

We first establish the (sometimes useful) equality that for any classifier f , one has

$$\mathcal{E}_\ell(f) - \mathcal{E}_\ell^* = 2\mathbb{E}\left[\mathbf{1}\{f^*(X) \neq f(X)\}\left|\eta(X) - \frac{1}{2}\right|\right]. \quad (1.9)$$

Indeed, as suggested above, we condition with respect to $X = x$ and consider x as fixed. On the left-hand side, we have $\mathbb{E}[\mathbf{1}\{f \neq Y\} - \mathbf{1}\{f^* \neq Y\}|X = x]$. Assume $\eta \geq \frac{1}{2}$ (the other case is of course similar); then $f^* = 1$. If $f = f^* = 1$, then obviously

$$\mathbb{E}[\mathbf{1}\{f \neq Y\} - \mathbf{1}\{f^* \neq Y\}] = 0 = 2\mathbf{1}\{f^* \neq f\}\left|\eta - \frac{1}{2}\right|.$$

If $f = 0$, then

$$\mathbb{E}[\mathbf{1}\{f \neq Y\} - \mathbf{1}\{f^* \neq Y\}] = \eta(1 - 0) + (1 - \eta)(0 - 1) = 2\eta - 1 = 2\mathbf{1}\{f^* \neq f\} \left| \eta - \frac{1}{2} \right|.$$

Thus (1.9) is established pointwise conditionally to $X = x$, thus also in expectation.

Now, consider the specific case $\hat{f} = \mathbf{1}\{\hat{\eta} \geq \frac{1}{2}\}$. Still conditioning with respect to $X = x$, we have pointwise

$$\mathbf{1}\{f^* \neq f\} \left| \eta - \frac{1}{2} \right| \leq |\hat{\eta} - \eta|. \quad (1.10)$$

Indeed, consider again the cases $f^* = f$ (trivial since the left-hand side is 0), and $f^* \neq f$ (the inequality holds because f and f^* must be on opposite sides of $\frac{1}{2}$).

Altogether (1.9) and (1.10) (integrated over X) give the second inequality of the claim. Finally, for the last equality, notice again that, pointwise conditionally with respect to $X = x$, we have since $\eta(x) = \mathbb{E}[Y|X = x]$:

$$\mathbb{E}[(\hat{\eta} - Y)^2] - \mathbb{E}[(Y - \mathbb{E}[Y])^2] = (\hat{\eta} - \mathbb{E}[Y])^2 = (\hat{\eta} - \eta)^2,$$

(where all expectations are to be understood w.r.t. $P_{Y|X}(\cdot|x)$), giving the last equality after integration w.r.t. $X \sim P_X$. \square

Proposition 1.14 shows in particular that if the squared loss excess risk of the estimate $\hat{\eta}$ converges to zero, then so does the classification excess risk of the associated plug-in classifier rule. So convergence of $\hat{\eta}$ to the Bayes optimal decision rule η for the quadratic risk implies convergence of the plug-in estimate \hat{f} to the Bayes classifier, for the classification risk.

Remark: This implies that universal consistency of $\hat{\eta}$ implies universal consistency of the associated plug-in rule \hat{f} . However, for consistency over a specific model \mathcal{G} , things are more complicated. The result of Proposition 1.14 does **not** hold in general if we replace the optimal Bayes risks $\mathcal{E}_h^*, \mathcal{E}_\ell^*$ by the optimal risks $\mathcal{E}_{\mathcal{G},h}^*, \mathcal{E}_{\mathcal{G},\ell}^*$ over an arbitrary model \mathcal{G} of regression functions and their induced plug-in classification rules $\tilde{\mathcal{G}}$, because the proof heavily relied on the fact that we considered excess risk with respect to the (unrestricted) optimal decision, for which we had an explicit formula. For a given model \mathcal{G} , we can only expect this “transfer of consistency to plug-in classifier” property if we limit consistency to the distribution model $\mathcal{P}_{\mathcal{G}}$ containing only those distributions P such that the optimal decision belongs to \mathcal{G} . (Restricting one’s attention to a model and data distributions such that the optimal decision function belongs to that model is sometimes called *proper learning* scenario).

Exercise 1.9. Develop the reflections in the previous remark into a clean argument. Let \mathcal{G} be a subset of $\mathcal{F}(\mathcal{X}, [0, 1])$ and $\tilde{\mathcal{G}} = \{f : x \mapsto \mathbf{1}\{g(x) \geq \frac{1}{2}, g \in \mathcal{G}\}\} \subset \mathcal{F}(\mathcal{X}, \{0, 1\})$ the set of plug-in classification rules induced by \mathcal{G} . Justify that, if a sequence $\hat{\eta}^{(n)}$ of estimators are consistent (either in expectation, probability or almost surely) for the model \mathcal{G} and the set of distributions

$$\mathcal{P}_{\mathcal{G}} := \{P \in \mathcal{P}(\mathcal{X} \times \{0, 1\}) : \eta_P^* \in \mathcal{G}\},$$

(where η_P^* is the function $x \mapsto P(Y = 1|X = x)$), then the associated sequence of plug-in rules $\hat{f}^{(n)}$ is consistent for the model $\tilde{\mathcal{G}}$ and the same set of distributions. (Use directly the result of Proposition 1.14.)

On the other hand, find a counter-example as simple as possible showing that the result of Proposition 1.14 does not hold if we replace unrestricted optimal risks $\mathcal{E}_h^*, \mathcal{E}_\ell^*$ by the optimal risks $\mathcal{E}_{\mathcal{G},h}^*, \mathcal{E}_{\tilde{\mathcal{G}},\ell}^*$ over models $\mathcal{G}, \tilde{\mathcal{G}}$, even if we assume $\hat{\eta} \in \mathcal{G}$. *Hint: it is sufficient to exhibit a class \mathcal{G} and a distribution P for which the optimal decision $\eta_{P,\mathcal{G}}^*$ for the quadratic risk is such that the associated plug-in rule is not the optimal classifier in $\tilde{\mathcal{G}}$. For this it is enough to consider a 2-point space \mathcal{X} and a suitable class \mathcal{G} with just two elements.*

1.6 Negative results: the “no free lunch” theorem

In this section we show that we cannot hope to find a classification algorithm satisfying a non-trivial “universal” bound on the classification risk when trained on a sample of size n , for any n . Here “universal” means that the bound would hold for *any* data generating distribution.

Theorem 1.15 (“No free lunch”). *Let \mathcal{X} be a set of infinite cardinality, $\mathcal{Y} = \tilde{\mathcal{Y}} = \{0, 1\}$, and the loss function is the 0-1 classification loss. Then for any $n \in \mathbb{N}_{>0}$, and any classification estimator \hat{f} acting on training samples of size n , it holds*

$$\sup_{P \in \mathcal{P}_{XY}: \mathcal{E}^*(P)=0} \left(\mathbb{E}_{S_n \sim P^{\otimes n}} \left[\mathcal{E}(\hat{f}_{S_n}, P) \right] \right) \geq \frac{1}{2},$$

where \mathcal{P}_{XY} denotes the set of all probability distributions on $\mathcal{X} \times \mathcal{Y}$.

Before proving the theorem, let us reflect about what it says.

- 1/2 is the risk of a classifier function that would guess the class completely at random whatever the situation (such a “randomized” classifier would not strictly speaking enter into our framework, which only allows for fixed decision functions, but it is easy to see how to extend the setting to accommodate decision functions that include a random component). Thus, whatever the learning algorithm and the sample size, we can find a data distribution P such that our algorithm cannot be any better by any margin $\varepsilon > 0$ than the “stupid” random guess rule that learns nothing (although the Bayes optimal classifier for P has zero error).
- should it mean that trying to learn anything is always doomed to failure? No, it just means that aiming at a “universal learning” having non-trivial risk for any data generating distribution P and fixed sample size n is too ambitious. In order to get non-trivial risk bounds for fixed n , it will be necessary to restrict the class of possible

data generating distributions (for example, by assuming that the optimal decision function lies in a certain restricted class \mathcal{G}).

- should it mean that “universal consistency” is impossible? It may seem surprising at first, but it does not. Namely, in the above theorem n is fixed, so that if we require the risk to be larger than $1/4$ (say), the distribution P_n that will make learning algorithms “fail” by having risk larger than $1/4$ depends on n . It does not preclude that there exists a learning algorithm such that for any *fixed* data generating distribution P , the risk sequence $R_n(P) = \mathcal{E}(\hat{f}^{(n)}, P)$ converges to zero. In fact there exists such algorithms (at least for $\mathcal{X} = [0, 1]^d$). What the theorem says however, is that we can find P such that the n -th term of the sequence $R_n(P)$ is arbitrarily close to $1/2$ (which does not prevent the sequence to eventually converge to zero).
- in fact, this theorem seems quite intuitive and almost obvious. If we are allowed any data generating distribution, we can imagine a distribution drawing uniformly at random a point in a finite subset of \mathcal{X} of size $m \gg n$. Since the training sample is of size n , we can only observe the true classification function on a set of probability at most $n/m \ll 1$, and on the complementary of that set the true decision function can be arbitrary and we have no information about it, so how could we hope to be better than random guessing on that part? And this is actually how the proof works, though how to capture this mathematically is enlightening (and also provides insightful ideas for proving less obvious and more interesting related results to establish lower bounds on the risk in a “worst-case” sense).

Proof of Theorem 1.15. We will use the (now somewhat standard) compact notation $\llbracket k \rrbracket = \{1, \dots, k\}$ for a positive integer k .

Assume that the sample size n , and the classification estimator (learning algorithm) \hat{f} is fixed. Let m be some integer greater than n , whose value will be discussed later. Since \mathcal{X} is infinite, we can find a subset $\mathcal{X}_m = \{t_1, \dots, t_m\}$ of size m in \mathcal{X} . For any $r = (r_1, \dots, r_m) \in \{0, 1\}^m$, let P_r be the probability distribution on $\mathcal{X} \times \{0, 1\}$ such that:

- X is drawn uniformly at random in the set \mathcal{X}_m ;
- $P_r(Y = 1 | X = t_i) = r_i$ for any $i \in \llbracket m \rrbracket$ — note that since $r_i \in \{0, 1\}$, this implies $Y = r_i$ (almost surely) conditional to $X = t_i$.

Note that $\mathcal{E}^*(P_r) = 0$ for any r .

In order to represent more conveniently the draw of X we will assume that I is a uniform random variable in $\llbracket m \rrbracket$ and $X = t_I$, so that $P_r(Y = 1 | I = i) = P_r(Y = 1 | X = t_i) = r_i$. For independent draws (X_i) , $i \in \llbracket n \rrbracket$ similarly we will assume underlying i.i.d. uniform variables I_1, \dots, I_n such that $X_k = t_{I_k}$. With this representation because $r_i \in \{0, 1\}$, under P_r we have $Y_k = r_{I_k}$, $k \in \llbracket n \rrbracket$ and $Y = r_I$ (almost surely).

The following step is a key idea for lower risk bounds. Rather than trying to find a worst-case distribution for the estimator \hat{f} , we will consider its risk on average over the

family P_r when r is itself drawn at random. Clearly, the worst-case risk can only be higher than this average:

$$\begin{aligned} \sup_{P \in \mathcal{P}_{XY}: \mathcal{E}^*(P)=0} \left(\mathbb{E}_{S_n \sim P^{\otimes n}} \left[\mathcal{E}(\widehat{f}_{S_n}, P) \right] \right) &\geq \sup_{r \in \{0,1\}^n} \left(\mathbb{E}_{S_n \sim P_r^{\otimes n}} \left[\mathcal{E}(\widehat{f}_{S_n}, P_r) \right] \right) \\ &\geq \mathbb{E}_R \mathbb{E}_{S_n \sim P_R^{\otimes n}} \left[\mathcal{E}(\widehat{f}_{S_n}, P_R) \right]. \end{aligned}$$

where the outer expectation is over R drawn uniformly at random over $\{0,1\}^n$. We now rewrite and bound the right-hand side:

$$\mathbb{E}_R \mathbb{E}_{S_n \sim P_R^{\otimes n}} \left[\mathcal{E}(\widehat{f}_{S_n}, P_R) \right] = \mathbb{E}_R \mathbb{E}_{S_n \sim P_R^{\otimes n}} \mathbb{E}_{(X,Y) \sim P_R} \left[\mathbf{1} \left\{ \widehat{f}_{S_n}(X) \neq Y \right\} \right] = \mathbb{P} \left[\widehat{f}_{S_n}(X) \neq Y \right],$$

where the probability \mathbb{P} is over the draw of everything (the vector R is uniform in $\{0,1\}^n$, and conditional to R , $(S_n, (X, Y)) \sim P_R^{\otimes(n+1)}$). Next, by the law of total probability, we can write the latter probability as a sum over all possible values of the covariates $(X_k)_{k \in \llbracket n \rrbracket}$ (or rather of the corresponding indices $(I_k)_{k \in \llbracket n \rrbracket}$, remember $X_k = t_{I_k}$) observed in the training sample S_n :

$$\mathbb{P} \left[\widehat{f}_{S_n}(X) \neq Y \right] = \sum_{(i_1, \dots, i_n) \in \llbracket m \rrbracket^n} \mathbb{P} \left[\widehat{f}_{S_n}(X) \neq Y \mid I_1 = i_1, \dots, I_n = i_n \right] \mathbb{P} \left[I_1 = i_1, \dots, I_n = i_n \right].$$

We now bound the conditional probability in each of the terms of the previous sum, let us therefore consider a particular n -uple of indices $(i_1, \dots, i_n) \in \llbracket m \rrbracket^n$ (note that there may be repeated indices); to simplify notation denote the *event* $A := \{I_k = i_k, k \in \llbracket n \rrbracket\}$, and $\mathcal{I} = \{i_k, k \in \llbracket m \rrbracket\}$ the *set* of indices of observed elements (we insist that this is non-random subset, since we fixed a particular value of (i_1, \dots, i_n)).

By nested conditioning it holds

$$\begin{aligned} \mathbb{P} \left[\widehat{f}_{S_n}(X) \neq Y \mid A \right] &\geq \mathbb{P} \left[\widehat{f}_{S_n}(X) \neq Y, I \notin \mathcal{I} \mid A \right] \\ &= \mathbb{P} \left[\widehat{f}_{S_n}(X) \neq Y \mid I \notin \mathcal{I}, A \right] \mathbb{P} \left[I \notin \mathcal{I} \mid A \right]. \end{aligned} \quad (1.11)$$

We claim that, conditional to the event A and $I \notin \mathcal{I}$, label $Y = R_I$ is drawn at random with probability $1/2$ and is independent of the labels $Y_j = R_{I_j}, j \in \llbracket n \rrbracket$, observed in S_n , and of I . This is because under this condition, the underlying point t_I has not been observed in the training sample S_n and therefore the latent random variable R_I remains random Bernoulli $1/2$ conditional to events A and $I \notin \mathcal{I}$, and independent of the observed labels and the index I of the test point. We give a formal justification for this claim later below, but it is intuitive.

From this we deduce that, conditional to events $A, I \notin \mathcal{I}$, Y is Bernoulli $1/2$ and independent of S_n and I (and therefore of X since $X = t_I$), therefore $\mathbb{P} \left[\widehat{f}_{S_n}(X) \neq Y \mid I \notin \mathcal{I}, A \right] = 1/2$ by the grouping lemma (“coalition lemma”) for independent variables since $\widehat{f}_{S_n}(X)$ only depends on S_n and X .

As for the second factor in (1.11), since $|I| \leq n$ (there may be repeated values of the x_i s in the sample, hence the inequality), it holds

$$\mathbb{P}[I \notin \mathcal{I} | A] = 1 - \frac{|\mathcal{I}|}{m} \geq 1 - \frac{n}{m}.$$

Backtracking, we get $\mathbb{P}[\widehat{f}_{S_n}(X) \neq Y] \geq \frac{1}{2}(1 - \frac{n}{m})$, so this is also a lower bound on our initial supremum. Since we can choose m arbitrary large, the result follows.

Formal justification of the claim: “conditional to the event A and $I \notin \mathcal{I}$, label $Y = R_I$ is drawn at random with probability $1/2$ and is independent of the labels $Y_j = R_{I_j}, j \in \llbracket n \rrbracket$, observed in S_n , and of I .”

Formally, for any values $(z_k)_{k \in \mathcal{I}} \in \{0, 1\}^{|\mathcal{I}|}$ and y in $\{0, 1\}$, for any index $k \notin \mathcal{I}$:

$$\begin{aligned} \mathbb{P}[(Y_j = z_{i_j})_{j \in \llbracket n \rrbracket}; I = k; Y = y | I \notin \mathcal{I}, A] &\stackrel{(a)}{=} \mathbb{P}[(R_{i_j} = z_{i_j})_{j \in \llbracket n \rrbracket}; I = k; R_k = y | I \notin \mathcal{I}, A] \\ &= \mathbb{P}[(R_\ell = z_\ell)_{\ell \in \mathcal{I}}; I = k; R_k = y | I \notin \mathcal{I}, A] \\ &\stackrel{(b)}{=} \mathbb{P}[(R_\ell = z_\ell)_{\ell \in \mathcal{I}}; I = k; R_k = y | I \notin \mathcal{I}] \\ &= \mathbb{P}[(R_\ell = z_\ell)_{\ell \in \mathcal{I}}; R_k = y | I = k] \mathbb{P}[I = k | I \notin \mathcal{I}] \\ &\stackrel{(c)}{=} \mathbb{P}[(R_\ell = z_\ell)_{\ell \in \mathcal{I}}, R_k = y] \mathbb{P}[I = k | I \notin \mathcal{I}] \\ &\stackrel{(d)}{=} \frac{1}{2} \mathbb{P}[(R_\ell = z_\ell)_{\ell \in \mathcal{I}}] \mathbb{P}[I = k | I \notin \mathcal{I}] \\ &\stackrel{(e)}{=} \frac{1}{2} \mathbb{P}[(R_\ell = z_\ell)_{\ell \in \mathcal{I}}; I = k | I \notin \mathcal{I}, A], \end{aligned}$$

where:

- for (a), we used that $Y_j = R_{i_j}$ a.s.;
- for (b), we used that the latent variables R_i for fixed indices, as well as the variable I , are independent of A ;
- for (c), we used that the latent variables R_i for fixed indices are independent of I ;
- for (d), we used that $k \notin \mathcal{I}$ and the variables $(R_j)_{i \in \mathcal{I}}$ are independent Bernouli(1/2);
- for (e), we implicitly backtracked all of the previous steps using the same type of argument.

This indeed establishes the claim. □

2 Linear Discrimination: A brief overview of classical methods

In this chapter, we will exclusively focus on the classification problem (also called *discrimination*, especially in older literature). The methods presented here are classical and not particularly recent (the idea of Linear Discriminant dates back to R.A. Fisher in 1936, Rosenblatt's Perceptron is from 1956), but they still form the backbone of most machine learning toolboxes nowadays and are to be known in order to understand more recent developments. Furthermore, we will not study the question of convergence or statistical consistency in this chapter, this will be relegated to later chapters.

We then consider $\mathcal{Y} = \{0, 1, \dots, K-1\}$, with $K \geq 2$, each element of \mathcal{Y} being called a *class*. There are K classes, the case $K = 2$ is called binary classification.

We will also always assume that the input space is $\mathcal{X} \subset \mathbb{R}^d$.

2.1 Linear discrimination functions

**

Definition 2.1. The family of affine score functions $(s_y(\cdot))_{y \in \mathcal{Y}}$ based on vectors $\mathbf{w} = (w_y)_{y \in \mathcal{Y}} \in \mathbb{R}^{dK}$ and constants $\mathbf{b} = (b_y)_{y \in \mathcal{Y}} \in \mathbb{R}^K$ is given by

$$s_y(x) = \langle w_y, x \rangle + b_y, \quad y \in \mathcal{Y}.$$

An associated classification function (linear discriminant) is given by

$$f_{\mathbf{w}, \mathbf{b}}(x) = \underset{y \in \mathcal{Y}}{\text{Arg Max}} s_y(x). \quad (2.1)$$

Note: Strictly speaking, in (2.1) we should break ties in some way in the case the Arg Max contains more than one element, i.e. two or more scores are equal. In order to avoid uninteresting complications, we will always assume that in such cases the ties are broken in favor of the smallest class, i.e. the Arg Max is always replaced by its smallest element; note that is always non-empty since the number of classes is finite.

If we denote $\tilde{x} := (x, 1) \in \mathbb{R}^{d+1}$, it holds $\langle w_y, x \rangle + b_y = \langle \tilde{w}_y, \tilde{y} \rangle$ wherein $\tilde{w}_y := (w_y, b_y)$. Thus, if we consider the “augmented” input space \mathbb{R}^{d+1} and the score functions are linear functions of \tilde{x} . For this reason we will occasionally (depending on the context) drop the constants \mathbf{b} , and implicitly assume we have performed this augmentation operation. Also for this reason, we will with some abuse of language talk about “linear” score functions although they are strictly speaking affine.

In the binary classification case ($K = 2$), observe that we have

$$f_{\mathbf{w}, \mathbf{b}}(x) = \mathbf{1} \left\{ \underbrace{\langle w_1 - w_0, x \rangle}_w + \underbrace{(b_1 - b_0)}_b \geq 0 \right\}, \quad (2.2)$$

(we assume ties are broken in favor of class 1 to simplify). Therefore, in the case of binary classification, we will consider only one score $s(x) = \langle w, x \rangle + b$ to simplify. (A similar reduction in parameters can be achieved for $K > 1$, by choosing the class 0 as reference and defining modified scores $s'_y(x) = s_y(x) - s_0(x)$, $y \in \mathcal{Y}$. But this breaks the symmetry somewhat, so that generally the full parametrization is used).

The goal of this chapter is to give an overview of classical methods to construct a linear discrimination function from a training sample $S_n = ((x_i, y_i))_{1 \leq i \leq n}$. The most natural approach if the standard classification loss is the target is the associated **ERM**:

$$(\hat{w}, \hat{b}) = \underset{(w, b) \in \mathbb{R}^{K(d+1)}}{\text{Arg Min}} \hat{\mathcal{E}}(f_{w, b}) = \underset{(w, b) \in \mathbb{R}^{K(d+1)}}{\text{Arg Min}} \frac{1}{n} \sum_{i=1}^n \mathbf{1} \left\{ \underset{y \in \mathcal{Y}}{\text{Arg Max}} (\langle w_y, x_i \rangle + b_y) \neq y_i \right\}.$$

Unfortunately (even in the binary classification case), the above minimization problem is considered at best cumbersome and at worst quite intractable, in particular in high dimension d , with large training data size n , etc. This is due essentially to the fact that the empirical risk is a noncontinuous, piecewise constant function of the parameters, so that usual numerical optimization approaches such as (stochastic) gradient descent are not applicable.

It is fair to say that ERM in the above form for classification is almost never used in practice. Instead, several alternative approaches are used, falling roughly speaking in two categories:

1. A particular class of input space \mathcal{X} and/or generating distribution P_{XY} is assumed, allowing to write the theoretically optimal linear classifier $f^* = f_{w^*, b^*}$ in an explicit way and using this knowledge to estimate the parameters directly from the data. This is sometimes called a *generative approach*.
2. One considers directly the linear score functions (with output in $\tilde{\mathcal{Y}} = \mathbb{R}^{K+1}$), and uses a different loss ℓ on $\tilde{\mathcal{Y}} \times \mathcal{Y}$ which lends itself better to optimization (typically because it is convex in the first variable). This approach is called the use of a “proxy loss”.

** 2.2 The naive Bayes classifier

The “naive Bayes” method falls into the category of generative approaches. The setting assumes that the input space is $\mathcal{X} = \{0, 1\}^d$, i.e. the coordinates (also called “features”) of the predictor x are binary. Furthermore, the main assumption is the following:

Assumption (NB): Under the generative distribution P_{XY} , the coordinates of X are independent conditionally to Y .

We then have the following result:

Proposition 2.2. *Naive Bayes, binary classification case* Assume $\mathcal{X} = \{0, 1\}^d$, $K = 2$, and assumption **(NB)** is satisfied. Then the optimal (Bayes) classifier function takes the form

$$f^*(x) = \mathbf{1} \left\{ \sum_{k=1}^d w^{(k)} x^{(k)} + b \geq 0 \right\},$$

where, using the notation $p_{k,j} := \mathbb{P}[X^{(k)} = 1|Y = j]$ and $\pi_j := \mathbb{P}[Y = j]$ (which are assumed to belong to $(0,1)$):

$$w^{(k)} := \log \frac{p_{k,1}(1 - p_{k,0})}{(1 - p_{k,1})p_{k,0}}; \quad b := \sum_{k=1}^d \log \frac{1 - p_{k,1}}{1 - p_{k,0}} + \log \frac{\pi_1}{\pi_0}.$$

Proof. From the general formula of the Bayes classifier we have

$$f^*(x) = \mathbf{1} \{ \mathbb{P}[Y = 1|X = x] \geq \mathbb{P}[Y = 0|X = x] \} = \mathbf{1} \left\{ \log \frac{\mathbb{P}[Y = 1|X = x]}{\mathbb{P}[Y = 0|X = x]} \geq 0 \right\}. \quad (2.3)$$

We then notice:

$$\mathbb{P}[Y = i|X = x] = \mathbb{P}[X = x|Y = i] \frac{\mathbb{P}[Y = i]}{\mathbb{P}[X = x]}, \quad (2.4)$$

so that

$$\log \frac{\mathbb{P}[Y = 1|X = x]}{\mathbb{P}[Y = 0|X = x]} = \log \frac{\mathbb{P}[X = x|Y = 1]}{\mathbb{P}[X = x|Y = 0]} + \frac{\mathbb{P}[Y = 1]}{\mathbb{P}[Y = 0]}.$$

The second term above is $\log \frac{\pi_1}{\pi_0}$. As to the first term, using assumption **(NB)**, and the fact that all coordinates belong to $\{0, 1\}$:

$$\begin{aligned} \log \mathbb{P}[X = x|Y = i] &= \log \prod_{k=1}^d \mathbb{P}[X^{(k)} = x^{(k)}|Y = i] = \sum_{k=1}^d (x^{(k)} \log p_{k,i} + (1 - x^{(k)}) \log(1 - p_{k,i})) \\ &= \sum_{k=1}^d \left(x^{(k)} \log \frac{p_{k,i}}{1 - p_{k,i}} + \log(1 - p_{k,i}) \right). \end{aligned}$$

Replacing into (2.4), (2.3) we obtain the announced result. \square

Estimation from data: in order to estimate a Naive Bayes classifier in practice, the plug-in principle is used, that is, the theoretical parameters $p_{k,i}$ and π_i are replaced in the formula by their frequentist estimators from a sample $S_n = ((x_i, y_i)_{1 \leq i \leq n})$:

$$\hat{\pi}_i := \frac{|\{j : y_j = i\}|}{n}; \quad \hat{p}_{k,i} := \frac{|\{j : (x_j^{(k)}, y_j) = (1, i)\}|}{|\{j : y_j = i\}|},$$

assuming the last denominator is nonzero (i.e. there exists at least one training example in each class.)

Exercise 2.1. Generalize the above result of Proposition 2.2 to the case $K > 2$.

2.3 Gaussian generative distribution: LDA and QDA

We consider in this section a generative approach where all the class-conditional distributions are Gaussian, i.e.

Assumption (GD):

$$\mathbb{P}_{X|Y}[\bullet|Y = i] = \mathcal{N}(m_i, \Sigma_i), \quad i = 0, \dots, K - 1. \quad (2.5)$$

We will also focus on the particular case of equal covariance matrices:

Assumption (GDEC): Like (GD), but where $\Sigma_i = \Sigma$ for all i .

We will assume furthermore that the Gaussian distributions involved are non-degenerate, i.e. the covariance matrices Σ_i have full rank. In this case the distributions have respective densities

$$f_i(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2}(x - m_i)^T \Sigma_i^{-1} (x - m_i)\right), \quad i = 0, \dots, K - 1.$$

with respect to the Lebesgue measure on \mathbb{R}^d and from Exercise 1.2 we know that $f^*(x) = \text{Arg Max}_i (\pi_i f_i(x))$ is a Bayes classifier, where $\pi_i := \mathbb{P}[Y = i]$. Since the arg max is unchanged by monotone increasing transformation of the function to minimize, we can take the logarithm and obtain

Under (GD), the Bayes classifier is given by

$$f_{\text{QDA}}^*(x) = \text{Arg Max}_i s_i(x), \quad \text{where } s_i(x) := -\frac{1}{2}(x - m_i)^T \Sigma_i^{-1} (x - m_i) - \frac{1}{2} \log |\Sigma_i| + \log \pi_i. \quad (2.6)$$

Since the above score functions are quadratic polynomials in x , this is called *Quadratic Discriminant Analysis* (QDA). The main quadratic term $(x - m_i)^T \Sigma_i^{-1} (x - m_i)$ is called (squared) *Mahalanobis distance* of x to the center m_i of the Gaussian distribution $\mathcal{N}(m_i, \Sigma_i)$.

Under **(GDEC)**, we observe that the quadratic part of the above scores is identical for all scores. We can therefore subtract it from all scores without changing the arg max as above, and it comes

Under **(GDEC)**, the Bayes classifier is given by

$$f_{\text{LDA}}^*(x) = \underset{i}{\text{Arg Max}} s_i(x), \text{ where } s_i(x) := \underbrace{\langle x, \Sigma^{-1} m_i \rangle}_{w_i} + \underbrace{\log \pi_i - \frac{1}{2} m_i^t \Sigma^{-1} m_i}_{b_i}, \quad (2.7)$$

which is a linear discrimination function called *Linear Discriminant Analysis* (LDA).

In the two-class case, as argued before it is enough to consider the difference of scores

Under **(GDEC)**, in the binary classification case, the Bayes classifier is given by

$$f_{\text{LDA}}^* = \mathbf{1} \left\{ \langle x, \underbrace{\Sigma^{-1}(m_1 - m_0)}_{w_{\text{LDA}}} \rangle + (b_1 - b_0) \geq 0 \right\},$$

where b_0, b_1 are as in (2.7).

Estimation from data: similarly to the previous section, one uses a plug-in approach wherein the unknown population parameters Σ_i, m_i, π_i are estimated by their frequentist estimators counterparts from a sample S_n and just “plugged into” the formula (2.6) resp. (2.7) (denoting $n_i := |\{j : y_j = i\}|$). Thus, for **(GD)**:

$$\begin{aligned} \hat{\pi}_i &:= \frac{|\{j : y_j = i\}|}{n}; \\ \hat{m}_i &:= \frac{1}{n_i} \sum_{j \text{ s.t. } y_j = i} x_j; \\ \hat{\Sigma}_i &:= \frac{1}{n_i - 1} \sum_{j \text{ s.t. } y_j = i} (x_i - \hat{m}_j)^T (x_i - \hat{m}_j). \end{aligned} \quad (2.8)$$

Note that the $(n_i - 1)$ in the denominator of $\hat{\Sigma}$ is to make the estimator unbiased, assuming $n_i \geq 2$; see a classical statistics course. If n_i is used instead, it will not change much.

On the other hand, for **(GDEC)**, one uses the so-called *pooled estimator* for the common covariance matrix:

$$\hat{\Sigma} := \frac{1}{n - K} \sum_{i=1}^K (x_i - m_{y_i})^T (x_i - m_{y_i}). \quad (2.9)$$

Exercise 2.2. Prove that QDA and LDA using the above estimators from data are *covariant by (bijective) linear data transformation*. More precisely, let $\tilde{x}_i := Ax_i$, where A is an invertible linear operator of \mathbb{R}^d . Denote \hat{f} the classification function constructed by LDA or QDA from $S_n := ((x_i, y_i)_{1 \leq i \leq n})$, and \tilde{f} the one constructed from $\tilde{S}_n := ((\tilde{x}_i, y_i)_{1 \leq i \leq n})$. Then $\hat{f}(x) = \tilde{f}(Ax)$ for all $x \in \mathbb{R}^d$.

Practical tricks.

In practice, there are often some modifications to the above canvas for LDA and QDA:

1. While the ERM estimator for the 0-1 loss is infeasible as argued in the beginning of this chapter, in the binary classification case, it is fairly easy to minimize, for a fixed w , the empirical classification error as a function of the constant b , $b \mapsto \hat{\mathcal{E}}(f_{w,b})$. Namely, for fixed b the problem is reduced to a 1-dimensional one and one only has to try for b the intermediate values of the reordered set of $\{\langle x_j, w \rangle, j = 1, \dots, n\}$ and select the one minimizing the empirical classification error. This is often what is done in practice, so that the formula for the linear projection $w_{\text{LDA}} = \Sigma^{-1}(m_1 - m_0)$ is often considered as the most important of (binary) LDA, while the exact formula for the constant b is unimportant, because it is often replaced by the above ERM minimizer.
2. The most problematic part of LDA (and a fortiori QDA) in practice is the inversion of the estimated covariance matrix. If some of its estimated eigenvalues are close to 0, taking the inverse can be a highly unstable operation and lead to significant estimation errors and erratic behavior, especially if the dimension d is large. For this reason instead of $\hat{\Sigma}$ as defined in (2.8),(2.9), it is often suggested to use a “regularized” version of it, such as

$$\tilde{\Sigma} := (1 - \lambda)\hat{\Sigma} + \lambda\hat{\sigma}^2 I_d,$$

or

$$\tilde{\Sigma} := (1 - \lambda)\hat{\Sigma} + \lambda\hat{D},$$

where D is the diagonal matrix formed with the diagonal entries $\hat{\sigma}_i^2 := \hat{\Sigma}_{ii}$ of $\hat{\Sigma}$ (estimators of the variances of each coordinate), and $\hat{\sigma}^2 = d^{-1} \sum_{i=1}^d \hat{\sigma}_i^2$. Here $\lambda \in [0, 1]$ is a so-called “shrinking parameter” that has to be tuned, for instance by cross-validation (see next chapter).

* **2.4 Classification as regression**

In this approach we use a proxy loss function, namely the quadratic loss. More precisely, we follow the principle explained in Example 1.4 (quadratic loss for multiclass classification). A linear prediction function f_w depending on parameters $w = (w_0, \dots, w_{K-1})$ outputs the scores in \mathbb{R}^K given by

$$f_w(x) = (\langle x, w_0 \rangle, \dots, \langle x, w_{K-1} \rangle)$$

(as explained earlier in this chapter, we disregard the constant parameters by implicitly “augmenting” the data x to $(x, 1)$). Furthermore we consider the quadratic loss $\ell(f_{\mathbf{w}}(x), y) = \|f_{\mathbf{w}}(x) - e_y\|^2$, where e_k denotes the k -th canonical basis vector in \mathbb{R}^K (recall $(k \mapsto e_k)$ is called “one-hot encoding” in K -class classification). The ERM is then given by

$$\begin{aligned}\hat{\mathbf{w}} &= \text{Arg Min}_{\mathbf{w} \in \mathbb{R}^{dK}} \sum_{j=1}^n \|f_{\mathbf{w}}(x_j) - e_{y_j}\|^2 \\ &= \text{Arg Min}_{\mathbf{w} \in \mathbb{R}^{dK}} \sum_{k=0}^{K-1} \sum_{j=1}^n (\langle w_k, x_j \rangle - \mathbf{1}\{y_j = k\})^2.\end{aligned}$$

Since the above function to minimize separates into K sums each involving only w_k , each sum can be minimized independently, given rise to

$$\hat{w}_k = \text{Arg Min}_{w \in \mathbb{R}^d} \sum_{j=1}^n (\langle x_j, w \rangle - \mathbf{1}\{y_j = k\})^2 = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \mathbf{Y}^{(k)},$$

using the notation introduced below (1.4), and with $\mathbf{Y}^{(k)} := (\mathbf{1}\{y_1 = k\}, \dots, \mathbf{1}\{y_n = k\}) \in \mathbb{R}^n$.

Thus in this approach we transform the initial multiclass classification problem into K distinct “one-versus-all” binary classification problems (classify class k against all the others), each using a different linear predictor (and we recall that the final class prediction is the the class attaining the maximum of these K linear scores).

Discussion: The unrestricted optimal prediction function for the loss function $\ell(y', y) = \|y' - e_y\|^2$ is given (again by separation into independent sums) by

$$f^*(x) = (\mathbb{P}[Y = 0|X = x], \dots, \mathbb{P}[Y = K - 1|X = x]) \in [0, 1]^K,$$

and it can seem a rather questionable idea to try to approximate a function from \mathbb{R}^d to $[0, 1]^K$ by a linear function, which is by definition unbounded. This is actually the main reason why this quadratic loss approach is actually almost never used in practice with linear prediction. Much more popular is the logistic regression.

Exercise 2.3. In this exercise we use explicitly the affine representation with parameters (w, b) for affine scores. We focus on the binary classification case ($K = 2$).

Establish that the above approaches reduces to a single regression problem, since it holds $(\hat{w}_1, \hat{b}_1) = (-\hat{w}_0, 1 - \hat{b}_0)$.

Prove that the direction of the estimated vector \hat{w}_1 using the quadratic regression approach coincides with the direction found by the (binary) LDA approach. (The constants found by the two approaches b differ, and the above property does not hold any more for $K \geq 3$).

**

2.5 Linear logistic regression

As we have seen, usual quadratic regression models the class probability functions $\eta_k(x) := \mathbb{P}[Y = k|X = x]$ as linear functions, which is problematic. The idea of *logistic regression* is to use a suitable transform, the “logit” transform, which is a log-ratio of probabilities. Here we will use the class 0 as a reference and implicitly assume that $\eta_k(x) \neq 0$ for all $k = 0, \dots, K - 1$.

Define the ideal logistic score functions as

$$s_k(x) := \log \frac{\mathbb{P}[Y = k|X = x]}{\mathbb{P}[Y = 0|X = x]}, \quad k = 0, \dots, K - 1. \quad (2.10)$$

(note that of course $s_0(x)$ is identically 0),

and observe that $f^*(x) := \text{Arg Max}_{k=0, \dots, K-1} s_k(x)$ is an optimal classifier. The score functions s_k can range anywhere in \mathbb{R} .

The principle of logistic regression is to model these as linear functions of x :

$$s_k(x) = \langle w_k, x \rangle, \quad k = 0, \dots, K - 1, \quad (2.11)$$

with $w_0 := 0$ and $(w_i)_{1 \leq i \leq K-1}$ arbitrary.

Proposition 2.3. *If the logistic scores defined in (2.10) satisfy (2.11), then it holds conversely*

$$\eta_k(x) = \mathbb{P}[Y = k|X = x] = \frac{\exp\langle w_k, x \rangle}{\sum_{\ell=0}^{K-1} \exp\langle w_\ell, x \rangle}. \quad (2.12)$$

Proof: left to the reader.

Observe that (2.12) constitutes a partial *statistical model*: it specifies the form of the conditional distribution $\mathbb{P}_{\mathbf{w}}(Y|X)$ of Y given X , depending on a finite number of parameters $\mathbf{w} = (w_1, \dots, w_{K-1}) \in \mathbb{R}^{d(K-1)}$. However, the distribution of X itself is not modeled and can be arbitrary. This can be seen as a generative approach for conditional probabilities only. Following the principle of the *maximum likelihood* estimation in classical statistics, in order to estimate the parameters from a sample $S_n = ((x_i, y_i)_{1 \leq i \leq n})$, it is proposed to use a *maximum conditional likelihood* principle, i.e. find

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\text{Arg Max}} L(\mathbf{w}),$$

where $L(\cdot)$ is the log-conditional-likelihood based on the sample S_n :

$$\begin{aligned}
L(\mathbf{w}) &:= \log \mathbb{P}_{\mathbf{w}}^{\otimes n}[Y_1 = y_1, \dots, Y_n = y_n | X_1 = x_1, \dots, X_n = x_n] \\
&= \sum_{j=1}^n \log \mathbb{P}_{\mathbf{w}}[Y_j = y_j | X_j = x_j] \\
&= \sum_{j=1}^n \left(\langle w_{y_j}, x_j \rangle - \log \left(1 + \sum_{\ell=1}^{K-1} \exp \langle w_{\ell}, x_j \rangle \right) \right). \tag{2.13}
\end{aligned}$$

Unlike quadratic regression, the expression (2.13) does not separate into independent sums for each parameter: the optimization problem has to be solved for all parameters (vectors w_k) jointly.

In the particular case of binary classification, this reduces to

$$L(w) = \sum_{j=1}^n (y_j \langle w, x_j \rangle - \log(1 + \exp \langle w, x_j \rangle)). \tag{2.14}$$

Under the above form (multi-class or binary class), one can notice that $-L(w)$ can be interpreted as an empirical risk with loss function (written in the binary case for simplicity)

**

$\ell_{\text{logit}}(f(x), y) := \log(1 + \exp f(x)) - yf(x), \quad y \in \{0, 1\} \tag{2.15}$
--

which is then applied to linear prediction functions. In this sense, the logistic regression can also be interpreted as a *proxy loss* method.

Practical implementation: The maximum of $L(\mathbf{w})$, or equivalently the minimum of $-L(\mathbf{w})$, does not have a closed-form formula. Still, $\mathbf{w} \mapsto -L(\mathbf{w})$ is a convex function, so various methods of convex optimization can be used. If the dimension d is not too large, a common approach is to use Newton-Raphson iterations

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \underbrace{\left(\frac{d^2 L}{d\mathbf{w}d\mathbf{w}^t} \right)_{|\mathbf{w}_k}}_{\text{Hessian}}^{-1} \underbrace{\left(\frac{dL}{d\mathbf{w}} \right)_{|\mathbf{w}_k}}_{\text{Gradient}}.$$

After some calculations, these iterations can be put under a form where they are interpreted as a repeated “weighted least squares”, where the weights are updated along the iterations.

Exercise 2.4. For binary classification, it is common to encode the two classes as $\mathcal{Y} = \{-1, 1\}$ which is more symmetrical. With this convention, verify that the logit loss (2.15) can be rewritten as

$$\ell_{\text{logit}}(f(x), y') = \log(1 + \exp -(y' f(x))), \quad y' \in \{-1, 1\}.$$

2.6 Hinge-loss based methods: Perceptron and Support Vector Machine

We consider the binary classification setting with $\mathcal{Y} = \{-1, 1\}$, $\tilde{Y} = \mathbb{R}$ and the proxy loss function called “hinge loss”

$$\ell_\varepsilon(y', y) = (\varepsilon - yy')_+,$$

where $(t)_+ := \max(0, t)$ is the positive part, and $\varepsilon \geq 0$ a fixed parameter.

Interpretation as “large margin prediction”. In the case of linear binary classification, a decision function $f_{w,b}$ of the form (2.2) separates the space into two half-spaces separated by a $(d-1)$ -dimensional hyperplane with normal vector w and offset b from the origin. If we assume $\|w\| = 1$, the corresponding score function $s(x) = \langle w, x \rangle + b$ can be geometrically interpreted as the (signed) distance of point x to the separating hyperplane.

Because of this, the quantity $s(x)y$ is often called the “margin” of prediction: a positive margin indicates a correct classification and negative margin, a classification error; and the absolute value of the margin indicates the distance to the separating hyperplane. As a general rule, correct predictions that have margin too close to 0 are penalized, and incorrect predictions are penalized all the more if the distance to the separating hyperplane is high. Loss functions of the form $\ell(y, y') = L(yy')$ for various functions $L : \mathbb{R} \rightarrow \mathbb{R}_+$, are often called “margin-based loss functions”, and decision functions trained by minimizing the associated risk “maximum margin classifiers”.

Perceptron. The perceptron algorithm (Rosenblatt, 1956) is based on a stochastic gradient descent for the empirical risk using linear functions. More precisely, using the above loss the empirical risk based on sample S_n can be written

$$\widehat{\mathcal{E}}(f) = \frac{1}{n} \sum_{\substack{j \text{ s.t.} \\ f(x_j)y_j \leq \varepsilon}} (\varepsilon - f(x_j)y_j),$$

and for linear functions, the associated gradient (disregarding non-differentiability in 0 of $t \rightarrow (t)_+$):

$$\partial_w \widehat{\mathcal{E}}(f_w) = \frac{1}{n} \sum_{\substack{j \text{ s.t.} \\ \langle w, x_j \rangle y_j \leq \varepsilon}} (-y_j x_j).$$

The perceptron algorithm is an early use of stochastic gradient descent: to avoid recomputing the above gradient at each optimization step, it is proposed to select randomly one of the terms in the above sum and to make a step in that direction. This gives rise to the following very simple procedure:

**

Perceptron algorithm

1. Initialize $w = 0$.
2. Choose uniformly at random $j \in \{1, \dots, n\}$.
3. If $\langle w, x_j \rangle y_j > \varepsilon$, return to step 2.
4. Update $w \leftarrow w + y_j x_j$.
5. Go to step 2.

Nowadays, the Perceptron algorithm is not so widely used, more modern approaches such as the linear Support Vector Machine (see below) are used. One advantage remains its simplicity.

An famous early result studying the convergence of the empirical risk of the above algorithm in the simplest case $\varepsilon = 0$ and when the training data is linearly separable is the following.

**

Theorem 2.4 (Novikov, 1962). *Let $S_n = ((x_i, y_i)_{1 \leq i \leq n})$ be a fixed training sample for binary classification. Assume that*

- *The training sample is linearly separable with margin $\gamma > 0$, meaning:*

$$\exists w_* \in \mathbb{R}^d, \text{ s. t. } \|w_*\| = 1, \text{ and } \gamma > 0 : \forall i \in \{1, \dots, n\} \langle w_*, x_i \rangle y_i \geq \gamma.$$

- *For all $i = 1, \dots, n$ it holds $\|x_i\| \leq R$.*

Then: the Perceptron algorithm (run with $\varepsilon = 0$) finds a vector w separating perfectly the data (i.e. such that $\langle w, x_i \rangle y_i > 0$ for all i) after at most $(R/\gamma)^2$ effective update operations (i.e. passes through step 4).

*

Support Vector Machine. The (linear) Support Vector Machine (Boser, Guyon, Vapnik 1992) uses the hinge loss with $\varepsilon = 1$. Several efficient algorithms have been developed to optimize the resulting optimization problem, that are preferred to the perceptron, and we won't enter in detail here. It is nowadays a standard method of machine learning toolboxes.

Exercise 2.5. Justify that in the binary classification case, the “classification as regression” approach (Section 2.4) and the logistic regression approach (Section 2.5) both can be seen

as “large margin classification” methods, in the sense that they are based on minimizing a certain *margin-based loss* (which can be written explicitly).

Exercise 2.6. Prove Thm. 2.4. Proceed as follows: let $\eta > 0$ be a fixed positive number, and consider the quantity $\Delta_k^2 := \|w_k - \eta w_*\|^2$, where w_k denotes the vector w after the k th effective update step. Compare Δ_{k+1}^2 to Δ_k^2 , and choose an appropriate value of η to establish that it must hold $\Delta_k^2 \leq \Delta_0^2 - kR^2$, then conclude. **Note:** this proof is an elementary example of the use of a well-chosen *Lyapunov function* decreasing along the iterations (here $\|w_k - \eta w_*\|^2$). It is a powerful technique to study convergence of iterative optimization schemes.

* 2.7 Regularization

In most linear methods, standard approaches tend to become unstable when the dimension d is high. This has, roughly speaking, to do with the overfitting phenomenon: intuitively, the number of free parameters is d and overfitting becomes more likely when there are more parameters to fit relative to the amount of data available. A common approach to “stabilize” ERM methods is to consider a *regularized* version

$$\hat{w}_\lambda \in \underset{w \in \mathbb{R}^d}{\text{Arg Min}} \left(\hat{\mathcal{E}}_\ell(f_w) + \lambda \Omega(w) \right),$$

where $\Omega : \mathbb{R}^d \rightarrow \mathbb{R}$ is a *regularization* or *penalization* function which is generally assumed to be convex, so that the above problem lends itself well to numerical optimization. A very common choice is $\Omega(w) = \|w\|^2$, though many others can be found in the literature.

The methods considered in previous sections using ERM with the squared loss (Section 2.4, the logit loss (Section 2.5), and the hinge loss (Section 2.6) all admit regularized versions in this sense, which are generally the ones used in practice.

Let us study the particular case of linear regression in the binary classification case ($\mathcal{Y} = \{0, 1\}$), which reduces to usual linear regression (see (1.3)). The corresponding regularized ERM is

$$\hat{w}_\lambda = \underset{w \in \mathbb{R}^d}{\text{Arg Min}} \left(\sum_{i=1}^n (y_i - \langle w, x_i \rangle)^2 + \lambda \|w\|^2 \right),$$

and it can be checked that it takes the explicit form (compare to (1.4)):

$$\hat{w}_\lambda = \left(\hat{\Sigma} + \lambda I_d \right)^{-1} \hat{\gamma}, \quad \text{where } \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^n x_i x_i^T, \quad \text{and } \hat{\gamma} := \frac{1}{n} \sum_{i=1}^n x_i y_i; \quad (2.16)$$

a classical equivalent form is $\hat{w}_\lambda = (\mathbf{X}^t \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^t \mathbf{Y}$, where \mathbf{X} is the (n, d) matrix whose rows are x_1^t, \dots, x_n^t , and $\mathbf{Y} = (y_1, \dots, y_n)^t$. This is also often called *ridge regression*.

The effect of regularization is therefore intuitively clear in the above formula: adding $\lambda > 0$ on the diagonal of a matrix before inverting it should stabilize the inverse operation, at the expense of adding some bias if λ is too large. A similar idea was used for regularized LDA/QDA (Section 2.3).

In practice, the parameter $\lambda > 0$ has to be tuned in order to get a good compromise between too much and too little regularization; we will see how in the coming chapter.

Exercise 2.7. Justify the formula (2.16) for regularized linear regression.

3 Introduction to statistical learning theory (part 2): elementary bounds

3.1 Controlling the error of a single decision function and Hold-Out principle

The main goal in this section is to get a theoretically justified control of the true risk $\mathcal{E}(\hat{f})$ of an estimator by using only observable quantities (i.e. that can be computed from the available data). In statistical terms, this means we are looking for a *confidence (upper) bound* on $\mathcal{E}(\hat{f})$ (we are primarily interested in an upper bound, i.e. a guarantee on the generalization error). A confidence bound is a quantity that can be computed from the data, and is indeed larger than the (unknown) quantity of interest with a prescribed probability (called *coverage probability*).

As we have discussed in Section 1.3, the empirical risk $\hat{\mathcal{E}}(\hat{f})$ is not (in general) a reliable approximation of $\mathcal{E}(\hat{f})$ because of the overfitting phenomenon: the same data is used to learn \hat{f} and to evaluate the empirical risk, resulting in a bias (that can possibly be very large).

On the other hand, we have argued that for a *fixed* decision function, it holds $\mathbb{E}[\hat{\mathcal{E}}(f)] = \mathcal{E}(f)$, moreover we have by the law of large numbers $\hat{\mathcal{E}}(f) = \frac{1}{n} \sum_{i=1}^n Z_i \rightarrow \mathcal{E}(f)$ in probability (or a.s.) as $n \rightarrow \infty$, where $Z_i := \ell(f(X_i), Y_i)$ are i.i.d.

This idea underlies the co-called “Hold-out” principle: learn \hat{f} and evaluate the empirical error on *different* samples. Thus, if S_n and $T_m = (X'_i, Y'_i)_{1 \leq i \leq m}$ are independent i.i.d. samples, we consider the quantity

$$\hat{\mathcal{E}}^{HO}(\hat{f}) = \hat{\mathcal{E}}(\hat{f}_{S_n}, T_m) = \frac{1}{m} \sum_{i=1}^m \ell(\hat{f}_{S_n}(X'_i), Y'_i).$$

In practice, S_n and T_m can be obtained from an arbitrary separation in two parts of a single sample. S_n is called *learning sample* and T_m *validation sample* (or “hold-out” sample, since it has been held out from the learning phase). Observe that conditionally to S_n , we can consider \hat{f}_{S_n} as a fixed function, thus $\mathbb{E}[\hat{\mathcal{E}}(\hat{f}_{S_n}, T_m) | S_n] = \mathcal{E}(\hat{f}_{S_n})$, furthermore we can apply the law of large numbers (LLN) conditionally to S_n .

This principle justifies that it is already of interest to get a mathematical control of the error of a single fixed decision function f . For this, we need more elaborate tools than the LLN, which gives an limiting value but not quantification of the speed of convergence. A traditional (asymptotic) tool for this is the Central Limit Theorem (CLT), which we recall here: if W_1, \dots, W_n are i.i.d. real random variables such that $\sigma^2 = \text{Var}[W_1]$ exists, then it holds

$$\sqrt{n} \frac{\frac{1}{n} \sum_{i=1}^n (W_i - \mathbb{E}[W_1])}{\sigma} \rightarrow \mathcal{N}(0, 1), \text{ in distribution, as } n \rightarrow \infty. \quad (3.1)$$

This can be used for deriving *asymptotic* confidence intervals (in the statistical learning context, remember the quantity of interest is typically $\mathcal{E}(f)$). An asymptotic confidence

interval or bound has the property that the coverage inequality only converges asymptotically to a prescribed value, say $1 - \alpha$.

However, learning theory generally focuses on obtaining *nonasymptotic* bounds, that is, whose validity (coverage probability) is ensured for any n . There are several motivations for this, but we will in particular stress that the CLT (3.1) is not valid in a uniform sense in general. Consider a situation where $X_i(p)$ are i.i.d. Bernoulli with parameter p (for instance, $W_i(p) = \mathbf{1}\{f(X_i) \neq Y_i\}$ for a classification problem and a given classifier function f with generalization error p). If the limit in (3.1) was uniform with respect to the parameter p , we would choose an arbitrary sequence p_n depending on n and still have the convergence (3.1). Yet, it is known that for $p_n = c/n$, we have the Poisson limit

$$\sum_{i=1}^n W_i\left(\frac{c}{n}\right) \longrightarrow \text{Poisson}(c), \text{ in distribution, as } n \longrightarrow \infty. \quad (3.2)$$

Therefore, we deduce (by usual continuity arguments for the convergence in distribution, and recalling the variance of Bernoulli variable with parameter p is $p(1 - p)$) that

$$\frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n (W_i(\frac{c}{n}) - \mathbb{E}[W_1(\frac{c}{n})])}{\sigma} \longrightarrow \frac{\text{Poisson}(c) - c}{\sqrt{c}}, \text{ in distribution, as } n \longrightarrow \infty,$$

which is obviously different from the standard CLT Gaussian limit. This situation could in principle happen if we are given a sequence of classifiers f_n whose classification error converges to zero fast enough as $n \rightarrow \infty$.

Exercise 3.1. Prove the convergence (3.2) by considering the Laplace transform $F_S(\lambda) := \mathbb{E}[\exp(\lambda S)]$ on both sides. It is known that the pointwise convergence of the Laplace transform implies the convergence in distribution.

3.2 A sharp but inconvenient bound in the Bernoulli case: the Clopper-Pearson bound

We consider in this section an exact confidence bound for the parameter of a Bernoulli variable, which we recall is relevant for the classification problem in order to get a bound on the generalization error $\mathcal{E}(f) = p \in [0, 1]$ of a fixed classifier f . In this situation, $W_i = \mathbf{1}\{f(X_i) \neq Y_i\}$ is a Bernoulli variable with parameter p , and the number of errors on the training sample is $n\hat{\mathcal{E}}(f) = \sum_{i=1}^n W_i$ which has a Binom(n, p) distribution.

* **Proposition 3.1.** *Let $F(p, n, k) = \mathbb{P}[B_{n,p} \leq k] = \sum_{i \leq k} \binom{n}{i} p^i (n - p)^{n-i}$ be the cumulative distribution function (cdf) of a Binom(n, p) variable $B_{n,p}$. Let*

$$\mathcal{B}(u, n, \alpha) = \max\{p \in [0, 1] : F(p, n, u) \geq \alpha\}.$$

Then given $B_{n,p}$, the quantity $\mathcal{B}(B_{n,p}, n, \alpha)$ is an upper confidence bound on p at coverage level $1 - \alpha$, which is to say

$$\forall p \in [0, 1] : \mathbb{P}[p \leq \mathcal{B}(B_{n,p}, n, \alpha)] \geq 1 - \alpha.$$

To prove this, we'll use the following classical result:

** **Lemma 3.2.** *Let X be a real-valued variable, and $F_X(t) := \mathbb{P}[X \leq t]$ its cdf. Then*

$$\forall \alpha \in [0, 1] : \mathbb{P}[F_X(X) \leq \alpha] \leq \alpha,$$

with equality if F_X is a continuous function. We say that $F_X(X)$ is stochastically lower bounded by a $\text{Unif}([0, 1])$ variable (and is exactly distributed as uniform in $[0, 1]$ if F_X is continuous).

(Proof of the Lemma). Let $s := \sup\{x : F_X(x) \leq \alpha\}$. We consider two cases:

(1) s is a maximum: then $F_X(s) = \alpha$ (since F_X is right-continuous), and $F_X(x) \leq \alpha \Leftrightarrow x \leq s$ by monotonicity of F_X . Then

$$\mathbb{P}[F_X(X) \leq \alpha] = \mathbb{P}[X \leq s] = F_X(s) = \alpha.$$

(2) s is not a maximum: then $F_X(x) \leq \alpha \Leftrightarrow x < s$, and

$$\mathbb{P}[F_X(X) \leq \alpha] = \mathbb{P}[X < s] = \lim_{x \nearrow s} F_X(x) \leq \alpha.$$

Note that case (2) can only happen if F_X is not (left)-continuous at point s , hence if F_X is continuous we are always in case (1). \square

(Proof of the proposition). Observe that $p > \mathcal{B}(u, n, \alpha) \Rightarrow F(p, n, u) < \alpha$ by definition of $\mathcal{B}(u, n, \alpha)$. Hence

$$\mathbb{P}[p > \mathcal{B}(B_{n,p}, n, \alpha)] \leq \mathbb{P}[F(p, n, B_{n,p}) < \alpha] \leq \alpha,$$

where the last inequality is from the Lemma. \square

The above confidence bound, called Clopper-Pearson bound, is sharp because it is based on the exact inversion of the cdf. However, it is not easy to qualitatively understand nor to manipulate. We will now consider a method to construct more explicit bounds.

3.3 The Chernov method

The Chernov method, also called the Laplace transform method or exponential moment method, is a generic device allowing to obtain non-asymptotic confidence bounds.

*** **Proposition 3.3.** *Let X be a real-valued random variable, and define successively for $\lambda \in \mathbb{R}$:*

$$F_X(\lambda) := \mathbb{E}[\exp(\lambda X)] \in (0, \infty] ; \quad (\text{Laplace transform})$$

$$\Psi_X(\lambda) := \log F_X(\lambda) \in (-\infty, \infty] ; \quad (\text{Log-Laplace transform})$$

$$\Psi_X^*(t) := \sup_{\lambda \in \mathbb{R}} (\lambda t - \Psi_X(\lambda)) \in [-\infty, \infty] ; \quad (\text{Legendre dual of } \Psi_X, \text{ Cramér transform}).$$

Then for any $t \geq \mathbb{E}[X]$:

$$\mathbb{P}[X \geq t] \leq \exp -\Psi_X^*(t). \quad (3.3)$$

Furthermore, if X_1, \dots, X_n are i.i.d. with the same distribution as X and $S_n := \sum_{i=1}^n X_i$, then for any $u \geq \mathbb{E}[X]$:

$$\mathbb{P}\left[\frac{1}{n}S_n \geq u\right] \leq \exp -n\Psi_X^*(u). \quad (3.4)$$

Note: the principle of the proof is so simple and useful that it is as important as the theorem itself.

Proof. Let us begin with assuming that $\lambda \geq 0$, so that $x \mapsto \exp(\lambda x)$ is a nondecreasing function. We have for any $t \in \mathbb{R}$:

$$\begin{aligned} \mathbb{P}[X \geq t] &\leq \mathbb{P}[\exp(\lambda(X - t)) \geq 1] \quad (\text{inequality only to account for the case } \lambda = 0) \\ &\leq \mathbb{E}[\exp(\lambda(X - t))] \quad \text{by Markov's inequality} \\ &= F_X(\lambda) \exp(-\lambda t) \\ &= \exp(\Psi_X(\lambda) - \lambda t). \end{aligned}$$

Optimizing over $\lambda \geq 0$, we thus get

$$\mathbb{P}[X \geq t] \leq \exp(-\sup_{\lambda \geq 0}(\lambda t - \Psi_X(\lambda))).$$

Now we justify why we can replace the above $\sup_{\lambda \geq 0}$ by a supremum over all $\lambda \in \mathbb{R}$, provided $t \geq \mathbb{E}[X]$. Observe that in this case, for $\lambda < 0$ and by Jensen's inequality, we have

$$\psi_X(\lambda) - \lambda t = \log \mathbb{E}[\exp(\lambda X)] - \lambda t \geq \lambda(\mathbb{E}[X] - t) \geq 0.$$

Hence we have (since a probability is always less than 1!)

$$\begin{aligned} \mathbb{P}[X \geq t] &\leq \min(1, \exp(-\sup_{\lambda \geq 0}(\lambda t - \Psi_X(\lambda)))) \\ &\leq \exp\left(-\max\left(\sup_{\lambda < 0}(\lambda t - \Psi_X(\lambda)), \sup_{\lambda \geq 0}(\lambda t - \Psi_X(\lambda))\right)\right) \\ &\leq \exp -\Psi_X^*(t). \end{aligned}$$

In the case where we are interested in the deviations of $\frac{1}{n}S_n = \frac{1}{n} \sum_{i=1}^n X_i$ with X_1, \dots, X_n i.i.d., notice $F_{S_n}(\lambda) = F_X(\lambda)^n$ by independence – a crucial property of the Laplace transform (like the characteristic function), hence $\Psi_{S_n}(\lambda) = n\Psi_X(\lambda)$ and $\Psi_{S_n}^*(t) = n\Psi_X^*\left(\frac{t}{n}\right)$. Applying Chernov's bound (3.3) to S_n , with $t = nu$ yields (3.4). \square

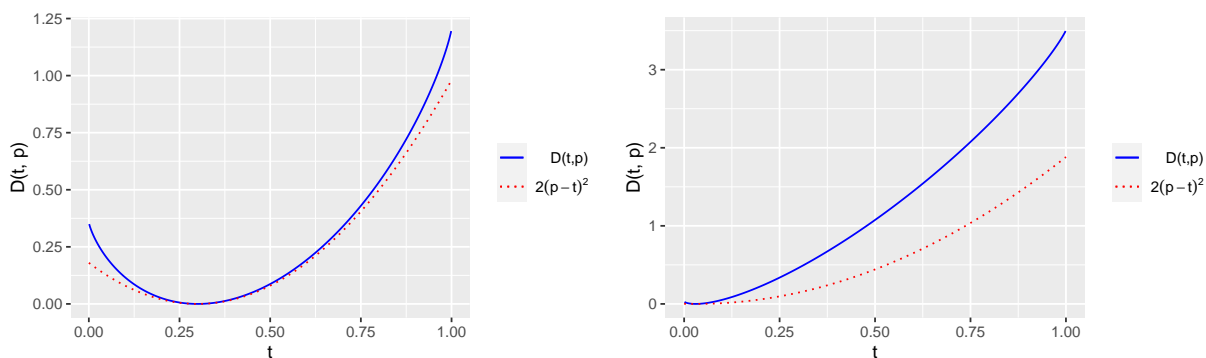


Figure 1: The function $t \mapsto D(t, p)$ and its lower bound $t \mapsto 2(p - t)^2$ (left for $p = 0.3$, right for $p = 0.03$); see Exercise 3.2.

Chernov's method for a Binomial random variable. An important application of Chernov's method is the following theorem bounding the deviations of a binomial random variable from its mean.

*

Proposition 3.4. Let $D(t, p) := t \log\left(\frac{t}{p}\right) + (1 - t) \log\left(\frac{1-t}{1-p}\right)$, defined for $(t, p) \in [0, 1] \times (0, 1)$ (with the convention $0 \log 0 = 0$). Let B_p denote a $\text{Binom}(n, p)$ random variable. Then

$$t \geq p \Rightarrow \mathbb{P}\left[\frac{1}{n}B_p \geq t\right] \leq \exp(-nD(t, p)); \quad (3.5)$$

$$t \leq p \Rightarrow \mathbb{P}\left[\frac{1}{n}B_p \leq t\right] \leq \exp(-nD(t, p)). \quad (3.6)$$

This implies, denoting $\hat{p} = \frac{1}{n}B_p$, for any $\alpha \in (0, 1]$:

$$\mathbb{P}\left[\hat{p} \geq p \text{ and } D(\hat{p}, p) \geq \frac{-\log \alpha}{n}\right] \leq \alpha; \quad (3.7)$$

$$\mathbb{P}\left[\hat{p} \leq p \text{ and } D(\hat{p}, p) \geq \frac{-\log \alpha}{n}\right] \leq \alpha; \quad (3.8)$$

$$\mathbb{P}\left[D(\hat{p}, p) \geq \frac{-\log \alpha}{n}\right] \leq 2\alpha. \quad (3.9)$$

The function $q \mapsto D(q, p)$ is convex with a minimum of 0 in $q = p$ (see Figure 1), hence is decreasing on $q \in [0, p]$ and increasing on $q \in [p, 1]$.

Exercise 3.2. Prove the quadratic lower bound $D(q, p) \geq 2(p - q)^2$ (see Fig. 1).

Proof of (3.5)-(3.6). We apply (3.4) and thus only need to compute $\Psi_X^*(t)$ where X is a Bernoulli variable of parameter p . In this case

$$F_X(\lambda) = p \exp(\lambda) + (1 - p) \text{ and } \Psi_X(\lambda) = \log((1 - p) + p \exp(\lambda)).$$

We deduce that $\Psi_X'(\lambda) = p \exp(\lambda) / ((1 - p) + p \exp(\lambda))$, finding the solution of $\Psi_X'(\lambda) = t \in (0, 1)$ is thus $\lambda = \log \frac{(1-p)t}{p(1-t)}$, so that $\Psi_X^*(t) = D(t, p)$ for $t \in (0, 1)$ (and $\Psi_X^*(t) = \infty$ otherwise), giving the conclusion.

Let us now prove the implication from (3.5) to (3.7). Recall that the function $t \mapsto D(t, p)$ is nonnegative and (strictly) increasing from $[p, 1]$ onto $[0, \log(1/p)]$. Hence we have the following equality of events, for any $t \geq p$:

$$\{\widehat{p} \geq t\} = \{D(\widehat{p}, p) \geq D(t, p) \text{ and } \widehat{p} \geq p\}.$$

For any $u \in [0, \log(1/p)]$, if we choose t as the unique solution in the interval $[p, 1]$ of $D(t, p) = u$, rewriting (3.5) in the light of the above event equality yields

$$\mathbb{P}[\widehat{p} \geq p \text{ and } D(\widehat{p}, p) \geq u] \leq \exp(-nu).$$

The latter inequality still holds trivially true if $u > \log(1/p)$, since in this case the considered event has probability zero (since $D(t, p) \leq \log(1/p)$ for all $t \geq p$). Taking $u = -\log(\alpha)/n$ yields (3.7).

Similarly (3.6) implies (3.8), and (3.7)-(3.8) imply (3.9) by a union bound. □

3.4 Sub-Gaussian random variables and Hoeffding's inequality

For a binomial variable, the Chernov bound gives a more easy to manipulate non-asymptotic bound than the Clopper-Pearson interval. However, the function $D(q, p)$ is somewhat cumbersome to handle (even if can be lower bounded by a quadratic function). Furthermore, we would like to have a deviation control applying to more general variables.

We will start with the following definition:

*** **Definition 3.5** (Sub-Gaussian random variable). A sub-Gaussian real-valued random variable X with parameter σ^2 is such that

$$\forall \lambda \in \mathbb{R} : \quad F_{(X - \mathbb{E}[X])}(\lambda) = \mathbb{E}[\exp \lambda(X - \mathbb{E}[X])] \leq \exp\left(\frac{\sigma^2 \lambda^2}{2}\right).$$

Using Prop. 3.3, we obtain immediately the following deviation control:

Proposition 3.6. *Let X be a sub-Gaussian variable with parameter σ . Then for any $\alpha \in (0, 1]$:*

$$\begin{aligned}\mathbb{P}\left[X \geq \mathbb{E}[X] + \sigma\sqrt{2\log(\alpha^{-1})}\right] &\leq \alpha; \\ \mathbb{P}\left[X \leq \mathbb{E}[X] - \sigma\sqrt{2\log(\alpha^{-1})}\right] &\leq \alpha; \\ \mathbb{P}\left[|X - \mathbb{E}[X]| \geq \sigma\sqrt{2\log(\alpha^{-1})}\right] &\leq 2\alpha.\end{aligned}$$

Proof. Without loss of generality we may assume $\mathbb{E}[X] = 0$. We apply Prop. 3.3. We have by assumption $F_X(\lambda) \leq \exp(\lambda^2\sigma^2/2)$, hence it holds $\Psi_X^*(t) \geq t^2/(2\sigma^2)$. Solving $\exp -\Psi_*(t) = \alpha$ for t , we obtain the first inequality. The second inequality is obtained by applying the same argument to $-X$, and the final one is a union bound applied with the events appearing in the first two inequalities. \square

It is straightforward to notice that sum of independent sub-Gaussian variables X_i with parameters σ_i is itself sub-Gaussian with parameter σ such that $\sigma^2 = \sum_i \sigma_i^2$. It follows:

**

Corollary 3.7. *If X_1, \dots, X_n are independent sub-Gaussian variables with respective parameters $\sigma_1^2, \dots, \sigma_n^2$, then putting $S_n := \sum_{i=1}^n X_i$, for any $\alpha \in (0, 1)$:*

$$\mathbb{P}\left[\frac{S_n}{n} \geq \mathbb{E}\left[\frac{S_n}{n}\right] + \bar{\sigma}\sqrt{\frac{2\log \alpha^{-1}}{n}}\right] \leq \alpha; \quad \mathbb{P}\left[\frac{S_n}{n} \leq \mathbb{E}\left[\frac{S_n}{n}\right] - \bar{\sigma}\sqrt{\frac{2\log \alpha^{-1}}{n}}\right] \leq \alpha, \quad (3.10)$$

where $\bar{\sigma}^2 := \frac{1}{n} \sum_{i=1}^n \sigma_i^2$.

The following proposition links boundedness to the sub-Gaussian property and is the foundation of Hoeffding's inequality coming next:

Proposition 3.8. *A random variable with values in $[a, b]$ is sub-Gaussian with parameter $(b - a)^2/4$. In particular, a random variable X such that $|X| \leq B$ a.s. is sub-Gaussian with parameter $\sigma^2 = B^2$.*

From this it follows directly from Corollary 3.7:

Corollary 3.9 (Hoeffding's inequality). *Let X_1, \dots, X_n be independent random variables taking values in the interval $[a, b]$ with $|b - a| \leq B$. Then for any $t \geq 0$:*

$$\mathbb{P} \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq t \right] \leq \exp \left(-\frac{2nt^2}{B^2} \right);$$

equivalently, for any $\alpha \in (0, 1]$:

$$\mathbb{P} \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq B \sqrt{\frac{-\log \alpha}{2n}} \right] \leq \alpha.$$

Proof of Proposition 3.8. Put $B := |b - a|$. Without loss of generality (i.e. possibly replacing X by $X' = X - \frac{a+b}{2}$), we can assume X is taking values in the interval $[-B/2, B/2]$. We start with upper bounding $\Psi_X(\lambda)$. We notice that

$$\begin{aligned} \Psi'_X(\lambda) &= \frac{\mathbb{E}[X \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]}, \\ \Psi''_X(\lambda) &= \frac{\mathbb{E}[X^2 \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]} - \left(\frac{\mathbb{E}[X \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]} \right)^2, \end{aligned}$$

where the justification of the exchange of expectation (over X) and derivation (over λ) is left as an exercise. We deduce that $\Psi_X(0) = 0$, $\Psi'_X(0) = \mathbb{E}[X]$, $\Psi''_X(\lambda) \leq B^2/4$, and by Taylor's formula with exact rest:

$$\Psi_X(\lambda) = \Psi_X(0) + \lambda \Psi'_X(0) + \frac{\lambda^2}{2} \Psi''_X(c) \leq \lambda \mathbb{E}[X] + \frac{\lambda^2 B^2}{8}.$$

Finally observe that $\Psi_{(X - \mathbb{E}[X])}(\lambda) = \Psi_X(\lambda) - \lambda \mathbb{E}[X] \leq \frac{\lambda^2}{2} \left(\frac{B}{2}\right)^2$. □

Notation: To lighten notation in the sequel, we will use the following shortcut notation:

$$\varepsilon(\delta, n) := \sqrt{\frac{-\log(\delta)}{2n}}. \tag{3.11}$$

Hoeffding's inequality allows us to bound with high probability the risk of a single prediction function from its empirical risk, provided the loss function is bounded:

Corollary 3.10. Consider a prediction setting where the loss function $\ell : \tilde{Y} \times \mathcal{Y} \rightarrow [0, B]$ is bounded by $B > 0$. Let f be a fixed prediction function, $\mathcal{E}(f)$ its risk and $\widehat{\mathcal{E}}(f)$ its empirical risk with respect to an i.i.d. sample S_n of size n . Then for any $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ with respect to the draw of the sample S_n :

$$\mathcal{E}(f) \leq \widehat{\mathcal{E}}(f) + B\varepsilon(\delta, n). \quad (3.12)$$

Similarly, it holds with probability at least $1 - \delta$:

$$\widehat{\mathcal{E}}(f) \leq \mathcal{E}(f) + B\varepsilon(\delta, n). \quad (3.13)$$

and it holds with probability at least $1 - \delta$:

$$\left| \widehat{\mathcal{E}}(f) - \mathcal{E}(f) \right| \leq B\varepsilon(\delta/2, n). \quad (3.14)$$

Proof. The first inequality is an immediate consequence of Hoeffding's inequality, since $\widehat{\mathcal{E}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i)$; put $W_i := \ell(f(X_i), Y_i)$, then the W_i s are i.i.d., taking values in $[0, B]$ and $\mathbb{E}[W_i] = \mathcal{E}(f)$. The second inequality is derived similarly, using $-W_i$ instead of W_i , and the third is obtained from the two first ones each applied with $\delta' = \delta/2$ and a union bound argument. \square

3.5 Uniform bounds over a finite class of prediction functions

We now turn to the situation where we wish to get bounds on the risks of several prediction functions using their empirical risks, uniformly over a class \mathcal{F} . To be specific, from 3.10 we know that (when the loss function is bounded by B)

$$\forall f \in \mathcal{F} : \quad \mathbb{P}\left[\mathcal{E}(f) \leq \widehat{\mathcal{E}}(f) + B\varepsilon(\delta, n)\right] \geq 1 - \delta; \quad (3.15)$$

however we wish to have a statement of the form

$$\mathbb{P}\left[\forall f \in \mathcal{F} : \mathcal{E}(f) \leq \widehat{\mathcal{E}}(f) + \mathcal{B}(f, \delta)\right] \geq 1 - \delta, \quad (3.16)$$

for some appropriate bound function \mathcal{B} . Notice the difference between (3.15) and (3.16): the second form will guarantee that *all* bounds, simultaneously, are valid with probability at least $1 - \delta$, while there is no such uniformity in the first statement.

We will start with a finite class \mathcal{F} .

Proposition 3.11. *Assume a prediction problem with bounded loss function taking values in $[0, B]$. Let \mathcal{F} be a finite set of prediction functions of cardinality K . Assume empirical risks are computed on an i.i.d. sample S_n of size n . Then it holds with probability at least $1 - \delta$ with respect to the draw of the sample S_n :*

$$\forall f \in \mathcal{F} : \quad \mathcal{E}(f) \leq \widehat{\mathcal{E}}(f) + B\sqrt{\frac{\log K}{2n}} + B\varepsilon(\delta, n). \quad (3.17)$$

Similarly, it holds with probability at least $1 - \delta$:

$$\forall f \in \mathcal{F} : \quad \mathcal{E}(f) \leq \widehat{\mathcal{E}}(f) + B\sqrt{\frac{\log K}{2n}} + B\varepsilon(\delta, n). \quad (3.18)$$

and it holds with probability at least $1 - \delta$:

$$\forall f \in \mathcal{F} : \quad \left| \mathcal{E}(f) - \widehat{\mathcal{E}}(f) \right| \leq B\sqrt{\frac{\log 2K}{2n}} + B\varepsilon(\delta, n). \quad (3.19)$$

Proof. This is a direct consequence of the union bound. Denote $A(f, \delta)$ the event where the bound $\mathcal{E}(f) \leq \widehat{\mathcal{E}}(f) + B\varepsilon(\delta, n)$ is satisfied. From Corollary (3.10), and (3.15) it holds that $\forall f \in \mathcal{F} \quad \mathbb{P}[A^c(f, \delta)] \leq \delta$ for any $\delta \in (0, 1)$. Therefore

$$\mathbb{P} \left[\bigcap_{f \in \mathcal{F}} A_{f, \delta} \right] = 1 - \mathbb{P} \left[\bigcup_{f \in \mathcal{F}} A_{f, \delta}^c \right] \geq 1 - \sum_{f \in \mathcal{F}} \mathbb{P}[A_{f, \delta}^c] \geq 1 - K\delta.$$

Replacing δ by δ/K in the above inequality, we therefore have

$$\mathbb{P} \left[\bigcap_{f \in \mathcal{F}} A_{f, \delta/K} \right] \geq 1 - \delta.$$

This gives the first announced inequality, just noticing that

$$\varepsilon \left(\frac{\delta}{K}, n \right) = \sqrt{\frac{\log K + \log \delta^{-1}}{2n}} \leq \sqrt{\frac{\log K}{2n}} + \varepsilon(\delta, n).$$

The second and third inequalities are obtained similarly from the corresponding single function inequalities of Corollary (3.10) and a union bound. \square

A first consequence of Proposition (3.11) is that we can derive a bound on the risk of an estimator \widehat{f} taking values in a finite class \mathcal{F} .

Corollary 3.12. Consider the same assumptions as in Proposition 3.11. Let \hat{f} be an estimator taking its values in the finite class \mathcal{F} of cardinality K . Then with probability at least $(1 - \delta)$ of the draw of the sample S_n , it holds

$$\mathcal{E}(\hat{f}_{S_n}) \leq \hat{\mathcal{E}}(\hat{f}_{S_n}, S_n) + B\sqrt{\frac{\log K}{2n}} + B\varepsilon(\delta, n). \quad (3.20)$$

Proof. We consider event (3.17), whose probability is at least $1 - \delta$. If this event is satisfied, we can in particular specialize the inequality which holds for all $f \in \mathcal{F}$ for the particular choice $\hat{f} \in \mathcal{F}$ (note that it does not matter that \hat{f} depends on the data now). This yields the conclusion. \square

We now consider more specifically the ERM over the finite class \mathcal{F} , and get the following result as a further consequence of Proposition 3.11.

Proposition 3.13. Assume a prediction problem with bounded loss function taking values in $[0, B]$. Let \mathcal{F} be a finite set of prediction functions of cardinality K . Assume empirical risks are computed on an i.i.d. sample S_n of size n . Consider the ERM estimator $\hat{f} = f_{\hat{k}}$, where

$$\hat{k} \in \underset{1 \leq k \leq K}{\text{Arg Min}} \hat{\mathcal{E}}(f_k).$$

Then for any $\delta \in (0, 1)$ it holds with probability at least $1 - \delta$:

$$\mathcal{E}(f_{\hat{k}}) \leq \mathcal{E}_{\mathcal{F}}^* + 2B\sqrt{\frac{\log K}{2n}} + 2B\varepsilon(\delta/2, n). \quad (3.21)$$

Observe that inequality (3.21) relates the risk of the ERM to the best risk in the class $\mathcal{E}_{\mathcal{F}}^* = \min_{1 \leq k \leq K} \mathcal{E}(f_k)$, hence it is an *excess risk* bound with respect to the class \mathcal{F} . It is trivial but useful to rewrite it in the following way to bound the excess risk with respect to all possible decision functions:

$$\mathcal{E}(f_{\hat{k}}) - \mathcal{E}^* \leq (\mathcal{E}_{\mathcal{F}}^* - \mathcal{E}^*) + 2B\sqrt{\frac{\log K}{2n}} + 2B\varepsilon(\delta/2, n). \quad (3.22)$$

Observe in particular the role of n (the sample size) and of K (the size of class \mathcal{F}). As expected, larger “complexity” of the class (in this simple scenario the complexity is simply the cardinality) results in a worse bound for the excess risk in (3.21), however we can expect that a more complex (i.e. larger) class also has in advantage in that it will have a smaller value for $(\mathcal{E}_{\mathcal{F}}^* - \mathcal{E}^*)$, in the bound (3.22).

The good news is that the cardinality K of the class enter only logarithmically in the bound. For instance taking $K = n^p$ (for any fixed p possibly much larger than 1) still results in a bound on the excess risk behaving in $\mathcal{O}(\sqrt{\log(n)/n})$.

Proof. We apply (the last inequality of) Proposition 3.11 so that the following event has probability at least $1 - \delta$:

$$\forall f \in \mathcal{F} : \quad \left| \mathcal{E}(f) - \widehat{\mathcal{E}}(f) \right| \leq B \sqrt{\frac{\log K}{2n}} + B\varepsilon(\delta/2, n). \quad (3.23)$$

In the remainder of the proof, we assume that the latter event holds, and do not repeat everytime “with probability at least $1 - \delta$ ”. We can in particular apply the bound (3.23) to any specific $f \in \mathcal{F}$, even data-dependent, as is $f_{\widehat{k}}$; this is precisely the purpose of having a uniform bound. Thus, we deduce that

$$\mathcal{E}(f_{\widehat{k}}) \leq \widehat{\mathcal{E}}(f_{\widehat{k}}) + B \sqrt{\frac{\log K}{2n}} + B\varepsilon(\delta/2, n). \quad (3.24)$$

Let $k^* \in \text{Arg Min}_{1 \leq k \leq K} \mathcal{E}(f_k)$ be fixed, and apply also (3.23) to f_{k^*} :

$$\widehat{\mathcal{E}}(f_{k^*}) \leq \mathcal{E}(f_{k^*}) + B \sqrt{\frac{\log(2\delta^{-1})}{2n}}. \quad (3.25)$$

Now using (3.24), (3.25), and the definition of $f_{\widehat{k}}$ (minimizing the empirical risk) and f_{k^*} (minimizing the risk), respectively, we obtain

$$\begin{aligned} \mathcal{E}(f_{\widehat{k}}) &\leq \widehat{\mathcal{E}}(f_{\widehat{k}}) + B \sqrt{\frac{\log K}{2n}} + B\varepsilon(\delta/2, n) \\ &\leq \widehat{\mathcal{E}}(f_{k^*}) + B \sqrt{\frac{\log K}{2n}} + B\varepsilon(\delta/2, n) \\ &\leq \mathcal{E}(f_{k^*}) + 2B \sqrt{\frac{\log K}{2n}} + 2B\varepsilon(\delta/2, n) \\ &= \min_{1 \leq k \leq K} \mathcal{E}(f_k) + 2B \sqrt{\frac{\log K}{2n}} + 2B\varepsilon(\delta/2, n). \end{aligned}$$

□

**

Hold-out estimator selection. The situation where we consider a finite set of candidate prediction functions is found commonly when we have several estimators $\widehat{f}_1, \dots, \widehat{f}_K$, amongst which we want to select one. We assume here that these estimators are “black boxes”, i.e. they may be overfitting the data they are trained on. To select one of these we again use the idea of Hold-Out presented at the beginning of the chapter:

1. Obtain two independent samples S_n, T_m of i.i.d. data. (Possibly split an existing sample into two separate sub-samples to do so).

2. Train each of the estimators $\widehat{f}_1, \dots, \widehat{f}_K$ using the sample S_n , resulting in the decision functions $\widehat{f}_{1,S_n}, \dots, \widehat{f}_{K,S_n}$.

3. Pick an estimator

$$\widehat{k} \in \underset{1 \leq k \leq K}{\text{Arg Min}} \widehat{\mathcal{E}}(\widehat{f}_{k,S_n}, T_m). \quad (3.26)$$

Observe that, *conditionally to* S_n (i.e. considering the sample S_n “fixed”), the decision functions $\widehat{f}_{1,S_n}, \dots, \widehat{f}_{K,S_n}$ can be considered as fixed too (since they depend only on S_n , and not on T_n). Therefore, (3.26) can be seen as an ERM method conditional to S_n , over the class $\mathcal{F}(S_n) := \left\{ f_k = \widehat{f}_{k,S_n}, k = 1, \dots, K \right\}$.

This *Hold-Out Selection method* is used commonly in the case where we have an estimator method \widehat{f}_λ depending on a “tuning parameter” λ (see for instance the regularization methods introduced in Section 2.7) that we want to choose in order to minimize the risk. In this case we can restrict the values of λ to discretized set $\{\lambda_1, \dots, \lambda_K\}$ and we use the previous strategy with $\widehat{f}_k = \widehat{f}_{\lambda_k}$.

* **Cross-validation.** In practice, it is considered somewhat wasteful to split the data into a training sample S_n and a “validation sample” T_m , as in the Hold-Out method. We mention the very common *cross-validation* approach which can be seen as an elaboration of the hold-out. We still assume to have a family of estimators $\widehat{f}_1, \dots, \widehat{f}_K$.

- Fix an integer $V \geq 2$.

- Given a sample S_n , of size n split it in (approximately) equal size, disjoint sub-samples $S^{(1)}, \dots, S^{(V)}$, each of size $\geq \lfloor \frac{n}{V} \rfloor$.

- Denote $S^{(-j)}$ the sample made of the reunion of $(S^{(i)})_{i \neq j}$.

- Define

$$\widehat{f}_k^{(-j)} = \widehat{f}_{k,S^{(-j)}}$$

the k -th estimator trained on all the samples except those of $S^{(j)}$.

- Let

$$\widehat{k} \in \underset{1 \leq k \leq K}{\text{Arg Min}} \sum_{j=1}^V \widehat{\mathcal{E}}(\widehat{f}_k^{(-j)}, S^{(j)}). \quad (3.27)$$

- as a final estimator, pick $\widehat{f} := \widehat{f}_{\widehat{k},S_n}$, which corresponds to estimator \widehat{k} trained on the entire sample S_n .

Note that (3.27) can be seen as a “multiple” hold-out where the disjoint samples $S^{(j)}, S^{(-j)}$ take the role of training and validation samples in the standard Hold-out, for all $j = 1, \dots, V$ in turn.

Each term in the sum (3.27) has expectation $\mathcal{E}(\widehat{f}_{k,n'})$, where $\widehat{f}_{k,n'}$ denotes estimator \widehat{f}_k trained using a sample of size $n\frac{(V-1)}{V}$, and it is possible to apply the previous arguments based on Hoeffding's inequality to bound it. However, the hope is that the "aggregation" of hold-out errors in (3.27) results in a yet sharper estimate.

A precise theoretical analysis of cross-validation, and why it may outperform Hold-out in practice, is a delicate subject that we won't touch here. (Notice in particular that we compare the empirical risks of estimators \widehat{f}_k trained on samples of size $n(V-1)/V$, while the final estimator is trained on the total sample of size n . Therefore, a minima one must make some kind of assumptions on the fact that the estimators trained on the full sample and on a sub-sample are "close" in some way).

Still, cross-validation is the "default" method in practice to pick the tuning parameters of a learning method. The particular case $V = n$ is called "leave-one-out" (one removes, in turn, a single data point of the sample, trains each estimator on the remaining $(n-1)$ data, and monitors its error on the point that has been left out).

3.6 Uniform bounds over a countable class of prediction functions; regularized ERM

In this section we present an extension of the arguments used previously, allowing to deal with countably infinite classes.

Proposition 3.14. *Assume a prediction problem with bounded loss function taking values in $[0, B]$. Let \mathcal{F} be a finite or countably infinite set of prediction functions. Assume that π is a set of real weights over \mathcal{F} , such that $\pi(f) \in [0, 1]$ and:*

$$\sum_{f \in \mathcal{F}} \pi(f) \leq 1. \quad (3.28)$$

Assume empirical risks are computed on an i.i.d. sample S_n of size n . Then for any $\delta \in (0, 1]$, it holds with probability at least $1 - \delta$ with respect to the draw of the sample S_n :

$$\forall f \in \mathcal{F} : \quad \mathcal{E}(f) \leq \widehat{\mathcal{E}}(f) + B\sqrt{\frac{\log(\pi(f)^{-1})}{2n}} + B\varepsilon(\delta, n). \quad (3.29)$$

Proof. The proof is a minor variation on the proof of Prop. 3.11. Denote $t(\delta, n) = B\sqrt{\log(\delta^{-1})/(2n)}$, and $A(f, \delta)$ the event where the bound $\mathcal{E}(f) \leq \widehat{\mathcal{E}}(f) + t(\pi(f)\delta, n)$ is satisfied. From Corollary (3.10), and (3.15) it holds that $\forall f \in \mathcal{F} \quad \mathbb{P}[A^c(f, \delta)] \leq \pi(f)\delta$ for any $\delta \in (0, 1)$. Therefore

$$\mathbb{P}\left[\bigcap_{f \in \mathcal{F}} A_{f,\delta}\right] = 1 - \mathbb{P}\left[\bigcup_{f \in \mathcal{F}} A_{f,\delta}^c\right] \geq 1 - \sum_{f \in \mathcal{F}} \mathbb{P}[A_{f,\delta}^c] \geq 1 - \delta \sum_{f \in \mathcal{F}} \pi(f) \geq 1 - \delta.$$

Elementary bounding for $t(\pi(f)\delta, n)$ then yields the claim. \square

Remarks.

- It is an extension of Proposition 3.11, since the latter can be obtained via the uniform weight choice $\pi(f) = 1/K$ for all $f \in \mathcal{F}$ (finite class).
- It is always better to have the constraint (3.28) satisfied with equality (if not, replace $\pi(f)$ by $\pi'(f) = \pi(f)/(\sum_{g \in \mathcal{F}} \pi(g))$, which improves (3.17). With the equality constraint, it is possible to interpret π as a discrete probability distribution over \mathcal{F} . However since there is no probabilistic argument used, we prefer the term “weights”.
- The choice of π is arbitrary, but must be fixed in advance (i.e. it cannot depend on the data). One can interpret $\log(\pi(f)^{-1})$ as a “complexity” of function f but it is not an intrinsic notion since we can decide freely π . Rather, when we pick π we have to decide which functions are considered “less complex” (=are given a higher weight), and we are limited by the sum 1 constraint: we can consider many functions as “complex” but not too many as “simple”.
- The fact that we have a bound on an infinite set of functions is somewhat of an illusion: it \mathcal{F} is countably infinite, it means that only a finite number of them can have weight larger than any given t . On the other hand we observe that (3.17) is trivial for all functions f with $\pi(f) \leq \exp(-2n)$ (since the risk is always trivially bounded by B). So for any given n , bound (3.17) gives non-trivial information only for a finite subset of functions of \mathcal{F} . On the other hand, the set of functions with a non-trivial bound grows with n (and will contain any given function f with $\pi(f) > 0$ for n big enough). In a way, this can be seen as a way to have a class \mathcal{F}_n that depends on n , but since it is implicit, it is more elegant.

Regularized ERM. Assume that we work under the same assumptions as Proposition 3.14. Bound (3.17) cannot be used to analyze the ERM on the (countably infinite) class \mathcal{F} in a meaningful way, because, if we try to follow the argument of the proof of Prop. 3.13 (we recommend it as an exercise to see where it fails) we will have a remaining term of the form $\sqrt{-\log \pi(\hat{f}^{ERM})}/2n$ in the bound; and we have no means to control it, it can be arbitrary large.

Instead, we can use bound (3.17) to *design* an estimator which will consist in minimizing that bound (for a fixed choice of π). Let us therefore define the *regularized ERM* (based on the regularization induced by π):

$$\hat{f}^\pi \in \operatorname{Arg Min}_{f \in \mathcal{F}} \left(\hat{\mathcal{E}}(f) + B \sqrt{\frac{-\log(\pi(f))}{2n}} \right). \quad (3.30)$$

Lemma 3.15. *If the loss function $\ell(\cdot, \dots)$ takes its values in $[0, B]$, the estimator \hat{f}^π always exists, i.e. the Arg Min in (3.30) is never empty.*

Proof. The property is obvious if \mathcal{F} is finite; so here we assume that \mathcal{F} is countably infinite. Since the weights $\pi(f), f \in \mathcal{F}$ are positive and summable, they can be ranked by nonincreasing order, i.e. there exists an indexation $\mathcal{F} = \{f_i, i \in \mathbb{N}\}$ such that $\pi(f_i)$ forms a nonincreasing sequence. Since $\ell(\cdot, \cdot) \in [0, B]$, it holds $\widehat{\mathcal{E}}(f) \in [0, B]$ for all $f \in \mathcal{F}$.

It follows that

$$\widehat{\mathcal{E}}(f_0) + B\sqrt{\frac{-\log(\pi(f_0))}{2n}} \leq B\left(1 + \sqrt{\frac{-\log(\pi(f_0))}{2n}}\right),$$

while for all $f_i \in \mathcal{F}$:

$$\widehat{\mathcal{E}}(f_i) \geq B\sqrt{\frac{-\log(\pi(f_i))}{2n}}.$$

Let $i_* = \min\{i \in \mathbb{N} : \sqrt{-\log(\pi(f_i))} > \sqrt{2n} + \sqrt{-\log(\pi(f_0))}\}$, then for all $i \geq i_*$ it holds $\widehat{\mathcal{E}}(f_i) \geq B\sqrt{\frac{-\log(\pi(f_i))}{2n}} > \widehat{\mathcal{E}}(f_0) + B\sqrt{\frac{-\log(\pi(f_0))}{2n}}$. It follows that a minimum for the expression (3.30) exists, and must be attained for some f_i with $i < i_*$. \square

Proposition 3.16. *Consider the same setting as in Proposition 3.14 and let \widehat{f}^π be the estimator defined by (3.30). Then for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ it holds*

$$\mathcal{E}(\widehat{f}^\pi) \leq \min_{f \in \mathcal{F}} \left(\mathcal{E}(f) + 2B\sqrt{\frac{-\log(\pi(f))}{2n}} \right) + 2B\varepsilon(\delta/2, n). \quad (3.31)$$

Proof. The proof is very similar to that of Prop. 3.13. We apply Prop. 3.14 and obtain that with probability at least $1 - \delta/2$, we have

$$\forall f \in \mathcal{F} : \quad \mathcal{E}(f) \leq \widehat{\mathcal{E}}(f) + B\sqrt{\frac{\log(\pi(f)^{-1})}{2n}} + B\varepsilon(\delta/2, n), \quad (3.32)$$

Let f_*^π achieving the minimum on the right-hand side of (3.31). Note that the minimum exists by using the same type of argument as in Lemma 3.15 (replacing the role of the empirical risk by the true risk). The decision function f_*^π is non-random and we can apply the simple Hoeffding's inequality to it, so that with probability at least $1 - \delta/2$ it holds

$$\widehat{\mathcal{E}}(f_*^\pi) \leq \mathcal{E}(f_*^\pi) + B\varepsilon(\delta/2, n). \quad (3.33)$$

Now using (3.32), (3.33) (holding simultaneously with probability at least $1 - \delta$), and the

definition of \widehat{f}^π, f_π^* , respectively, we obtain

$$\begin{aligned}
\mathcal{E}(\widehat{f}^\pi) &\leq \widehat{\mathcal{E}}(\widehat{f}^\pi) + B\sqrt{\frac{-\log \pi(\widehat{f}^\pi)}{2n}} + B\varepsilon(\delta/2, n) \\
&\leq \widehat{\mathcal{E}}(f_\pi^*) + B\sqrt{\frac{-\log \pi(f_\pi^*)}{2n}} + B\varepsilon(\delta/2, n) \\
&\leq \mathcal{E}(f_\pi^*) + B\sqrt{\frac{-\log \pi(f_\pi^*)}{2n}} + 2B\varepsilon(\delta/2, n) \\
&= \min_{f \in \mathcal{F}} \left(\mathcal{E}(f) + B\sqrt{\frac{-\log \pi(f)}{2n}} \right) + 2B\varepsilon(\delta/2, n).
\end{aligned}$$

□

In particular, we obtain the following consistency property on a countable family of decision functions.

Corollary 3.17. *Consider the same settings as in Propositions 3.14 and 3.16, let π be set of **strictly** positive real weights on the countable family \mathcal{F} satisfying (3.28). Let \widehat{f}_n^π be the estimator defined by (3.30) from an i.i.d. sample S_n of size n . Then the sequence \widehat{f}_n^π is consistent in probability on \mathcal{F} , that is, $\mathcal{E}(\widehat{f}_n^\pi)$ converges to $\mathcal{E}_\mathcal{F}^*$ in probability, as $n \rightarrow \infty$.*

Proof. Let $t > 0$ be fixed. Let $f_t^* \in \mathcal{F}$ be such that $\mathcal{E}(f_t^*) \leq \inf_{f \in \mathcal{F}} \mathcal{E}(f) + \frac{t}{2}$. Let us now fix any $\delta \in (0, 1)$. Then event (3.31) (holding with probability at least $1 - \delta$) implies that

$$\begin{aligned}
\mathcal{E}_\mathcal{F}^* \leq \mathcal{E}(\widehat{f}_n^\pi) &\leq \left(\mathcal{E}(f_t^*) + 2B\sqrt{\frac{-\log(\pi(f_t^*))}{2n}} \right) + 2B\varepsilon(\delta/2, n) \\
&\leq \mathcal{E}_\mathcal{F}^* + \frac{t}{2} + 2B\sqrt{\frac{-\log(\pi(f_t^*))}{2n}} + 2B\varepsilon(\delta/2, n).
\end{aligned}$$

The two last terms in the above bound converge to zero as $n \rightarrow \infty$, so for any δ , for any n large enough, it holds that

$$\mathbb{P} \left[\left| \mathcal{E}(\widehat{f}_n^\pi) - \mathcal{E}_\mathcal{F}^* \right| > t \right] \leq \delta,$$

in other words $\mathbb{P} \left[\left| \mathcal{E}(\widehat{f}_n^\pi) - \mathcal{E}_\mathcal{F}^* \right| > t \right] \rightarrow 0$ as $n \rightarrow \infty$; this is true for any $t > 0$, hence the conclusion. □

* **An application to interval-based classification.** Consider a binary classification problem ($\mathcal{Y} = \widetilde{\mathcal{Y}} = \{0, 1\}$, 0-1 loss) for $\mathcal{X} = [0, 1]$ and denote \mathcal{F}_K the set of piecewise constant classification functions, constant on each of the K intervals of the form $[\frac{j-1}{K}, \frac{j}{K}]$,

$j \in \llbracket K \rrbracket$. Since classification functions can only take two values, we have $|\mathcal{F}_K| = 2^K$. Furthermore, we define a family of weights on $\mathcal{F} := \bigcup_{K \geq 1} \mathcal{F}_K$ via

$$\pi(f) = \frac{1}{cK^2} 2^{-K}, \text{ if } f \in \mathcal{F}_K,$$

for $c = \pi^2/6$. (It is possible that the same decision function belongs to \mathcal{F}_K for several K s, in which case we just take the smallest K having this property). It is easy to check that these weights satisfy (3.28). Applying Proposition 3.16, the regularized ERM estimator \widehat{f}^π using these weights satisfies with probability at least $1 - \delta$:

$$\mathcal{E}(\widehat{f}^\pi) \leq \min_{K \geq 1} \min_{f \in \mathcal{F}_K} \left(\mathcal{E}(f) + 2B \sqrt{\frac{K \log 2}{2n}} + 2B \sqrt{\frac{\log c + 2 \log K}{2n}} \right) + 2B\varepsilon(\delta/2, n).$$

Observe that if we choose a fixed K in advance, and consider the ERM on \mathcal{F}_K , applying Prop. 3.13 yields a bound similar to the above (without the third term), albeit for this fixed K only. As we have seen in Proposition 1.13 for piecewise-constant functions in a slightly different context, choosing K of the right order is important in order to obtain fast convergence rates. By contrast, above inequality for the regularized ERM tells us that we get a bound as good as what we would obtain for the “best” choice K for the ERM on \mathcal{F}_K – the price to pay for this *adaptivity* is the additional third term in the bound, which is modest since it is negligible with respect to the second term for large K . The above type of bound is sometimes called an *oracle inequality*: the risk of the regularized estimator is (almost) as good as if an “oracle” had told us in advance which \mathcal{F}_K to choose to minimize the corresponding ERM risk.

◇ **An application to decision trees.** We will consider an application of the previous principle to (a certain type of) decision trees. Decision trees are a certain class of decision functions that are piecewise constant on the input space \mathcal{X} and whose pieces are defined by recursive partitioning of \mathcal{X} . More formally, a decision tree decision function is given by:

- A complete binary tree structure T ; complete means that each node either has 2 daughter-nodes (it is then called an interior node) or has no descendents (it is then called a leaf). Let \mathring{T} denote the set of interior nodes, and ∂T the set of leaves.
- For each interior node $s \in \mathring{T}$, a *question* q_s which is a function $\mathcal{X} \rightarrow \{0, 1\}$. We will assume that $q_s \in \mathcal{Q}$, where \mathcal{Q} is a finite library of questions.
- For each leaf $t \in \partial T$, a prediction value $\tilde{y}_t \in \tilde{Y}$.

The set $\mathcal{T}(\mathcal{Q})$ of possible triplets $(T, (q_s)_{s \in \mathring{T}}, (\tilde{y}_t)_{t \in \partial T})$ will be called the set of decision trees (with questions belonging to \mathcal{Q}).

Given the above structure and parameters $\overline{T} = (T, (q_s)_{s \in \mathring{T}}, (\tilde{y}_t)_{t \in \partial T})$, the associated decision function $f_{\overline{T}}$ can be defined algorithmically as follows: for a given input point x :

1. Let s be the root node of the tree T .
2. If s is a leaf, return the value $f_{\bar{T}}(x) = \tilde{y}_s$.
3. Otherwise, s is an interior node. Then compute $q_s(x) \in \{0, 1\}$.
4. If $q_s(x) = 1$, replace s by its right daughter-node, otherwise replace s by its left daughter-node.
5. Return to step 2.

There are several classical methods available to build a decision tree function $\hat{f} = f_{\hat{T}}$, corresponding to a certain triplet $\hat{T} \in \mathcal{T}(\mathcal{Q})$, from a data sample S_n . We will not present them in detail here. If we consider the output of such a method \hat{f}_{S_n} , we would like to be able to have a confidence bound on its risk without knowing the internal details of the method. Here we will not use a hold-out sample approach but rather use the same sample S_n , this is why we resort to finding confidence bounds that are valid for all functions of \mathcal{T} with high probability (following the same approach as in Proposition 3.16 and Corollary 3.12 for finite classes).

Proposition 3.18. *Assume a prediction setting with a loss function bounded by B . Let \mathcal{Q} be a finite set of questions, and assume $\tilde{\mathcal{Y}}$ is a finite set. Let S_n be an i.i.d. training sample of size n which will be used to compute all empirical risks. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds:*

$$\forall \bar{T} \in \mathcal{T} : \quad \mathcal{E}(f_{\bar{T}}) \leq \hat{\mathcal{E}}(f_{\bar{T}}) + B\sqrt{\frac{C|\bar{T}|}{2n}} + B\sqrt{\frac{\log \delta^{-1}}{2n}}, \quad (3.34)$$

where $C := \log|\mathcal{Q}| + \log|\tilde{\mathcal{Y}}| + \log 4$, and for $\bar{T} = (T, (q_s)_{s \in \bar{T}}, (\tilde{y}_{t \in \partial T}))$ we denote $|\bar{T}| := |\partial T|$.

As a consequence, if \hat{f} is an estimator taking values in $\{f_{\bar{T}}, \bar{T} \in \mathcal{T}\}$, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds

$$\mathcal{E}(\hat{f}_{S_n}) \leq \hat{\mathcal{E}}(\hat{f}_{S_n}, S_n) + B\sqrt{\frac{C|\hat{f}_{S_n}|}{2n}} + B\sqrt{\frac{\log \delta^{-1}}{2n}},$$

where we denoted $|\hat{f}_{S_n}| = |\bar{T}|$, $\bar{T} \in \mathcal{T}$.

Proof. We want to apply Prop. (3.14), and for this need to choose a weight function π over \mathcal{T} (we consider directly a weight function over \mathcal{T} , which induces of course a weight function on $\{f_{\bar{T}}, \bar{T} \in \mathcal{T}\}$). Although we insisted that there is no probabilistic argument linked to π , it is easier to describe π as a probability distribution. We construct π as a probability distribution in the following way:

- The marginal of the binary tree structure T is taken to be a Galton-Watson process where each node had a probability ρ to have 0 descendents, and $(1 - \rho)$ to have 2 daughters (with $\rho \in (0, \frac{1}{2}]$). Thus

$$\pi(T) := \rho^{|\mathring{T}|} (1 - \rho)^{|\partial T|} = (1 - \rho)^{-1} [\rho(1 - \rho)]^{|\partial T|},$$

since $|\partial T| = |\mathring{T}| + 1$ for a complete binary tree (check this property!)

- Conditional to T , we take the distribution on the questions in the interior nodes of T , $(q_s)_{s \in \mathring{T}}$ and of the predictions in the leaves of T , $(\tilde{y}_t)_{t \in \partial T}$ to be independent uniform over \mathcal{Q} and $\tilde{\mathcal{Y}}$, respectively.

We plug this choice (taking $\rho = \frac{1}{2}$) into (3.29), and it holds

$$\begin{aligned} \log(\pi(\bar{T})^{-1}) &= -\log \pi(T) - \log \pi((q_s)_{s \in \mathring{T}}, (\tilde{y}_t)_{t \in \partial T} | T) \\ &= -\log(2) + |\partial T| \log 4 + (|\partial T| - 1) \log |\mathcal{Q}| + |\partial T| \log |\tilde{\mathcal{Y}}| \\ &\leq |\partial T| (\log(4) + \log |\mathcal{Q}| + \log |\tilde{\mathcal{Y}}|), \end{aligned}$$

which leads to the conclusion. □

Observe that with the above choice of weights, the size of the tree $|\partial T|$ plays a natural role of “complexity” of the associated decision function. Moreover, in practice we may want to pick a decision tree that minimizes (at least as much as possible, its exact minimization is not exactly possible) the bound (3.34), which is regularized ERM (or approximate ERM if exact minimization is not possible) with the regularization penalty $\Omega(\bar{T}) = B \sqrt{C |\bar{T}| / 2n}$.

4 The Nearest Neighbors method

In this section we introduce and propose an elementary analysis of nearest-neighbors (NN) methods, which are part of the classical toolbox for prediction methods. In a nutshell, the output of the NN method from a training sample S_n is a decision function f such as the prediction $f(x)$ at a point x is obtained by looking up the neighbors of x in the training set, and taking a decision based on a local average or majority vote among the labels associated to these neighbors.

4.1 Basic notation and definitions

We assume that \mathcal{X} is a Polish space (complete, separable metric space, made into a measurable space by endowing it with its Borel σ -algebra), we denote $d(\cdot, \cdot)$ the metric on \mathcal{X} . We assume given a training sample $S_n = (X_i, Y_i)_{1 \leq i \leq n}$ of size n (as usual, assumed to be drawn i.i.d. from a generating distribution \mathbb{P}_{XY}).

Given a point $x \in \mathcal{X}$, there exists a permutation σ_x (depending on S_n , therefore random) such that:

$$d(x, X_{\sigma_x(1)}) \leq d(x, X_{\sigma_x(2)}) \leq \dots \leq d(x, X_{\sigma_x(n)}).$$

(We assume that in case of ties, the tied points are ordered by their indices in the training set. It can be checked that it makes the permutation σ_x a measurable function of the training set S_n).

We introduce the notation

$$X^{(k)}(x) := X_{\sigma_x(k)}, \quad Y^{(k)} := Y_{\sigma_x(k)}(x), \quad k = 1, \dots, n;$$

$X^{(k)}(x)$ is called the k -th nearest neighbor of x (in the sample S_n), and $Y^{(k)}(x)$ is its label.

In the regression setting ($\mathcal{Y} = \tilde{\mathcal{Y}} = \mathbb{R}$), for a given integer k the k -nearest-neighbor (k -NN) prediction function based on the sample S_n is defined as

$$\hat{f}_{k\text{-NN}}(x) := \frac{1}{k} \sum_{i=1}^k Y^{(i)}(x).$$

In the classification setting ($\mathcal{Y} = \tilde{\mathcal{Y}} = \{1, \dots, K\}$), a k -NN classifier function is defined as

$$\hat{f}_{k\text{-NN}}(x) \in \underset{c=1, \dots, K}{\text{Arg Max}} \left(\sum_{i=1}^k \mathbf{1}\{Y^{(i)}(x) = c\} \right),$$

i.e. $\hat{f}_{k\text{-NN}}$ predicts the majority class among the neighbors of x in the sample (in case of ties, once again one can decide to predict among the tied classes the one with the smallest index).

The definition of the k -NN decision function is simple and natural, and there exists a vast literature on the subject going back to the 1960s. In this chapter, we will concentrate on the following questions concerning the asymptotic behavior of the k -NN prediction (more precisely in the case of binary classification):

- what is the behavior of the risk $\mathcal{E}(\widehat{f}_{k\text{-NN}})$, as the sample size n grows but k is fixed?
- what is the behavior of the risk $\mathcal{E}(\widehat{f}_{k\text{-NN}})$, as the sample size n grows and $k(n)$ is allowed to grow with n ?

Note that it is natural to let $k(n)$ grow with n , since intuitively local averages are more accurate if we use more data (but not too many since we want to remain in a neighborhood of the prediction point).

4.2 Analysis for k fixed

As announced, from now on we will focus on binary classification $\mathcal{Y} = \widetilde{\mathcal{Y}} = \{0, 1\}$ with the 0-1 classification loss. We denote, as in previous chapters, $\eta(x) = \mathbb{P}[Y = 1|X = x]$. We will make the following assumption:

$$\eta \text{ is a continuous function.} \tag{Cont}$$

We introduce some additional notation: for $p \in [0, 1]$, denote

$$Q_k(p) := \mathbb{P}\left[B_{k,p} < \frac{k}{2}\right], \text{ where } B_{k,p} \sim \text{Binom}(k, p).$$

**

Theorem 4.1. *In the binary classification setting, assuming (Cont) holds, we have:*

1. $\forall x \in \text{Supp}(\mathbb{P}_X), \mathbb{E}_{S_n} \mathbb{E}\left[\ell(\widehat{f}_{k\text{-NN}}^{(n)}(x), Y)|X = x\right] \rightarrow \alpha_k(\eta(x)), \text{ as } n \rightarrow \infty.$
2. $\mathbb{E}_{S_n}\left[\mathcal{E}(\widehat{f}_{k\text{-NN}}^{(n)})\right] \rightarrow \mathcal{E}_{k\text{-NN}}^* := \mathbb{E}[\alpha_k(\eta(X))], \text{ as } n \rightarrow \infty,$

where the superscript (n) is a reminder for the sample size used to construct $\widehat{f}_{k\text{-NN}}^{(n)}$, and

$$\alpha_k(\eta) := \eta Q_k(\eta) + (1 - \eta)(1 - Q_k(\eta)). \tag{4.1}$$

Before proving this theorem, we give a general intuition of why it holds (the proof will consist in making this intuition mathematically rigorous):

- As $n \rightarrow \infty$, the number of sample points in a given neighborhood of x will grow to infinity (provided that x is in the support of \mathbb{P}_X). Therefore, the distance of the k th-NN of x to x should tend to zero.
- Conditional to the sample points $(X_i)_{1 \leq i \leq n}$, the labels $Y^{(1)}(x), \dots, Y^{(k)}(x)$ of the k neighbors of x will be distributed as Bernoulli variables with respective parameters $\eta(X^{(1)}(x)), \dots, \eta(X^{(k)}(x))$. However, since the neighbors of x tend to x itself and η is continuous, these labels will essentially behave as i.i.d. $\text{Ber}(\eta(x))$ variables.

- The decision function $\widehat{f}_{k\text{-NN}}(x)$ will predict the majority class among the k neighbors. By the previous point, we expect the number of neighbors of class 1 to behave as $\text{Binom}(k, \eta(x))$ variable, so f will behave as a randomized decision predicting 0 with probability $Q_k(x)$ and 1 with probability $1 - Q_k(x)$, giving rise to the average error $\alpha_k(x)$ at point x .

We now proceed to the proof. The following lemmas will roughly speaking correspond to the successive points in the above intuition.

**

Lemma 4.2. *Let X_1, \dots, X_n, \dots be i.i.d. points from the distribution \mathbb{P}_X on a Polish space \mathcal{X} . For a given $x \in \mathcal{X}$ we denote $X_n^{(k)}(x)$ the k -th nearest neighbor of x among the points X_1, \dots, X_n . Then it holds:*

1. *For any fixed $x \in \text{Supp}(\mathbb{P}_X)$, and any fixed integer k , it holds $d(X_n^{(k)}, x) \rightarrow 0$ as $n \rightarrow \infty$, in probability and a.s.*
2. *Let $X \sim \mathbb{P}_X$, independent of the $(X_i)_{i \geq 1}$. Then it holds $d(X_n^{(k)}, X) \rightarrow 0$ as $n \rightarrow \infty$, in probability and a.s.*

Proof. Assume $x \in \text{Supp}(\mathbb{P}_X)$, by definition this means that for any $\varepsilon > 0$, $p(x, \varepsilon) := \mathbb{P}[X \in B(x, \varepsilon)] > 0$, where $B(x, \varepsilon)$ is the open ball of center x and radius ε .

Let us fix $\varepsilon > 0$. Denote $N_{\varepsilon, n}(x) := \#\{i \in \{1, \dots, n\} : X_i \in B(x, \varepsilon)\}$; by the law of large numbers, it holds

$$\frac{N_{\varepsilon, n}(x)}{n} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i \in B(x, \varepsilon)\} \rightarrow p(x, \varepsilon) > 0,$$

as $n \rightarrow \infty$, in probability. Furthermore, note the following implications:

$$d(X_n^{(k)}(x), x) \geq \varepsilon \Rightarrow X_n^{(k)}(x) \notin B(x, \varepsilon) \Rightarrow N_{\varepsilon, n} < k;$$

hence

$$\mathbb{P}[d(X_n^{(k)}(x), x) \geq \varepsilon] \leq \mathbb{P}\left[\frac{N_{\varepsilon, n}}{n} < \frac{k}{n}\right].$$

Since $\frac{N_{\varepsilon, n}}{n} \rightarrow p(x, \varepsilon) > 0$ in probability, but $\frac{k}{n} \rightarrow 0$, in particular for n big enough $\frac{k}{n} \leq p(x, \varepsilon)/2 < p(x, \varepsilon)$, and $\mathbb{P}\left[\frac{N_{\varepsilon, n}}{n} < \frac{k}{n}\right] \leq \mathbb{P}\left[\frac{N_{\varepsilon, n}}{n} < p(x, \varepsilon) - p(x, \varepsilon)/2\right] \rightarrow 0$ by definition of convergence in probability. We have proved that $d(X_n^{(k)}(x), x) \rightarrow 0$ in probability.

However, this implies also convergence a.s. by a monotonicity argument: observe that (for any fixed $\omega \in \Omega$ determining the sequence $(X_i(\omega))_{i \geq 1}$), $d(X_n^{(k)}(x), x)(\omega)$ is a decreasing sequence in n (adding more points to the set $\{X_1, \dots, X_n\}$ can only make the k -th neighbor of x possibly closer to x). It has therefore has a (nonnegative) limit $L(\omega)$ which can only be

0 (a.s.); indeed for any $t > 0$, by right-continuity $\lim_{n \rightarrow \infty} \mathbf{1}\{d(X_n^{(k)}(x), x) \geq t\} = \mathbf{1}\{L \geq t\}$, hence by dominated convergence

$$\mathbb{P}[L \geq t] = \lim_{n \rightarrow \infty} \mathbb{P}[d(X_n^{(k)}(x), x) \geq t] = 0,$$

so $L = 0$ a.s.

For the point 2, observe that for any $\varepsilon > 0$

$$\mathbb{P}_{X, (X_i)_{1 \leq i \leq n}} [d(X_n^{(k)}(X), X) > \varepsilon] = \mathbb{E}_X \left[\mathbb{P} \left[d(X_n^{(k)}(X), X) > \varepsilon \mid X \right] \right].$$

In the point 1. we have proved that $F_\varepsilon^{(n)}(x) := \mathbb{P} \left[d(X_n^{(k)}(X), X) > \varepsilon \mid X = x \right] = \mathbb{P} \left[d(X_n^{(k)}(x), x) > \varepsilon \right]$ converges to 0 for \mathbb{P}_X -almost all x , since in a Polish space it holds $\mathbb{P}[X \in \text{Supp}(\mathbb{P}_X)] = 1$. Therefore, since it is bounded by 1, its integral (the above expression) converges to 0 by dominated convergence. This means that $d(X_n^{(k)}(X), X) \rightarrow 0$ as $n \rightarrow \infty$, in probability, and implies convergence a.s. by the same monotonicity argument as in point 1. \square

The following lemma based on a coupling argument will allow to formalize that the labels of the close neighbors of a point x behave almost as if they were drawn with the Bernoulli parameter $\eta(x)$.

* **Lemma 4.3.** *Let k be an integer and Ψ an integrable function $\{0, 1\}^k \rightarrow [0, 1]$. Assume Y_1, \dots, Y_k are independent Bernoulli random variables with parameters η_1, \dots, η_k and Y'_1, \dots, Y'_k are i.i.d. Bernoulli random variables with parameters η . Then it holds*

$$|\mathbb{E}[\Psi(Y_1, \dots, Y_k)] - \mathbb{E}[\Psi(Y'_1, \dots, Y'_k)]| \leq \sum_{i=1}^k |\eta_i - \eta|.$$

(Note: equivalently the above is a bound on the total variation between the distribution of $(Y_i)_{1 \leq i \leq k}$ and that of $(Y'_i)_{1 \leq i \leq k}$).

Proof. Since the two expectations of $\Psi(Y_1, \dots, Y_k)$ and $\Psi(Y'_1, \dots, Y'_k)$ only depend on the distributions of the one and the other k -uple of random variables, it is possible to construct a *coupling* between these two k -uples. More precisely, we construct two k -uples of random variables $(\tilde{Y}_i)_{1 \leq i \leq k}$ and $(\tilde{Y}'_i)_{1 \leq i \leq k}$ so that $(\tilde{Y}_i)_{1 \leq i \leq k}$ has the same distribution as $(Y_i)_{1 \leq i \leq k}$, i.e. $\bigotimes_{i=1}^k \text{Ber}(\eta_i)$, and similarly for $(\tilde{Y}'_i)_{1 \leq i \leq k}$, but so that $\tilde{Y}_i = \tilde{Y}'_i$ as “often” as possible.

The construction is as follows: let U_1, \dots, U_k be i.i.d. $\text{Unif}[0, 1]$, and define $\tilde{Y}_i = \mathbf{1}\{U_i \leq \eta_i\}$; $\tilde{Y}'_i = \mathbf{1}\{U_i \leq \eta\}$. Then it can be checked easily that:

- $(\tilde{Y}_i)_{1 \leq i \leq k} \sim \bigotimes_{i=1}^k \text{Ber}(\eta_i)$;
- $(\tilde{Y}'_i)_{1 \leq i \leq k} \sim \bigotimes_{i=1}^k \text{Ber}(\eta)$;
- $\mathbb{P} \left[\tilde{Y}_i \neq \tilde{Y}'_i \right] = \mathbb{P} \left[U_i \in [\min(\eta, \eta_i), \max(\eta, \eta_i)] \right] = |\eta_i - \eta|.$

Therefore, since ψ takes values in $[0, 1]$:

$$\begin{aligned}
|\mathbb{E}[\Psi(Y_1, \dots, Y_k)] - \mathbb{E}[\Psi(Y'_1, \dots, Y'_k)]| &= \left| \mathbb{E}[\Psi(\tilde{Y}_1, \dots, \tilde{Y}_k)] - \mathbb{E}[\Psi(\tilde{Y}'_1, \dots, \tilde{Y}'_k)] \right| \\
&\leq \mathbb{E} \left[\left| \Psi(\tilde{Y}_1, \dots, \tilde{Y}_k) - \Psi(\tilde{Y}'_1, \dots, \tilde{Y}'_k) \right| \right] \\
&\leq \mathbb{P} \left[\exists i \in \{1, \dots, k\} : \tilde{Y}_i \neq \tilde{Y}'_i \right] \\
&\leq \sum_{i=1}^k \mathbb{P} \left[\tilde{Y}_i \neq \tilde{Y}'_i \right] \\
&\leq \sum_{i=1}^k |\eta - \eta_i|.
\end{aligned}$$

□

We can now assemble the previous results to prove the theorem:

Proof of Theorem 4.1. Write

$$\mathbb{E}_{S_n} \mathbb{E} \left[\ell(\hat{f}_{k\text{-NN}}^{(n)}(x), Y) \mid X = x \right] = \mathbb{E}_{X_1, \dots, X_n} \left[\mathbb{E} \left[\ell(\hat{f}_{k\text{-NN}}^{(n)}(x), Y) \mid X_1, \dots, X_n; X = x \right] \mid X = x \right], \quad (4.2)$$

where the internal conditional expectation is over (Y_1, \dots, Y_n, Y) . Since $\hat{f}_{k\text{-NN}}^{(n)}$ is constructed using only the values $(y^{(1)}(x), \dots, y^{(k)}(x))$ of the labels for the k nearest neighbors of x , we can write

$$\ell(\hat{f}_{k\text{-NN}}^{(n)}(x), y) =: \psi_{\mathbf{x}}(y^{(1)}(x), \dots, y^{(k)}(x), y).$$

Introduce the abbreviated notation $\mathbf{X}^{(n)} = (X_1, \dots, X_n, X)$, $\mathbf{x} = (x_1, \dots, x_n, x)$ and $\mathbf{Y}_{\mathbf{x}}^{(k)} := (Y^{(1)}(x), \dots, Y^{(k)}(x), Y)$. Conditionally to $\mathbf{X}^{(n)} = \mathbf{x}$, the tuple $\mathbf{Y}_{\mathbf{x}}^{(k)}$ has the distribution

$$\mathbb{P}_{\mathbf{Y}^{(k)} \mid \mathbf{X}^{(n)} = \mathbf{x}} := \left(\bigotimes_{i=1}^k \text{Ber}(\eta(x^{(i)}(x))) \right) \otimes \text{Ber}(\eta(x)).$$

On the other hand, it can be checked that $\alpha_k(x)$ defined by (4.1) is the expectation of $\psi_{\mathbf{x}}((\mathbf{Y}^{(k)})')$, for $(\mathbf{Y}^{(k)})' \sim \bigotimes_{i=1}^{k+1} \text{Ber}(\eta(x))$ (to see why, recall the third point in the general intuition discussion following Theorem 4.1). Applying Lemma 4.3, we obtain for any \mathbf{x} :

$$\begin{aligned}
|\mathbb{E}[\psi_{\mathbf{x}}(\mathbf{Y}^{(k)}) \mid \mathbf{X}^{(n)} = \mathbf{x}] - \alpha_k(x)| &= |\mathbb{E}_{\mathbf{Y}^{(k)} \mid \mathbf{X}^{(n)} = \mathbf{x}}[\psi_{\mathbf{x}}(\mathbf{Y}^{(k)})] - \mathbb{E}[\psi_{\mathbf{x}}((\mathbf{Y}^{(k)})')]| \\
&\leq \sum_{i=1}^k |\eta(x^{(i)}(x)) - \eta(x)|.
\end{aligned}$$

Now, assuming $x \in \text{Supp}(\mathbb{P}_X)$, by Lemma 4.2, and continuity of η , since $d(X^{(i)}(x), x) \rightarrow 0$ in probability for $i = 1, \dots, k$ as $n \rightarrow \infty$, and since η is bounded by 1, the expectation of the above right-hand side with respect to X_1, \dots, X_n converges to 0, which in view of (4.2) establishes the first part of the theorem. The second part follows immediately by integration over $X \sim \mathbb{P}_X$ and dominated convergence. □

We now examine closer the function α_k from (4.1) which determines the asymptotic error, for k fixed, of the k -NN method. We look at the cases $k = 1, 3, 5$ (note that it is reasonable to choose k odd to avoid ties). In each case, we examine the behavior of $\alpha(\eta)$ as η is close to 0, and compare it to the “local” optimal risk, which is $\min(\eta, 1 - \eta) = \eta$. By symmetry (since $\alpha_k(\eta) = \alpha_k(1 - \eta)$ by definition, for k odd), the same behavior holds as a function of $(1 - \eta)$ if η is close to 1.

This behavior is relevant in a situation where the classification problem is well-separable, i.e. $\eta(x)$ is always close to 0 or 1.

- $k = 1$: then $\alpha_1(\eta) = 2\eta(1 - \eta)$. Remember that the Bayes optimal error in classification is given by $\mathcal{E}^* = \mathbb{E}[\min(\eta(X), 1 - \eta(X))]$. If we denote $a(x) := \min(\eta(x), 1 - \eta(x))$, it therefore holds

$$\mathcal{E}_{1\text{-NN}}^* = \mathbb{E}[2a(X)(1 - a(X))] \leq 2\mathbb{E}[a(X)]\mathbb{E}[1 - a(X)] = 2\mathcal{E}^*(1 - \mathcal{E}^*),$$

where the above inequality is Jensen’s, since $x \mapsto x(1 - x)$ is concave on $[0, 1]$. We see that when \mathcal{E}^* is close to 0, the risk of 1-NN is bounded by twice the Bayes risk. This factor 2 is unavoidable, since in a well-separable situation, if η is close to 0 we have

$$\alpha_1(\eta) \sim 2\eta,$$

which is twice the (local) optimal risk.

- $k = 3$: then $Q_3(\eta) = (1 - \eta)^3 + 3(1 - \eta)^2\eta$, and $\alpha_3(\eta) = (1 - \eta)^3\eta + \eta^3(1 - \eta) + 6(1 - \eta)^2\eta^2$. So

$$\alpha_3(\eta) - \eta \sim 3\eta^2, \text{ as } \eta \rightarrow 0.$$

- $k = 5$: then with some additional tedious computations we get

$$\alpha_5(\eta) - \eta \sim 10\eta^3, \text{ as } \eta \rightarrow 0.$$

We see that in well-separable situations, when for all x , $\eta(x) \in [0, p] \cup [1 - p, 1]$ with p close to 0, we have $\mathcal{E}_{1\text{-NN}}^* \approx 2\mathcal{E}^*$ while $\mathcal{E}_{3\text{-NN}}^* \approx \mathcal{E}^*$. We see clearly the advantage of taking more neighbors in improving the asymptotic error. The 5-NN method asymptotically gives an even better approximation of the optimal Bayes error, still the 3-NN method may be sufficient in such a well-separable situation.

4.3 Consistency of the k -nearest-neighbors method

In this section we turn to analyzing the case where the number of neighbors $k(n)$ can grow with the sample size n . The main result is the following.

**

Theorem 4.4. *We consider a binary classification problem (with the usual 0-1 loss) under either of the following assumptions:*

(A) \mathcal{X} is a Polish space, and the function $\eta(x) = \mathbb{P}[Y = 1|X = x]$ is continuous.

(B) $\mathcal{X} = \mathbb{R}^d$ for some finite dimension d , endowed with the Euclidean distance, and there is no assumption on η .

Then, if $k(n)$ is such that $k(n) \rightarrow \infty$ and $k(n)/n \rightarrow 0$ and $n \rightarrow \infty$, it holds either under setting (A) or (B) that

$$\mathcal{E}(\widehat{f}_{k(n)\text{-NN}}^{(n)}) \rightarrow \mathcal{E}^*, \text{ in probability as } n \rightarrow \infty.$$

Notice in particular the remarkable fact that in setting (B) (\mathbb{R}^d endowed with the Euclidean distance), the k -NN method with suitably increasing $k(n)$ is **universally consistent**, i.e. its risk converges asymptotically towards the Bayes risk without any assumption whatsoever on the generating distribution \mathbb{P}_{XY} .

We start by revisiting Lemma 4.2.

**

Lemma 4.5. *Let X_1, \dots, X_n, \dots be i.i.d. points from the distribution \mathbb{P}_X on a Polish space \mathcal{X} . For a given $x \in \mathcal{X}$ we denote $X_n^{(k)}(x)$ the k -th nearest neighbor of x among the points X_1, \dots, X_n . Then if $k(n)/n \rightarrow 0$, as $n \rightarrow \infty$, it holds:*

1. *For any fixed $x \in \text{Supp}(\mathbb{P}_X)$, $d(X_n^{(k(n))}, x) \rightarrow 0$ as $n \rightarrow \infty$, in probability.*

2. *Let $X \sim \mathbb{P}_X$, independent of the $(X_i)_{i \geq 1}$. Then it holds $d(X_n^{(k(n))}, X) \rightarrow 0$ as $n \rightarrow \infty$, in probability.*

Proof. The proof is actually the same as for Lemma 4.2. To wit, for any $\varepsilon > 0$ and fixed k and n , we have seen that

$$\mathbb{P}[d(X_n^{(k)}(x), x) \geq \varepsilon] \leq \mathbb{P}\left[\frac{N_{\varepsilon, n}}{n} < \frac{k}{n}\right],$$

and furthermore $\frac{N_{\varepsilon, n}}{n} \rightarrow p(x, \varepsilon) > 0$ in probability as $n \rightarrow \infty$. So if $\frac{k(n)}{n} \rightarrow 0$ as $n \rightarrow \infty$, for n large enough it holds $\frac{k(n)}{n} < p(x, \varepsilon)/2$ implying that the right-hand side above converges to zero. The statement of point 2 is obtained by integration over x since the statement of point 1 is true for \mathbb{P}_X -almost all x , because $\mathbb{P}[X \in \text{Supp}(\mathbb{P}_X)] = 1$ in a Polish space. \square

Proof of Theorem 4.4. . Denote

$$\widehat{\eta}(x) := \frac{1}{k(n)} \sum_{i=1}^{k(n)} Y^{(i)}(x).$$

Observe that $\widehat{f}_{k\text{-NN}}$ can be seen as a plug-in classifier $\widehat{f}_{k\text{-NN}} := \mathbf{1}\{\widehat{\eta}(x) > \frac{1}{2}\}$. From the result on risk comparison for plug-in classifiers (Proposition 1.14), we know

$$\mathcal{E}_\ell(\widehat{f}_{k\text{-NN}}) - \mathcal{E}_\ell^* \leq 2\mathbb{E}_X[|\widehat{\eta}(X) - \eta(X)|].$$

Now define

$$\widetilde{\eta}(x) := \frac{1}{k(n)} \sum_{i=1}^{k(n)} \eta(X^{(i)}(x)),$$

and bound

$$\mathbb{E}_X[|\widehat{\eta}(X) - \eta(X)|] \leq \underbrace{\mathbb{E}_X[|\widehat{\eta}(X) - \widetilde{\eta}(X)|]}_{(II)} + \underbrace{\mathbb{E}_X[|\widetilde{\eta}(X) - \eta(X)|]}_{(I)}.$$

Estimating term (II). We have for any fixed x :

$$\widehat{\eta}(x) - \widetilde{\eta}(x) = \frac{1}{k(n)} \sum_{i=1}^n (Y^{(i)}(x) - \eta(X^{(i)}(x))).$$

Since $\mathbb{E}[Y^{(i)}(x)] = \eta(X^{(i)}(x))$ and conditionally to X_1, \dots, X_n , the labels $Y^{(1)}(x), \dots, Y^{(k)}(x)$ are independent Bernoulli variables with respective parameters $\eta(X^{(i)}(x))$, it holds

$$\mathbb{E}_{X, S_n} [(\widetilde{\eta}(x) - \widehat{\eta}(x))^2 | X_1, \dots, X_n; X = x] = \frac{1}{k(n)^2} \sum_{i=1}^{k(n)} \eta(X^{(i)}(x))(1 - \eta(X^{(i)}(x))) \leq \frac{1}{4k(n)}.$$

Furthermore, by integration

$$\mathbb{E}_{S_n}[(II)] = \mathbb{E}_{X, S_n}[|\widetilde{\eta}(x) - \widehat{\eta}(x)|] \leq \mathbb{E}_{X, S_n}[(\widetilde{\eta}(x) - \widehat{\eta}(x))^2]^{\frac{1}{2}} \leq \frac{1}{2\sqrt{k(n)}} \xrightarrow{n \rightarrow \infty} 0.$$

Hence term II tends to 0 in expectation over S_n and therefore also in probability, since it is nonnegative.

Estimating Term (I), setting (A). We have for any $x \in \mathcal{X}$:

$$\eta(x) - \widetilde{\eta}(x) = \frac{1}{k(n)} \sum_{i=1}^{k(n)} (\eta(x) - \eta(X^{(i)}(x))).$$

Since we assumed under setting (A) that η is continuous, for any fixed $\varepsilon > 0$ there exists $\delta > 0$ such that for any $x', d(x, x') < \delta$ implies $|\eta(x) - \eta(x')| \leq \varepsilon$. Therefore, since η and $\widetilde{\eta}$ are bounded by 1:

$$|\eta(x) - \widetilde{\eta}(x)| \leq \varepsilon + \mathbf{1}\{d(X^{(k(n))}, x) > \delta\};$$

hence

$$\mathbb{E}_{S_n}[(I)] = \mathbb{E}_{S_n, X}[|\eta(X) - \widetilde{\eta}(X)|] \leq \varepsilon + \mathbb{P}[d(X^{(k(n))}, X) > \delta],$$

and from Lemma 4.5 we know that the second term tends to 0. Since this holds for any $\varepsilon > 0$, the term I converges to 0 in expectation over S_n , and therefore also in probability, since it is nonnegative. This concludes the proof for the setting (A).

Estimating Term (I), setting (B). In this setting $\mathcal{X} = \mathbb{R}^d$, but η is not necessarily continuous. However, we know that the set $\mathcal{C}(\mathbb{R}^d)$ of continuous functions on \mathbb{R}^d is dense in $L^1(\mathbb{R}^d, \mathbb{P}_X)$. Let $\varepsilon > 0$ be fixed, and pick then η_ε continuous such that $\mathbb{E}[|\eta(X) - \eta_\varepsilon(X)|] \leq \varepsilon$. We define

$$\tilde{\eta}_\varepsilon(x) := \frac{1}{k(n)} \sum_{i=1}^{k(n)} \eta_\varepsilon(X^{(i)}(x)),$$

and write

$$(I) = \mathbb{E}_X[|\eta(X) - \tilde{\eta}(X)|] \leq \underbrace{\mathbb{E}_X[|\eta(X) - \eta_\varepsilon(X)|]}_{\leq \varepsilon} + \underbrace{\mathbb{E}_X[|\eta_\varepsilon(X) - \tilde{\eta}_\varepsilon(X)|]}_{(Ia)} + \underbrace{\mathbb{E}_X[|\tilde{\eta}_\varepsilon(X) - \tilde{\eta}(X)|]}_{(Ib)}.$$

The term (Ia) can be shown to converge in probability to zero, with exactly the same argument as used in the setting (A), since η_ε is continuous.

Finally, in order to estimate term (Ib), we will use a clever geometrical lemma which will be stated precisely below, which will prove the following:

$$\begin{aligned} \mathbb{E}_{S_n}[(Ib)] &= \mathbb{E}_{S_n, X}[|\tilde{\eta}_\varepsilon(X) - \tilde{\eta}(X)|] \\ &\leq \mathbb{E}_{S_n, X} \left[\frac{1}{k(n)} \sum_{i=1}^{k(n)} |\tilde{\eta}_\varepsilon(X^{(i)}(X)) - \tilde{\eta}(X^{(i)}(X))| \right] \\ &\leq \gamma_d \mathbb{E}[|\eta(X) - \eta_\varepsilon(X)|] \\ &\leq \gamma_d \varepsilon, \end{aligned} \tag{4.3}$$

where γ_d is a factor that only depends on d . Overall, since (Ia) converges to 0 in expectation over S_n , it also converges in probability, and the proof is done. \square

The following lemma is used to prove inequality (4.3):

* **Lemma 4.6** (Stone's lemma). *Let $\mathcal{X} = \mathbb{R}^d$ endowed with the Euclidean distance; X, X_1, \dots, X_n be i.i.d. variables of distribution \mathbb{P} , on \mathcal{X} , and $f \in L^1_+(\mathbb{R}^d, \mathbb{P})$ a nonnegative integrable function. Then there exists a factor γ_d , only depending on d , such that for any integer $k > 0$:*

$$\mathbb{E} \left[\sum_{i=1}^k f(X^{(i)}(X)) \right] \leq k \gamma_d \mathbb{E}[f(X)]. \tag{4.4}$$

Proof. Assume $k > 0$ is a fixed integer. Let us denote $X_0 = X$, and for $i = 0, \dots, n$, introduce the notation

$$\text{NN}_k(X_i) := \{j \neq i : X_j \text{ is one of the } k \text{ nearest neighbors of } X_i \text{ among } X_0, \dots, X_n\}.$$

Then we have

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^k f(X^{(i)}(X_0)) \right] &= \mathbb{E} \left[\sum_{i=1}^n f(X_i) \mathbf{1}\{i \in \text{NN}_k(X_0)\} \right] \\ &= \sum_{i=1}^n \mathbb{E}[f(X_i) \mathbf{1}\{i \in \text{NN}_k(X_0)\}] \\ &= \sum_{i=1}^n \mathbb{E}[f(X_0) \mathbf{1}\{0 \in \text{NN}_k(X_i)\}] \quad (*) \\ &= \mathbb{E} \left[f(X_0) \sum_{i=1}^n \mathbf{1}\{0 \in \text{NN}_k(X_i)\} \right] \\ &= \mathbb{E}[f(X_0) \#\{i : 0 \in \text{NN}_k(X_i)\}]. \quad (**) \end{aligned}$$

Observe that the clever step (*) is obtained by symmetry: in each separate expectation we can exchange the role of X_0 and X_i , while the distribution of the $(n+1)$ -tuple (X_0, \dots, X_n) remains unchanged. Finally, the next lemma will establish that

$$\#\{i : 0 \in \text{NN}_k(X_i)\} \leq k\gamma_d,$$

for a factor γ_d only depending on d ; this will conclude the proof, since we can plug this upper bound into (**) (observe that it is at that point that we must make use of the fact that f is nonnegative). \square

Lemma 4.7. *Let (x_0, \dots, x_n) be $(n+1)$ points in \mathbb{R}^d and*

$$\text{NN}_k(x_i) := \{j \neq i : x_j \text{ is one of the } k \text{ nearest neighbors of } x_i \text{ among } x_0, \dots, x_n\},$$

where ties are broken by taking the smallest index. Then

$$\#\{i : 0 \in \text{NN}_k(x_i)\} \leq k\gamma_d.$$

with γ_d a factor only depending on d .

Proof. Without loss of generality we assume $x_0 = \mathbf{0}$. Let us consider a fixed open cone \mathcal{C}_0 starting from the origin (x_0) and of angle $2\theta \leq \pi/3$. For any two points y and z in \mathcal{C}_0 , assume that $\|y\| \leq \|z\|$, then since the angle between y and z is strictly less than 2θ :

$$\|y - z\|^2 < \|y\|^2 + \|z\|^2 - 2\|y\|\|z\| \underbrace{\cos(2\theta)}_{\geq 1/2} \leq \|z\|^2 \left(1 + \underbrace{\frac{\|y\|^2}{\|z\|^2} - \frac{\|y\|}{\|z\|}}_{\leq 1} \right) \leq \|z\|^2. \quad (4.5)$$

Let x_{i_1}, \dots, x_{i_k} be the elements of $\{x_1, \dots, x_n\}$ belonging to \mathcal{C}_0 and closest to the origin x_0 (if there are only $k' < k$ such elements, take only those; if there are more because of ties, take the ones with the k smallest indices). Now, notice that for any other $x_j \in \mathcal{C}_0$ with $j \notin \{i_1, \dots, i_k\}$, we have $\|x_j\| \geq \|x_{i_\ell}\|$ for all $\ell = 1, \dots, k$, thus by (4.5) we have

$$\|x_j - x_0\|^2 = \|x_j\|^2 > \sup_{\ell=1, \dots, k} \|x_j - x_{i_\ell}\|^2.$$

Therefore, for any such x_j , the point x_0 is *not* among the k nearest neighbors of x_j , and we must have $0 \notin NN_k(x_j)$.

To summarize: in any such open cone \mathcal{C}_0 , there are *at most* k indices i_ℓ such that $0 \in NN_k(x_{i_\ell})$. Now the space \mathbb{R}^d can be covered by a finite number γ_d of such open cones (note that it is enough to cover the unit ball by homogeneity, and then used compactness). Overall there are *at most* $k\gamma_d$ indices i_ℓ such that $0 \in NN_k(x_{i_\ell})$. This implies the conclusion. \square

Exercise 4.1. Prove the bound $\gamma_d \leq \left(1 + \frac{1}{\sin(\pi/12)}\right)^d \leq 5^d$. For this, assume that the unit ball is covered by open cones of angle $\pi/3$, with principal axes given by the direction of vectors x_1, \dots, x_M , with $\|x_i\| = 1$. Furthermore, assume that this covering is of minimal cardinality. Prove then that it must hold $\|x_i - x_j\| \geq 2 \sin \frac{\pi}{12} =: r$ for $i \neq j$. Conclude by a volume argument: the balls $B(x_i, \frac{r}{2})$ must be disjoint and are all contained in $B(0, 1 + \frac{r}{2})$, entailing that the sum of their volumes is less than the volume of this containing ball. Conclude.

Exercise 4.2. It is possible to also prove a.s. convergence in Lemma 4.5 as in Lemma 4.2, but the argument of Lemma 4.2 has to be modified since a.s. monotonicity of $d(X_n^{(k(n))})(x), x)$ does not necessary hold when $k(n)$ depends on n .

Establish the a.s. convergence property in Lemma 4.5 by considering $U_n := \sup_{m \geq n} d(X_m^{(k(m))})(x), x)$ and establishing that $U_n \rightarrow 0$ in probability. Then use the monotonicity argument since U_n is now a.s. decreasing in n .

5 Reproducing kernel methods

5.1 Motivation

Linear methods after a feature mapping. In the chapter on linear classification methods, a central role was played by linear (or affine) score functions which were linear forms $f_w(x) = \langle x, w \rangle$; such linear forms are also the class of predictors considered for linear regression. Now imagine we would like to consider as prediction (or score) functions the class of functions of the form

$$\mathcal{F} := \left\{ f_\alpha(x) := \sum_{i=1}^M \alpha_i f_i(x), \alpha = (\alpha_1, \dots, \alpha_M) \in \mathbb{R}^M \right\},$$

where $\{f_1, \dots, f_M\}$ is a known, fixed finite set of real-valued functions $\mathcal{X} \rightarrow \mathbb{R}$. For instance, we could consider that f_1, \dots, f_M are monomials in the coordinates of x of degree up to m , so that \mathcal{F} is the set of polynomials in x of degree up to m . Or the f_i could be trigonometric functions; or in fact any finite “library” of functions that we consider relevant for the problem at hand. (Note in particular that \mathcal{X} does not have to be a subset of \mathbb{R}^d ; here \mathcal{X} could be something like a sequence of characters, a graph. . . .)

In this setting each function f_i is called a fixed “feature”. While the above setting might seem much more general than linear functions, we can subsume it into linear methods by considering the “feature mapping”

$$\Phi(x) : \mathcal{X} \rightarrow \mathbb{R}^M, \quad x \mapsto (f_1(x), \dots, f_M(x)),$$

if we use the shorthand notation $\tilde{x} := \Phi(x)$, then we observe that functions in \mathcal{F} which are nonlinear in x are linear functions of \tilde{x} :

$$f_\alpha(x) := \sum_{i=1}^M \alpha_i f_i(x) = \langle \alpha, \Phi(x) \rangle = \langle \alpha, \tilde{x} \rangle. \quad (5.1)$$

Therefore, we can in principle apply any linear learning method (regression, or one of the linear classification methods seen in Section 2) to the modified input data \tilde{x} in order to learn a prediction function in the class \mathcal{F} .

An important point to notice right away is that the feature mapping Φ can often be high-dimensional (as exemplified by polynomial regression: if $\mathcal{X} = \mathbb{R}^d$, the vector space of polynomial functions of degree up to m has dimension $(m+1)^d$). In fact it can commonly be the case that we would like to consider a “feature space” (the image space of Φ) of dimensionality M larger than the data sample size n . This has two important consequences:

1. It is essential to consider *regularized* methods (see in particular Section 2.7), otherwise we are certain to run into overfitting.
2. It can computationally inconvenient to store the data as its explicit feature mapping $(\tilde{x}_1, \dots, \tilde{x}_n) = (\phi(x_1), \dots, \phi(x_n))$, both in terms of computation time of this mapping, and of memory size.

Scalar products are sufficient. Concerning the second point above, an important point to remark is that all linear methods (possibly in regularized form) we have seen have the following property:

1. The learnt linear function (or functions, in the case of multi-class classification) $f_{\hat{w}}$ has a parameter which can be written (regardless of the dimension) as a linear combination of the input data:

$$\hat{w} = \sum_{i=1}^n \beta_i X_i; \quad (5.2)$$

2. In order to compute the coefficients β_i above, it is sufficient to know the scalar products $\langle X_i, X_j \rangle$, $i = 1, \dots, n$, along with the labels Y_1, \dots, Y_n .
3. In order to compute $f_{\hat{w}}(x)$ for a new test point x , given the coefficients $(\beta_1, \dots, \beta_n)$ of the representation (5.2), it is sufficient to know the scalar products $\langle X_i, x \rangle$, $i = 1, \dots, n$.

The last point is obvious since given (5.2) we have

$$f_{\hat{w}}(x) = \langle \hat{w}, x \rangle = \sum_{i=1}^n \beta_i \langle X_i, x \rangle. \quad (5.3)$$

For the two first points, we take first the example of the perceptron. Remember that for the perceptron algorithm (see Section 2.6), the main iteration is $\hat{w}_k = \hat{w}_{k-1} + X_{i_k} Y_{i_k}$. By recursion, assume \hat{w}_{k-1} satisfies points 1-2 above, that is to say, is of the form (5.2), with coefficients $\beta_i^{(k)}$ which can be computed only given the scalar products $\langle X_i, X_j \rangle$. Observe that the determination of the index i_k depends on finding which training examples are correctly classified or not by the classifier $\text{sign}(f_{\hat{w}_{k-1}})$, for which we only need to know the scalar products $\langle X_i, X_j \rangle$, by (5.3). As a consequence, \hat{w}_k also satisfies points 1-2 above.

Let us take ridge regression as the next example. Remember from (2.16) and below that this method outputs (for a fixed regularization parameter $\lambda > 0$) the linear predictor with parameter vector

$$\hat{w}_\lambda = (\mathbf{X}^t \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^t \mathbf{Y}, \quad (5.4)$$

where \mathbf{X} is the (n, d) matrix whose rows are x_1^t, \dots, x_n^t , and $\mathbf{Y} = (y_1, \dots, y_n)^t$. We have the following lemma:

* **Lemma 5.1.** *It holds for $\lambda > 0$:*

$$(\mathbf{X}^t \mathbf{X} + \lambda I_d)^{-1} \mathbf{X}^t = \mathbf{X}^t (\mathbf{X} \mathbf{X}^t + \lambda I_n)^{-1}.$$

As a consequence, the formula (5.4) implies that \hat{w}_λ is of the form (5.2), with

$$(\beta_1, \dots, \beta_n) = (\mathbf{X} \mathbf{X}^t + \lambda I_n)^{-1} \mathbf{Y}; \quad (5.5)$$

observe that $(\mathbf{X}\mathbf{X}^t)_{ij} = \langle X_i, X_j \rangle$, so finally points 1-2 hold in this setting too.

As apparent above, an object of central importance is $\mathbf{X}\mathbf{X}^t$, the *Gram matrix* associated with points (X_1, \dots, X_n) . Note that it is more economical to use the (n, n) Gram matrix than the (d, d) matrix $\mathbf{X}^t\mathbf{X}$ if $d < n$.

The conclusion of these observations is that it is enough to know how to compute scalar products $\langle x, x' \rangle$ in order to learn and apply prediction functions for classical linear method. If we combine this observation with the idea of a feature mapping exposed earlier, we see that we do not need to know the explicit feature mapping Φ : we only need to be able to compute $\langle \Phi(x), \Phi(x') \rangle$ for arbitrary $x, x' \in \mathcal{X}$.

This is in essence the principles underlying the construction of kernel methods, to summarize:

1. We can greatly extend the flexibility of linear methods by applying them after a *feature mapping* Φ in a possibly high-dimensional Euclidean vector space.
2. For standard methods, we don't need to know the feature mapping Φ explicitly, but only need to be able to compute scalar products $\langle \Phi(x), \Phi(x') \rangle$ for any $x, x' \in \mathcal{X}$.
3. It is important to always consider regularized versions of linear methods in this context, since the output space of the feature mapping Φ is generally high-dimensional.

Exercise 5.1. Prove Lemma 5.1 and justify (5.5).

5.2 Reproducing kernel Hilbert spaces

The previous considerations motivate the interest of being able to compute easily

$$k(x, x') := \langle \Phi(x), \Phi(x') \rangle \quad (5.6)$$

rather than explicitly computing $\Phi(x), \Phi(x')$. We will call such a function k a *kernel*. Since we only need to know the function k in order to apply various algorithms on the feature space, it is natural to ask the reciprocal question: under what conditions is a given function $k : \mathcal{X} \times \mathcal{X}$ the kernel associated to a feature mapping; i.e. when can we guarantee that there exists some feature space and a feature mapping Φ such that (5.6) holds?

Example. Let $\mathcal{X} = \mathbb{R}^d$ and $k(x, x') = (\langle x, x' \rangle + c)^2$, where $c \geq 0$ is a constant. Is it a kernel in the above sense? We have

$$\begin{aligned} (\langle x, x' \rangle + c)^2 &= \left(\sum_{i=1}^d (x_i x'_i) \right)^2 + 2c \sum_{i=1}^d x_i x'_i + c^2 \\ &= \sum_{i,j=1}^d (x_i x_j)(x'_i x'_j) + \sum_{i=1}^d (\sqrt{2cx_i})(\sqrt{2cx'_i}) + c^2, \end{aligned}$$

so (5.6) is satisfied with

$$\Phi(x) := \left[(x_i x_j)_{1 \leq i, j \leq d}; (\sqrt{2cx}); c \right].$$

Observe that Φ maps to all monomials of degree up to 2, so the associated function space defined via 5.1 is the vector space of polynomials of degree up to 2 in the coordinates of x . For this reason k is called *polynomial kernel* of order 2.

The following fundamental theorem gives a set of necessary and sufficient conditions on k in order for (5.6) to hold.

**

Theorem 5.2 (Characterization theorem). *Let \mathcal{X} be a nonempty set, and $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ a function.*

Then there exist a Hilbert space \mathcal{H}_\circ and a mapping $\Phi_\circ : \mathcal{X} \rightarrow \mathcal{H}_\circ$ with

$$k(x, x') = \langle \Phi_\circ(x), \Phi_\circ(x') \rangle_{\mathcal{H}_\circ} \quad (5.7)$$

if and only if the following conditions are satisfied:

1. *k is symmetric: $k(x, x') = k(x', x)$ for all $x, x' \in \mathcal{X}$.*
2. *k has positive type, that is, for any integer $n > 0$, and any n -uples $(x_1, \dots, x_n) \in \mathcal{X}^n$, and $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$, it holds*

$$\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

Note: *the above properties can be more compactly expressed equivalently as: for any integer $n > 0$, and any n -uple $(x_1, \dots, x_n) \in \mathcal{X}^n$, the matrix K given by $K_{ij} = k(x_i, x_j)$ is symmetric positive semi-definite. For this reason, we will call a kernel satisfying the above conditions a symmetric positive semi-definite (spsd) kernel.*

Proof. “Only if” direction: assume (5.7) holds. Then obviously k is symmetric, and

$$\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) = \sum_{i,j=1}^n \alpha_i \alpha_j \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}} = \left\| \sum_{i=1}^n \alpha_i \Phi(x_i) \right\|^2 \geq 0.$$

“If” direction: assume k is a spsd kernel. We need to construct \mathcal{H} and Φ . For any $x \in \mathcal{X}$, denote the real-valued function $k_x : \mathcal{X} \rightarrow \mathbb{R}, y \mapsto k(x, y)$ (we will alternatively use the notation $k_x := k(x, \cdot)$). Now define

$$\mathcal{H}_{\text{pre}} := \text{Span}\{k_x, x \in \mathcal{X}\}; \quad (5.8)$$

we stress that the above set is made of *finite* linear combinations of functions of the form $k(x_i, \cdot)$.

We define a bilinear form $[\cdot, \cdot]$ on \mathcal{H}_{pre} :

$$\text{for } f = \sum_{i \in I_1} \lambda_i k_{x_i}; \quad g = \sum_{j \in I_2} \mu_j k_{x_j} \quad \text{define } [f, g] := \sum_{\substack{i \in I_1 \\ j \in I_2}} \lambda_i \mu_j k(x_i, x_j). \quad (5.9)$$

We need to stress that this is a well-formed definition: indeed, it may be possible that the same function has another representation as a linear expansion, say $f = \sum_{i \in I'_1} \lambda'_i k_{x'_i}$. But it holds that $\sum_{\substack{i \in I_1 \\ j \in I_2}} \lambda_i \mu_j k(x_i, x_j) = \sum_{j \in I_2} \mu_j f(x_j)$ by definition, so the definition of $[\cdot, \cdot]$ does not depend on the particular representation of f . The same argument applies to g .

Now, it is easy to check that the property 1. (k symmetric) implies that $[\cdot, \cdot]$ is symmetric, and that property 2. (k has positive type) implies that for $f \in \mathcal{H}_{\text{pre}}$ having the representation as in (5.9), it holds $[f, f] = \sum_{i, j \in I_1} \lambda_i \lambda_j k(x_i, x_j) \geq 0$. Hence $[\cdot, \cdot]$ is a symmetric positive semidefinite form on the vector space \mathcal{H}_{pre} .

We finally check that it is definite. A symmetric positive semidefinite form satisfies the Cauchy-Schwarz inequality, so it holds (assuming again $f \in \mathcal{H}_{\text{pre}}$ with representation as in (5.9))

$$f(x) = \sum_{i \in I_1} \lambda_i k(x_i, x) = [f, k_x] \leq [f, f]^{\frac{1}{2}} [k_x, k_x]^{\frac{1}{2}}, \quad (5.10)$$

hence $[f, f] = 0$ implies that $f(x) = 0$ for all x , i.e. $f = 0$ as a function. Hence $[\cdot, \cdot]$ is a symmetric definite positive bilinear form on \mathcal{H}_{pre} .

Finally, define $\Phi_{\text{pre}}(x) : \mathcal{X} \rightarrow \mathcal{H}_{\text{pre}}$, $x \mapsto k_x$. Then it holds

$$[\Phi_{\text{pre}}(x), \Phi_{\text{pre}}(x')] = [k_x, k_{x'}] = k(x, x'). \quad (5.11)$$

We have just constructed a pre-Hilbert space \mathcal{H}_{pre} and a mapping Φ_{pre} such that (5.7) holds.

What is missing for a proper Hilbert space is completeness. But this space can be completed: there exists a complete Hilbert space \mathcal{H}_{\circ} , and an isometry $i: (\mathcal{H}_{\text{pre}}, [\cdot, \cdot]) \xrightarrow{i} (\mathcal{H}_{\circ}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ such that $i(\mathcal{H}_{\text{pre}})$ is dense in \mathcal{H}_{\circ} . (The completion operation is obtained by considering equivalence classes of Cauchy sequences in \mathcal{H}_{pre} ; this is a standard construction that we don't detail here.)

Correspondingly we can define $\Phi_{\circ}(x) := i \circ \Phi_{\text{pre}}(x)$, which satisfies (5.7) because of (5.11), since i is an isometry. This concludes the proof. \square

In the previous proof, \mathcal{H}_{pre} was specifically constructed as a pre-Hilbert space of real-valued functions on \mathcal{X} with $\Phi_{\text{pre}}(x) = k_x$. It is an important point that this property in fact carries over to its completion, and we highlight this in the next result.

**

Theorem 5.3 (and definition). *If k is a spsd kernel on the set \mathcal{X} (as in Theorem 5.2), the Hilbert space \mathcal{H} and the mapping Φ satisfying (5.7) can be constructed so that:*

1. \mathcal{H} is a vector space of real-valued functions \mathcal{X} ;
2. The mapping $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ is given by $\Phi : x \mapsto k_x$;
3. The following reproducing property is satisfied:

$$\forall f \in \mathcal{H} : \quad f(x) = \langle k_x, f \rangle_{\mathcal{H}}. \quad (5.12)$$

The space \mathcal{H} satisfying the above properties is unique and is called reproducing kernel Hilbert space on \mathcal{X} with kernel k .

Furthermore, \mathcal{H}_{pre} given by (5.8) is dense in \mathcal{H} .

Proof. We have constructed in the proof of Theorem 5.2 a pre-Hilbert space \mathcal{H}_{pre} satisfying the announced properties (observe that the reproducing property (5.12) holds in \mathcal{H}_{pre} by construction/definition of the form $[\cdot, \cdot]$). What about its completion \mathcal{H}_o ? We recall that there exists an isometry $i : \mathcal{H}_{\text{pre}} \rightarrow \mathcal{H}_o$ with $i(\mathcal{H}_{\text{pre}})$ dense in \mathcal{H}_o . We now construct the following mapping

$$\xi : \mathcal{H}_o \rightarrow \mathcal{F}(\mathcal{X}, \mathbb{R}) : \quad h \mapsto \xi(h) := (x \in \mathcal{X} \mapsto \langle i(k_x), h \rangle_{\mathcal{H}_o}).$$

Let us prove that ξ is injective. Assume $\xi(h) = 0$, which is to say, for all $x \in \mathcal{X}$ it holds $\langle i(k_x), h \rangle_{\mathcal{H}_o} = 0$. This implies by linearity that for any $f \in \mathcal{H}_{\text{pre}}$, $\langle i(f), h \rangle_{\mathcal{H}_o} = 0$. But since $i(\mathcal{H}_{\text{pre}})$ is dense in \mathcal{H}_o , it implies that for any $h' \in \mathcal{H}_o$, $\langle h', h \rangle_{\mathcal{H}_o} = 0$, hence $h = 0$.

Since ξ is linear, it defines a bijection between \mathcal{H}_o and $\xi(\mathcal{H}_o) \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$. We can therefore endow $\mathcal{H} = \xi(\mathcal{H}_o)$ with the scalar product $\langle f, f' \rangle := \langle \xi^{-1}(f), \xi^{-1}(f') \rangle$, so that \mathcal{H} is a Hilbert space of functions $\mathcal{X} \rightarrow \mathbb{R}$ which is isometric to \mathcal{H}_o .

Additionally, we observe that $\mathcal{H}_{\text{pre}} \subseteq \mathcal{H}$ since $\xi \circ i$ coincides with the identity: for any $x \in \mathcal{X}$, it holds

$$\xi(i(k_x)) = (y \mapsto \langle i(k_y), i(k_x) \rangle_{\mathcal{H}_o} = \langle k_y, k_x \rangle_{\mathcal{H}_{\text{pre}}} = k(x, y)) = k_x, \quad (5.13)$$

hence by linearity of $\xi \circ i$, we have $\mathcal{H}_{\text{pre}} \xrightarrow{\xi \circ i} \mathcal{H}$, which is an inclusion of Hilbert spaces since $\xi \circ i$ is an isometry by composition of isometries.

We can therefore define the feature mapping $\Phi(x) = k_x$, which satisfies

$$\langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = \langle k_x, k_{x'} \rangle_{\mathcal{H}} = \langle k_x, k_{x'} \rangle_{\mathcal{H}_{\text{pre}}} = k(x, x'),$$

by the above isometric inclusion.

Finally, we check the reproducing property: for any $f \in \mathcal{H} = \xi(\mathcal{H}_o)$ there exists $h \in \mathcal{H}_o$ with $f = \xi(h)$, hence $f = (x \mapsto \langle h, i(k_x) \rangle_{\mathcal{H}_o})$ and for any $x \in \mathcal{X}$:

$$f(x) = \langle h, i(k_x) \rangle_{\mathcal{H}_o} \text{ while } \langle f, k_x \rangle_{\mathcal{H}} = \langle \xi(h), \xi(i(k_x)) \rangle_{\mathcal{H}} = \langle h, i(k_x) \rangle_{\mathcal{H}_o},$$

hence (5.12) is satisfied.

We turn to unicity. Let \mathcal{H}' be another Hilbert space of real functions on \mathcal{X} satisfying the announced properties. By property 2. it holds that $k_x \in \mathcal{H}'$ for all $x \in \mathcal{X}$, and by consequence $\mathcal{H}_{\text{pre}} \subseteq \mathcal{H}'$. Furthermore property 3. implies that $\langle k_x, k_{x'} \rangle_{\mathcal{H}} = k(x, x') = [k_x, k'_{x'}]$, where $[\cdot, \cdot]$ is the bilinear form constructed on \mathcal{H}_{pre} in the proof of Theorem 5.2. By linearity $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ coincides with $[\cdot, \cdot]$ on \mathcal{H}_{pre} . Hence the identity mapping $\mathcal{H}_{\text{pre}} \hookrightarrow \mathcal{H}'$ is an isometry.

On the other hand we have established that $\mathcal{H}_{\text{pre}} \subseteq \mathcal{H}$ via the isometric inclusion $\xi \circ i$. Let $\overline{\mathcal{H}_{\text{pre}}}$ be the closure of \mathcal{H}_{pre} in \mathcal{H} . It can be checked that $\overline{\mathcal{H}_{\text{pre}}} = \mathcal{H}$, where \mathcal{H} was constructed above. Indeed, we know that $i(\mathcal{H}_{\text{pre}})$ is dense in \mathcal{H}_\circ , hence by isometry $\xi \circ i(\mathcal{H}_{\text{pre}}) = \mathcal{H}_{\text{pre}}$ is dense in $\xi(\mathcal{H}_\circ) = \mathcal{H}$.

Finally, observe that the closure of \mathcal{H}_{pre} in \mathcal{H} coincides with the closure of \mathcal{H}_{pre} in \mathcal{H}' . Indeed, any Cauchy sequence of functions $(f_n)_{n \geq 1}$ in \mathcal{H}_{pre} converges both in \mathcal{H} and \mathcal{H}' by completeness of both these spaces, and the limit point f is uniquely determined as a function, since for any $x \in \mathcal{X}$, $f(x) = \lim_{n \rightarrow \infty} [k_x, f_n]$, the right-hand side of the latter equality being a real-valued Cauchy sequence (by continuity) hence having a unique limit. So any such limit function f belongs to both \mathcal{H} and \mathcal{H}' , and since $\mathcal{H} = \overline{\mathcal{H}_{\text{pre}}}$, we have $\mathcal{H} \subseteq \mathcal{H}'$.

Since \mathcal{H} is closed in \mathcal{H}' we can write $\mathcal{H}' = \mathcal{H} \oplus \mathcal{H}_1$; but for any $f_1 \in \mathcal{H}_1$ we have since $\mathcal{H}_1 \perp \mathcal{H}$ that for any $h \in \mathcal{H}$, $\langle f_1, h \rangle_{\mathcal{H}'} = 0$. In particular, for any $x \in \mathcal{X}$, $k_x \in \mathcal{H}$ and $\langle f_1, k_x \rangle_{\mathcal{H}'} = f_1(x) = 0$ (by the assumed reproducing property on \mathcal{H}'), so $f_1 = 0$. Finally $\mathcal{H}' = \mathcal{H}$, proving unicity. \square

For a complete overview we also mention the following characterization of reproducing Hilbert kernel spaces.

**

Theorem 5.4. *Let \mathcal{H} be a Hilbert space of real-valued functions over a set \mathcal{X} . Then the following properties are equivalent:*

1. *For all $x \in \mathcal{X}$, the evaluation function*

$$\delta_x : \mathcal{H} \rightarrow \mathbb{R}; \quad f \mapsto f(x) \tag{5.14}$$

is continuous.

2. *There exists a (unique) function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that:*

(a) *for all $x \in \mathcal{X}$: $k(x, \cdot) \in \mathcal{H}$.*

(b) *for all $x \in \mathcal{X}$ and $f \in \mathcal{H}$: $\langle f, k(x, \cdot) \rangle = f(x)$.*

Furthermore, the function k in the last point is a spsd kernel, so that \mathcal{H} is the reproducing kernel Hilbert space with kernel k .

Proof. (1) \Rightarrow (2): since δ_x is continuous, by Riesz' theorem there exists a unique element $\zeta_x \in \mathcal{H}$ such that $\delta_x(f) = \langle \zeta_x, f \rangle$ for all $f \in \mathcal{H}$. Since \mathcal{H} is a space of real-valued functions, we define $k(x, y) = \zeta_x(y)$ for all x, y . Then $k(x, \cdot) = \zeta_x$, so that the announced properties (a) and (b) are satisfied.

(2) \Rightarrow (1): we have for any $f \in \mathcal{H}$ and $x \in \mathcal{X}$, that $\delta_x(f) = f(x) = \langle f, k_x \rangle$ which is continuous by continuity of the scalar product (which can be seen as a consequence of the Cauchy-Schwarz inequality).

The kernel k appearing in (2) is spsd: it holds $k(x, y) = k_x(y) = \langle k_x, k_y \rangle = k_y(x) = k(y, x)$, so k is symmetric. Furthermore

$$\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \sum_{i,j} \alpha_i \alpha_j \langle k_{x_i}, k_{x_j} \rangle = \left\| \sum_i \alpha_i k_{x_i} \right\|^2 \geq 0.$$

□

5.3 Construction of spsd kernels

Theorem 5.2 gave us a characterization of kernels that are scalar products of mappings of points of \mathcal{X} through a feature mapping Φ to an underlying feature space \mathcal{H} : these are exactly the spsd kernels. But it is not obvious to check if a given function is spsd. Instead, the following result will tell us how to construct many such kernels.

**

Theorem 5.5. *Let \mathcal{X} be a nonempty set.*

- (i) *If f is a real-valued function on \mathcal{X} , then $k(x, y) := f(x)f(y)$ is a spsd kernel on \mathcal{X} .*
- (ii) *If \mathcal{X} is a Euclidean or Hilbert space with inner product $\langle \cdot, \cdot \rangle$, then $k(x, y) = \langle x, y \rangle$ is a spsd kernel. (“Linear kernel”)*
- (iii) *If k_1 is a spsd kernel on \mathcal{X} and $c \geq 0$ is a real, then $k = ck_1$ is a spsd kernel.*
- (iv) *If k_1, k_2 are spsd kernels on \mathcal{X} , then $k = k_1 + k_2$ is a spsd kernel.*
- (v) *If k_1, k_2 are spsd kernels on \mathcal{X} , then $k = k_1 k_2$ is a spsd kernel.*
- (vi) *If \mathcal{X}' is another set, k_3 a spsd kernel on \mathcal{X}' and $F : \mathcal{X} \rightarrow \mathcal{X}'$ a mapping, then $k(x, y) := k_3(F(x), F(y))$ is a spsd kernel on \mathcal{X} .*
- (vii) *If $(k_i)_{i \geq 1}$ is a sequence of kernels on \mathcal{X} which converge pointwise (for any $x, y \in \mathcal{X}$) then the limiting function is a spsd kernel.*

The proof for all points of the above theorem are left as an exercise, with the exception of point (v), for which we provide the following lemma:

Lemma 5.6. *Let M, N be two (n, n) spsd real matrices. Then the (n, n) matrix A defined by $A_{ij} := M_{ij}N_{ij}$ is spsd.*

Proof. Obviously A is symmetric. Let $(e_k)_{1 \leq k \leq n}$ be a diagonalizing orthonormal basis of M corresponding to nonnegative eigenvalues $(\lambda_k)_{1 \leq k \leq n}$. Thus $M = \sum_{k=1}^n \lambda_k e_k e_k^t$, and $M_{ij} = \sum_{k=1}^n \lambda_k e_k^{(i)} e_k^{(j)}$. Similarly, let $(f_\ell)_{1 \leq \ell \leq n}$ be a diagonalizing basis of N corresponding to nonnegative eigenvalues $(\mu_\ell)_{1 \leq \ell \leq n}$. Thus $N_{ij} = \sum_{\ell=1}^n \mu_\ell f_\ell^{(i)} f_\ell^{(j)}$. For any $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$:

$$\begin{aligned} \sum_{i,j=1}^n \alpha_i \alpha_j A_{ij} &= \sum_{i,j=1}^n \alpha_i \alpha_j \left(\sum_{k=1}^n \lambda_k e_k^{(i)} e_k^{(j)} \right) \left(\sum_{\ell=1}^n \mu_\ell f_\ell^{(i)} f_\ell^{(j)} \right) \\ &= \sum_{i,j,k,l=1}^n \alpha_i \alpha_j \lambda_k \mu_\ell e_k^{(i)} e_k^{(j)} f_\ell^{(i)} f_\ell^{(j)} \\ &= \sum_{k,l=1}^n \lambda_k \mu_\ell \sum_{i,j=1}^n (\alpha_i e_k^{(i)} f_\ell^{(i)}) (\alpha_j e_k^{(j)} f_\ell^{(j)}) \\ &= \sum_{k,l=1}^n \lambda_k \mu_\ell \left(\sum_{i=1}^n \alpha_i e_k^{(i)} f_\ell^{(i)} \right)^2 \geq 0. \end{aligned}$$

Therefore A is spsd. □

We deduce the following corollaries from Theorem 5.5:

Corollary 5.7. *Let \mathcal{X} be a Euclidean or Hilbert space, and f be a real polynomial with nonnegative coefficients. Then $k(x, y) := f(\langle x, y \rangle)$ is a spsd kernel on \mathcal{X} .*

Corollary 5.8. *Let \mathcal{X} be a Euclidean or Hilbert space, and $F(t) = \sum_{i \geq 0} a_i t^i$ be an analytical function with real, nonnegative coefficients $a_i \geq 0$ and convergence radius $R > 0$. Then $k(x, y) := F(\langle x, y \rangle)$ is a spsd kernel on $B_{\mathcal{X}}(0, \sqrt{R}) = \{x \in \mathcal{X} : \|x\| < \sqrt{R}\}$.*

Proof. Corollary 5.7 is a direct consequences of points (ii)-(iii)-(iv)-(v) of Theorem 5.5. Corollary 5.8 is a consequence of the previous corollary and of point (vii) of Theorem 5.5, noticing that if $x, y \in B_{\mathcal{X}}(0, \sqrt{R})$ then $|\langle x, y \rangle| \leq \|x\| \|y\| < R$, so $\langle x, y \rangle$ is within the convergence radius of the power series defining $F(t)$, and therefore $F(\langle x, y \rangle)$ is the limit of the corresponding truncated series, while each such truncated series defines a spsd kernel by Corollary 5.7. □

Examples of spsd kernels. In each of the following examples \mathcal{X} is a subset of \mathbb{R}^d .

- $k(x, y) = (\langle x, y \rangle + c)^m$ for $m \in \mathbb{N}^*$, $c > 0$: polynomial kernel of order m .
- $k(x, y) = (1 - \langle x, y \rangle)^{-\alpha}$ for $\alpha > 0$, on $\mathcal{X} = B_{\mathbb{R}^d}(0, 1)$: negative binomial kernel.
- $k(x, y) = \exp(\lambda \langle x, y \rangle)$, for $\lambda > 0$: exponential kernel.
- $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$, for $\sigma > 0$: Gaussian kernel.

Exercise 5.2. Justify that the above kernels are spsd.

5.4 Kernel-based methods

We now turn back to our initial motivation: how to use a spsd kernel as a “proxy” for a scalar product, and adapt linear methods to the kernel setting. Remember that:

- using a spsd kernel k instead of the regular scalar product can be seen as (implicitly) mapping the x -data to a Hilbert space \mathcal{H} via a feature mapping Φ , and applying the linear method to the transformed data \tilde{x} .
- linear functions of the transformed data \tilde{x} are in general non-linear functions of the original data x .
- for each algorithm we want to find a suitable representation of the learnt function as an expansion of the form (5.2) of the transformed data, i.e.:

$$\hat{w} = \sum_{i=1}^n \beta_i \Phi(X_i), \quad (5.15)$$

thus we only need to store the n -vectors of the coefficients $(\beta_i)_{1 \leq i \leq n}$ and to determine how to compute them from the only information of the scalar products $(\langle X_i, X_j \rangle)_{1 \leq i, j \leq n}$ and the labels $(Y_i)_{1 \leq i \leq n}$.

In this section, we will assume that k is a given spsd on \mathcal{X} , \mathcal{H} the associated RKHS, and denote K the *kernel Gram matrix* of the x -data, i.e. $K_{ij} := \langle X_i, X_j \rangle$, $1 \leq i, j \leq n$.

If we are using the RKHS \mathcal{H} associated to k , due to the reproducing property, if $w \in \mathcal{H}$ we have

$$f_w(x) := \langle w, \Phi(x) \rangle = \langle w, k_x \rangle = w(x); \text{ hence } f_w = w,$$

in other words the function f_w associated to w is w itself, also the representation (5.15) becomes (here denoting \hat{f} instead of \hat{w} to emphasize that it is a function):

$$\hat{f} = \sum_{i=1}^n \beta_i k_{X_i}. \quad (5.16)$$

This form is sometimes called “kernel expansion” or “dual representation”.

* **Kernel perceptron.** Recall again the standard perceptron iteration ($\mathcal{Y} = \{-1, 1\}$):

$$\widehat{w}_0 = 0; \quad \widehat{w}_{\ell+1} = \widehat{w}_\ell + X_{i_\ell} Y_{i_\ell}, \text{ where } i_\ell \text{ is any index s.t. } Y_{i_\ell} \langle \widehat{w}_\ell, X_{i_\ell} \rangle \leq 0.$$

In the “kernelized perceptron”, we want to represent the vectors $\widehat{w}_\ell \in \mathcal{H}$ by means of its vector of coefficients $\beta^{(\ell)} \in \mathbb{R}^n$ in the representation (5.15). So the above becomes:

$$\beta^{(0)} = 0; \quad \beta^{(\ell+1)} = \beta^{(\ell)} + e_{i_\ell},$$

where (e_i) is the i -th canonical basis vector of \mathbb{R}^n , and i_ℓ is any index such that

$$Y_{i_\ell} \sum_{i=1}^m \beta_i^{(\ell)} \langle \Phi(X_{i_\ell}), \Phi(X_i) \rangle = Y_{i_\ell} \sum_{i=1}^m \beta_i^{(\ell)} k(X_{i_\ell}, X_i) = Y_{i_\ell} [K\beta^{(\ell)}]_{i_\ell} \leq 0.$$

After L iterations, the corresponding prediction function is

$$\text{sign}\left(\widehat{f}_{\widehat{w}_L}(x)\right) = \text{sign}\left(\left\langle \sum_{i=1}^n \beta_i^{(L)} \phi(X_i), x \right\rangle\right) = \text{sign}\left(\sum_{i=1}^n \beta_i^{(L)} k(X_i, x)\right).$$

In the case of the perceptron, regularization is obtained by early stopping, which is to say, stopping at an iteration before all training points are all classified correctly. The stopping iteration is typically determined from a predetermined set of values of K by hold-out or cross-validation.

** **Regularized kernel ERM.** We assume here that the prediction space is $\widetilde{Y} = \mathbb{R}$, and $\ell : \widetilde{\mathcal{Y}} \times \mathcal{Y} \rightarrow \mathbb{R}$ is a loss function. We consider ERM over prediction functions of the form $f_w(x) := \langle w, \Phi(x) \rangle$ for $w \in \mathcal{H}$; as we have seen it holds $f_w = w$, hence our class of prediction functions is the RKHS \mathcal{H} itself. Furthermore, we consider regularization by the squared RKHS norm. For a regularization parameter $\lambda > 0$, we therefore define

$$\widehat{f}_\lambda \in \underset{w \in \mathcal{H}}{\text{Arg Min}} \left(\sum_{i=1}^n \ell(Y_i, \langle w, \Phi(X_i) \rangle) + \lambda \|w\|_{\mathcal{H}}^2 \right) = \underset{f \in \mathcal{H}}{\text{Arg Min}} \left(\sum_{i=1}^n \ell(Y_i, f(X_i)) + \lambda \|f\|_{\mathcal{H}}^2 \right). \quad (5.17)$$

We want to prove that, in general, \widehat{f}_λ admits the representation (5.16) This will be established as a consequence of the following result.

Theorem 5.9 (Representation theorem). *Let \mathcal{H} be a RKHS on \mathcal{X} with kernel k . Let $n > 0$ be an integer and $\Psi : \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}$ be a mapping such that:*

for any $\mathbf{x} \in \mathbb{R}^n$, the function $t \in \mathbb{R}_+ \mapsto \Psi(\mathbf{x}, t)$ is nondecreasing.

For any $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$, denote

$$\mathcal{S}_{\mathbf{x}} := \text{Span}\{k_{x_i}, i = 1, \dots, n\} = \left\{ \sum_{i=1}^n \beta_i k_{x_i}, (\beta_1, \dots, \beta_n) \in \mathbb{R}^n \right\},$$

then it holds

$$\inf_{f \in \mathcal{H}} \Psi(f(x_1), \dots, f(x_n), \|f\|_{\mathcal{H}}) = \inf_{f \in \mathcal{S}_{\mathbf{x}}} \Psi(f(x_1), \dots, f(x_n), \|f\|_{\mathcal{H}}).$$

Furthermore, if the above infimum on the left-hand side is a minimum, then it is also a minimum on the right-hand side; in other words, the minimum over \mathcal{H} is attained at for an element of $\mathcal{S}_{\mathbf{x}}$.

Proof. Let $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$ be fixed. Observe that $\mathcal{S}_{\mathbf{x}}$ is a closed subset of \mathcal{H} since it is finite-dimensional. Hence there exists a well-defined orthogonal projector Π onto $\mathcal{S}_{\mathbf{x}}$.

For $f \in \mathcal{H}$, denote for short $\bar{\Psi}(f) := \Psi(f(x_1), \dots, f(x_n), \|f\|_{\mathcal{H}})$. Let f_{ε} be such that $\bar{\Psi}(f_{\varepsilon}) \leq \inf_{f \in \mathcal{H}} \bar{\Psi}(f) + \varepsilon$ for a fixed constant $\varepsilon > 0$. Consider the decomposition $f_{\varepsilon} = \tilde{f}_{\varepsilon} + f_{\varepsilon}^{\perp}$, where $\tilde{f}_{\varepsilon} := \Pi f_{\varepsilon}$; observe that $f_{\varepsilon}^{\perp} \perp \mathcal{S}_{\mathbf{x}}$, and therefore, by the reproducing property,

$$\forall i = 1, \dots, n: \quad f_{\varepsilon}^{\perp}(x_i) = \langle f_{\varepsilon}^{\perp}, k_{x_i} \rangle = 0,$$

so that $f_{\varepsilon}(x_i) = \tilde{f}_{\varepsilon}(x_i)$, for $i = 1, \dots, n$.

On the other hand, since an orthogonal projector is a contraction, it holds $\|\tilde{f}_{\varepsilon}\|_{\mathcal{H}} \leq \|f_{\varepsilon}\|_{\mathcal{H}}$. Therefore

$$\begin{aligned} \bar{\Psi}(\tilde{f}_{\varepsilon}) &:= \Psi(\tilde{f}_{\varepsilon}(x_1), \dots, \tilde{f}_{\varepsilon}(x_n), \|\tilde{f}_{\varepsilon}\|_{\mathcal{H}}) = \Psi(f_{\varepsilon}(x_1), \dots, f_{\varepsilon}(x_n), \|\tilde{f}_{\varepsilon}\|_{\mathcal{H}}) \\ &\leq \Psi(f_{\varepsilon}(x_1), \dots, f_{\varepsilon}(x_n), \|f_{\varepsilon}\|_{\mathcal{H}}) \\ &= \bar{\Psi}(f_{\varepsilon}), \end{aligned}$$

where the inequality comes from the monotonicity of Ψ in its last variable. This proves the first claim of the theorem and, in case the infimum over $f \in \mathcal{H}$ is a minimum, the same argument with $\varepsilon = 0$ holds to prove the second claim. \square

Examples of kernel ERM methods . We revisit here different (regularized) ERM methods studied in Chapter 2 in “kernelized” form, using the representation (5.16) which we know holds in general, due to Theorem 5.9 (we assume in each case that the minimum is attained.)

Kernel Ridge Regression. Kernel ridge regression is regularized least-squares regression using a RKHS as function space and the RKHS norm as regularization:

$$\text{for } \lambda > 0: \quad \hat{f}_{\lambda} \in \text{Arg Min}_{f \in \mathcal{H}} \left(\sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right). \quad (5.18)$$

From Theorem 5.9 we know that we can assume the representation (5.16), i.e. $\widehat{f}_\lambda = \sum_{i=1}^n \beta_{\lambda,i} k(X_i, \cdot)$. Denoting $\beta_\lambda := (\beta_{\lambda,1}, \dots, \beta_{\lambda,n}) \in \mathbb{R}^n$ the coefficients of this expansion, we observe that $\|\widehat{f}_\lambda\|_{\mathcal{H}}^2 = \sum_{i,j=1}^n \beta_{\lambda,i} \beta_{\lambda,j} k(X_i, X_j) = \beta_\lambda^T K \beta_\lambda$; also $(\widehat{f}_\lambda(X_1), \dots, \widehat{f}_\lambda(X_n)) = K \beta_\lambda$. Restricting the search for a minimum for functions of this form, (5.18) becomes:

$$\text{for } \lambda > 0 : \quad \beta_\lambda \in \underset{\beta \in \mathbb{R}^n}{\text{Arg Min}} (\|\mathbf{Y} - K\beta\|^2 + \lambda \beta^t K \beta), \quad (5.19)$$

where we recall $\mathbf{Y} = (Y_1, \dots, Y_n)^t$. By usual arguments (cancelling the first derivative wrt. β of the above function to minimize), we obtain the necessary and sufficient condition

$$K(K + \lambda I_n) \beta_\lambda = K \mathbf{Y};$$

hence a solution is

$$\beta_\lambda = (K + \lambda I_n)^{-1} \mathbf{Y}, \quad (5.20)$$

observe that we have recovered exactly the formula (5.5) discussed in the beginning in the chapter, but for the data mapped into the Hilbert space.

Kernel logistic regression. Recall that logistic regression (for a binary classification problem with label space $\mathcal{Y} = \{-1, 1\}$) can be seen as an ERM estimator with loss function

$$\ell_{\text{logit}}(f(x), y) = \log(1 + \exp(-(yf(x))), \quad y \in \{-1, 1\},$$

see Exercise 2.4. Again, for the “kernelized” (and regularized) version given by (5.17), and the representation (5.16) with the fact that $\widehat{f}_\lambda(X_i) = [K \beta_\lambda]_i$, the coefficient vector $\beta_\lambda \in \mathbb{R}^n$ is defined by

$$\beta_\lambda \in \underset{\beta \in \mathbb{R}^n}{\text{Arg Min}} \left(\sum_{i=1}^n \log(1 + \exp(-Y_i [K \beta]_i)) + \lambda \beta^t K \beta \right),$$

which is a convex optimization problem in β and can be solved by standard methods such as gradient descent, stochastic gradient descent, or Newton-Raphson iterations. The latter requires inversion of a (n, n) Hessian matrix at each step, which can be prohibitive, so the former methods might be preferred even if their convergence rate is not as fast.

In the multiclass case ($\mathcal{Y} = \{0, \dots, K-1\}$), a similar argument holds with an appropriate loss function, and the fact that we are looking for $(K-1)$ score functions $\widehat{f}_{\lambda,1}, \dots, \widehat{f}_{\lambda,K-1}$, it suffices to adapt (2.13) in the kernel setting.

Kernel Support vector machine. It is in all points similar to the previous argument for logistic regression (still for binary classification with $\mathcal{Y} = \{-1, 1\}$), but with the loss function $\ell_{\text{Hinge}}(f(x), y) := (1 - yf(x))_+$. Again, the optimization problem for the kernel expansion coefficients $\beta_\lambda \in \mathbb{R}^n$ is convex. There exists a number of implementations using further reformulations of the problem and using the particular form of the loss function for efficient computation of an approximate minimum. It is one of the most standard classification methods of machine learning toolboxes.

5.5 Regularity and approximation properties of functions in a RKHS

It is of general interest to understand the properties of the functions belonging to a RKHS with a given kernel, since the RKHS is the class of functions we use as predictors (or scores in the case of classification) in different learning settings.

From a practical point of view, we can observe that due to the representation (5.16) as a finite kernel expansion, the considered estimators belong (in general) to \mathcal{H}_{pre} . Therefore, whenever the kernel function $k(\cdot, \cdot)$ is measurable, resp. bounded, resp. continuous with respect to either of its variables, so are the functions on \mathcal{H}_{pre} , by finite linear combination. Still, it is of mathematical interest (for further mathematical analysis, use of Hilbertian analysis tools, etc.) to understand if this is also the case for the full RKHS obtained as the completion of \mathcal{H}_{pre} .

*

Theorem 5.10 (Measurability). *Let \mathcal{X} be a measurable space and k a spsd kernel on \mathcal{X} with RKHS \mathcal{H} . Then*

$$(\forall f \in \mathcal{H}, f \text{ is measurable}) \Leftrightarrow (\forall x \in \mathcal{X} : k(x, \cdot) \text{ is measurable}).$$

Proof. (\Rightarrow) trivial since $k(x, \cdot) = k_x \in \mathcal{H}$ for all $x \in \mathcal{X}$.

(\Leftarrow) by linearity, any function \mathcal{H}_{pre} is measurable. Now for $f \in \mathcal{H}$, remember that \mathcal{H}_{pre} is dense in \mathcal{H} (Theorem 5.3), so there exists a sequence $(f_n)_{n \geq 1}$ of elements of \mathcal{H}_{pre} such that $\|f_n - f\|_{\mathcal{H}} \rightarrow 0$ as $n \rightarrow \infty$. This implies pointwise convergence since by the reproducing property and Cauchy-Schwarz's inequality, for any $x \in \mathcal{X}$

$$|f_n(x) - f(x)| = |\langle f - f_n, k_x \rangle| \leq \|f - f_n\|_{\mathcal{H}} \|k_x\|_{\mathcal{H}} = \|f - f_n\|_{\mathcal{H}} \sqrt{k(x, x)}.$$

Therefore f is measurable as a pointwise limit of measurable functions. \square

*

Theorem 5.11 (Boundedness). *Let be k a spsd kernel on a nonempty set \mathcal{X} with RKHS \mathcal{H} . Then the following are equivalent:*

$$(i) \forall f \in \mathcal{H} : \sup_{x \in \mathcal{X}} |f(x)| < \infty.$$

$$(ii) \sup_{x \in \mathcal{X}} k(x, x) < \infty.$$

$$(iii) \sup_{(x, y) \in \mathcal{X}^2} |k(x, y)| < \infty.$$

In the case either of these properties are satisfied, the topology of the norm on \mathcal{H} is stronger than the supremum norm topology, more precisely

$$\forall f \in \mathcal{H} : \|f\|_{\infty} \leq \|f\|_{\mathcal{H}} \sup_{x \in \mathcal{X}} \sqrt{k(x, x)}. \quad (5.21)$$

Proof. (ii) \Rightarrow (i) because $|f(x)| = |\langle f, k_x \rangle| \leq \|f\|_{\mathcal{H}} \sqrt{k(x, x)}$; this also implies (5.21).

(ii) \Rightarrow (iii) because $|k(x, y)| = |\langle k_x, k_y \rangle| \leq \sqrt{k(x, x)} \sqrt{k(y, y)}$.

(iii) \Rightarrow (ii): trivial

(i) \Rightarrow (ii): can be seen as a consequence of the Banach-Steinhaus theorem. Namely, consider the family of linear forms on \mathcal{H} given by $\mathcal{L} := \{\delta_x, x \in \mathcal{X}\}$, where δ_x is the evaluation functional at point x given by (5.14); we have

$$\forall f \in \mathcal{F} : \sup_{L \in \mathcal{L}} |L(f)| = \sup_{x \in \mathcal{X}} |f_x| < \infty$$

by assumption. The Banach-Steinhaus theorem implies (since \mathcal{H} is a Banach space) that the pointwise bounded family of linear forms \mathcal{L} is uniformly bounded, therefore $\sup \|\delta_x\|_{\mathcal{H}^*} < \infty$. But by the reproduction property it holds $\delta_x = k_x^*$, therefore

$$\|\delta_x\|_{\mathcal{H}^*} = \|k_x^*\|_{\mathcal{H}^*} = \sup_{\|f\|_{\mathcal{H}}=1} \langle k_x, f \rangle = \sqrt{k(x, x)}.$$

□

*

Theorem 5.12 (Continuity). *Let \mathcal{X} be a topological space and k a spsd kernel on \mathcal{X} with RKHS \mathcal{H} . Then*

$$(k \text{ is continuous } \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}) \Rightarrow (\forall f \in \mathcal{H}, f \text{ is continuous}).$$

Proof. We have for any $f \in \mathcal{H}$:

$$|f(x) - f(y)| = |\langle f, k_x - k_y \rangle| \leq \|f\|_{\mathcal{H}} (k(x, x) + k(y, y) - 2k(x, y))^{\frac{1}{2}},$$

and the last factor converges to 0 as $x \rightarrow y$, by bivariate continuity of k .

□

Universal kernels. We end this chapter by important results on the approximation properties of RKHSs. Introduce the following definition:

**

Definition 5.13.

Let \mathcal{X} be a nonempty, compact topological space. Then a continuous spsd kernel k on $\mathcal{X} \times \mathcal{X}$ is called *universal* then the corresponding RKHS \mathcal{H} is dense (in the sense of the supremum norm) in the space $\mathcal{C}(X)$ of continuous real-valued functions on \mathcal{X} .

This definition extends to a non-compact topological space \mathcal{X} : then k is said to be universal if its restriction on any compact subset of \mathcal{X} is universal.

We begin with the following observation:

Proposition 5.14. *Let \mathcal{X} be a nonempty, compact topological space. Then a continuous spsd on \mathcal{X} is universal iff \mathcal{H}_{pre} is dense in $\mathcal{C}(X)$ for the supremum norm.*

Proof. (\Leftarrow): since $\mathcal{H}_{\text{pre}} \subseteq \mathcal{H}$, \mathcal{H}_{pre} dense in $\mathcal{C}(X)$ trivially implies that \mathcal{H} is also dense in $\mathcal{C}(X)$.

(\Rightarrow): we know that \mathcal{H}_{pre} is dense in \mathcal{H} in the sense of the \mathcal{H} -norm. Since k is continuous on the compact $\mathcal{X} \times \mathcal{X}$, it is bounded, therefore (5.21) applies and \mathcal{H}_{pre} is *a fortiori* dense in \mathcal{H} for the supremum norm, and therefore also dense in $\mathcal{C}(\mathcal{X})$ for the supremum norm, since \mathcal{H} is. \square

The following final result of this section is very useful to establish universality of a number of classical kernels on \mathbb{R}^d .

**

Theorem 5.15 (Universal Taylor kernels). *Let $F(t) = \sum_{i \geq 0} a_i t^i$ be a real-valued analytical function with real, strictly positive coefficients $a_i > 0$, and convergence radius $R > 0$.*

Let $\mathcal{X} = \mathbb{R}^d$. Then $k(x, y) := F(\langle x, y \rangle)$ is a universal spsd kernel on $B_{\mathcal{X}}(0, \sqrt{R}) = \{x \in \mathcal{X} : \|x\| < \sqrt{R}\}$.

This result implies in particular that the exponential kernel and the negative-binomial kernel introduced at the end of Section 5.3 are universal on \mathbb{R}^d and $B_{\mathbb{R}^d}(0, 1)$, respectively.

Lemma 5.16. *Let k be a spsd kernel on a nonempty set \mathcal{X} , and \mathcal{H}_\circ be a Hilbert space and Φ_\circ a mapping $\mathcal{X} \rightarrow \mathcal{H}_\circ$, so that it holds for all x, y in \mathcal{X} : $\langle \Phi_\circ(x), \Phi_\circ(y) \rangle_{\mathcal{H}_\circ} = k(x, y)$.*

Then, for any $w \in \mathcal{H}_\circ$, the function $x \mapsto \langle w, \Phi_\circ(x) \rangle_{\mathcal{H}_\circ}$ belongs to the RKHS \mathcal{H} associated to k .

Proof. Let us denote $\xi : \mathcal{H}_\circ \rightarrow \mathcal{F}(\mathcal{X}, \mathbb{R})$ the linear mapping given by

$$\xi(w) = (x \mapsto \langle w, \Phi_\circ(x) \rangle_{\mathcal{H}_\circ}).$$

If $w = \Phi_\circ(x)$ for some $x \in \mathcal{X}$, the function $x' \mapsto \langle w, \Phi_\circ(x') \rangle_{\mathcal{H}_\circ} = k(x, x')$ coincides with $k(x, \cdot)$, i.e. $\xi(\Phi_\circ(x)) = k_x \in \mathcal{H}_{\text{pre}} \subseteq \mathcal{H}$ and furthermore $\|\xi(\Phi_\circ(x))\|_{\mathcal{H}} = k(x, x) = \|\Phi_\circ(x)\|_{\mathcal{H}_\circ}$. By linearity, ξ is an isometry from $\mathcal{H}_1 := \text{Span}\{\Phi_\circ(x), x \in \mathcal{X}\}$ into \mathcal{H} (in fact into \mathcal{H}_{pre}), in particular $\xi(\mathcal{H}_1) \subseteq \mathcal{H}$.

For any sequence $(w_n)_{n \geq 1}$ of elements in \mathcal{H}_1 converging to w_* in \mathcal{H}_\circ , the sequence $(\xi(w_n))_{n \geq 1}$ is Cauchy in \mathcal{H} (by isometry) and therefore converges to a limit f^* . But it holds for any $x \in \mathcal{X}$ (using the definition of ξ , continuity of scalar products in \mathcal{H}_\circ and \mathcal{H} ,

the isometry property of ξ , and the reproducing property in \mathcal{H}):

$$\begin{aligned}\xi(w^*)(x) &= \left\langle \lim_{n \rightarrow \infty} w_n, \Phi_\circ(x) \right\rangle_{\mathcal{H}_\circ} = \lim_{n \rightarrow \infty} \langle w_n, \Phi_\circ(x) \rangle_{\mathcal{H}_\circ} \\ &= \lim_{n \rightarrow \infty} \langle \xi(w_n), \xi(\Phi_\circ(x)) \rangle_{\mathcal{H}} \\ &= \left\langle \lim_{n \rightarrow \infty} \xi(w_n), k_x \right\rangle_{\mathcal{H}} \\ &= f^*(x),\end{aligned}$$

hence $\xi(w^*) = f^*$. Therefore $\xi(\overline{\mathcal{H}}_1) \subseteq \mathcal{H}$.

Finally, since $\overline{\mathcal{H}}_1$ is closed in \mathcal{H}_\circ , it holds $\mathcal{H}_\circ = \overline{\mathcal{H}}_1 \oplus (\overline{\mathcal{H}}_1)^\perp$. But for any $w \in (\overline{\mathcal{H}}_1)^\perp$ and any $x \in \mathcal{X}$, since $w \perp \Phi_\circ(x) \in \mathcal{H}_1$ it holds

$$\xi(w)(x) = \langle w, \Phi_\circ(x) \rangle_{\mathcal{H}_\circ} = 0.$$

Finally we have established $\xi(\mathcal{H}_\circ) \subseteq \mathcal{H}$, which is the desired claim. \square

Proof of Theorem 5.15. In this proof we will construct a Hilbert space \mathcal{H}_\circ and Φ_\circ a mapping $\mathcal{X} \rightarrow \mathcal{H}_\circ$, with the property $\langle \Phi_\circ(x), \Phi_\circ(y) \rangle_{\mathcal{H}_\circ} = k(x, y)$, which will be different from the RKHS \mathcal{H} associated to k , but we will map it into \mathcal{H} using the previous lemma.

To simplify the next calculation we introduce the following notation: for $\mathbf{j} = (j_1, \dots, j_d) \in \mathbb{N}^d$, put $s(\mathbf{j}) = j_1 + \dots + j_d$; $c(\mathbf{j}) := \binom{s(\mathbf{j})}{j_1, \dots, j_d}$ the multinomial coefficient; and for $x \in \mathcal{X}$, $m_{\mathbf{j}}(x) := \prod_{i=1}^d x_i^{j_i}$ the monomial in the coordinates of x associated to the multi-index \mathbf{j} .

Observe that since $F(t) = \sum_{i \geq 0} a_i t^i$, we have (for $\max(\|x\|, \|y\|) \leq R$ ensuring convergence of the series below), by the multinomial formula:

$$\begin{aligned}k(x, y) &= \sum_{\ell \geq 0} a_\ell \langle x, y \rangle^\ell = \sum_{\ell \geq 0} a_\ell \left(\sum_{j=1}^d x_j y_j \right)^\ell = \sum_{\ell \geq 0} a_\ell \sum_{j_1 + \dots + j_d = \ell} c(j_1, \dots, j_d) \prod_{i=1}^d (x_i y_i)^{j_i} \\ &= \sum_{\mathbf{j} \in \mathbb{N}^d} a_{s(\mathbf{j})} c(\mathbf{j}) m_{\mathbf{j}}(x) m_{\mathbf{j}}(y) \\ &= \sum_{\mathbf{j} \in \mathbb{N}^d} \phi_{\mathbf{j}}(x) \phi_{\mathbf{j}}(y),\end{aligned}$$

where $\phi_{\mathbf{j}}(x) := \sqrt{a_{s(\mathbf{j})} c(\mathbf{j})} m_{\mathbf{j}}(x)$.

We therefore consider $\mathcal{H}_\circ := \ell_2(\mathbb{N}^d)$, and $\Phi_\circ(x) := (\phi_{\mathbf{j}}(x))_{\mathbf{j} \in \mathbb{N}^d}$. Note that absolute convergence of the power series defining F ensures that $\Phi_\circ(x) \in \mathcal{H}_\circ$, i.e.

$\sum_{\mathbf{j} \in \mathbb{N}^d} (\phi_{\mathbf{j}}(x))^2 < \infty$ for any $x \in \mathbb{R}^d$ such that $\|x\| < \sqrt{R}$.

We can apply the previous lemma and conclude that for any $w \in \mathcal{H}_\circ$, the function $x \mapsto \langle w, \Phi_\circ(x) \rangle_{\mathcal{H}_\circ}$ belongs to the RKHS \mathcal{H} . Let us choose, for an arbitrary multi-index $\mathbf{j} \in \mathbb{N}^d$, the vector $w \in \mathcal{H}_\circ$ such that the \mathbf{j} -coordinate of w is $1/\sqrt{a_{s(\mathbf{j})} c(\mathbf{j})}$ and the other coordinates 0. Then for any $x \in \mathcal{X}$:

$$\langle w, \Phi_\circ(x) \rangle_{\mathcal{H}_\circ} = m_{\mathbf{j}}(x).$$

We conclude that all monomial functions in the coordinates of x belong to \mathcal{H} ; then also all polynomials by linearity, and the conclusion is a consequence of the Stone-Weierstraß theorem. \square

◇ 5.6 Translation invariant kernels and random Fourier features

In this section, we will address recent developments of the kernel methodology that are relevant for practice. We have seen that the spsd kernel methodology allows to implicitly represent feature mappings of the x -data into an infinite-dimensional Hilbert space \mathcal{H} . However, all kernel-based methods require to store and manipulate the kernel Gram matrix K which is a (n, n) matrix. In typical modern applications the sample size n can be very large (hundreds of millions) which can make the computation and storage of such a matrix, with $\mathcal{O}(n^2)$ complexity, prohibitive; let alone manipulating it numerically.

For this reason, it has been proposed to construct *explicit, approximate* feature mappings $\tilde{\Phi} : \mathcal{X} \rightarrow \mathbb{R}^p$ (with $p \ll n$) such that $\langle \tilde{\Phi}(x), \tilde{\Phi}(y) \rangle \approx k(x, y)$. While it seems like we are now going backwards to the beginning of the chapter — and thus maybe could do completely without kernels at all — having gained a mathematical understanding of the properties of the mathematical object we are approximating (a RKHS) is very valuable, and we could not properly understand the methods considered below without having introduced RKHSs.

We start with a few reminders on Fourier transform

Basic facts on Fourier transform on \mathbb{R}^d

For any $f \in L^1(\mathbb{R}^d, \mathbb{C})$, the d -dimensional Fourier transform $\hat{f} = \mathcal{F}(f)$

$$\hat{f}(\omega) := \int_{\mathbb{R}^d} \exp(-i\langle x, \omega \rangle) f(x) dx ; \quad \omega \in \mathbb{R}^d \quad (5.22)$$

exists and satisfies:

- $\mathcal{F}(f) = \hat{f}$ is continuous on \mathbb{R}^d ;
- $\lim_{\|\omega\| \rightarrow \infty} \hat{f}(\omega) = 0$;
- $\|\hat{f}\|_{\infty} \leq \|f\|_{L^1}$;
- The inverse Fourier transform formula holds: if $\hat{f} \in L^1(\mathbb{R}^d, \mathbb{C})$ holds, then

$$f(x) \stackrel{\text{a.s.}}{=} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp(i\langle x, \omega \rangle) \hat{f}(\omega) d\omega ; \quad x \in \mathbb{R}^d. \quad (5.23)$$

- If $(x \mapsto \|\omega\|^k f(\omega)) \in L^1(\mathbb{R}^d)$, then \hat{f} is k times continuously differentiable and

$$\forall \ell \leq k, \quad (i_1, \dots, i_\ell) \in \llbracket d \rrbracket^\ell, \quad \partial_{i_1} \dots \partial_{i_\ell} \hat{f}(\omega) = \mathcal{F}(x \mapsto (-i)^\ell x_{i_1} \dots x_{i_\ell} f(x)).$$

Remark: there are several concurrent normalization conventions in the literature for the Fourier transform. Two convenient (and probably more common nowadays) alternative conventions are to put a factor $(2\pi)^{-d/2}$ in front of the integral defining the transform, or to add a 2π factor inside the complex exponential. Either of these conventions have the advantage that the inverse Fourier transform takes exactly the same form as the direct transform, but with i replaced by $-i$, which is more symmetric. In these notes we stick to the above definition because we won't use the inverse Fourier transform heavily.

Translation-invariant kernels. A fundamental relation between Fourier transform and translation invariant kernels is given in the following theorem.

Theorem 5.17. *Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$, continuous and let $k(x, y) = \varphi(x - y)$. Assume $\varphi = \widehat{f}$ for some $f \in L^1(\mathbb{R}^d, \mathbb{R})$, with $f(x) \geq 0$ a.s. Then k is a spsd kernel.*

Note: this direction is actually the “easy” one. There exists a converse (*Bochner's theorem*) stating that if k is a spsd, translation-invariant kernel — i.e. it is of the form $k(x, y) = k(x - y, 0) = \varphi(x - y)$, where $\varphi = k(0, \cdot)$ — then φ is the Fourier-Stieltjes transform of a finite (nonnegative) measure on \mathbb{R}^d .

Proof. It holds for any integer $n > 0$, $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ and $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$:

$$\begin{aligned} \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) &= \sum_{i,j=1}^n \alpha_i \alpha_j \varphi(x_i - x_j) \\ &= \sum_{i,j=1}^n \alpha_i \alpha_j \int_{\mathbb{R}^d} \exp(-i\langle x_i, \omega \rangle) \exp(i\langle x_j, \omega \rangle) f(\omega) d\omega \\ &= \int_{\mathbb{R}^d} \left| \sum_{i=1}^n \alpha_i \exp(-i\langle x_i, \omega \rangle) \right|^2 f(\omega) d\omega \geq 0. \end{aligned}$$

□

Note: we have assumed here as in the rest of the chapter that φ and therefore k are real-valued (which implies in particular that f must be symmetric around 0 in Theorem 5.17). This can be generalized to a more general theory of complex-valued spsd (Hermitian) kernels, *mutatis mutandis*.

Note: Because of the inverse Fourier formula, provided that φ is integrable we can identify the function f as the inverse Fourier transform of φ , given by (5.23).

Examples.

- We find another proof that the Gaussian kernel $k(x, y) := \exp(-\|x - y\|^2/(2\sigma^2))$ is spsd, with

$$f(t) = \frac{\sigma^d}{(2\pi)^d} \exp\left(-\frac{\sigma^2 t^2}{2}\right).$$

- The Laplace kernel $k(x, y) := \frac{1}{2} \exp(-\gamma|x - y|)$ (with $\gamma \geq 0$) is spsd, with

$$f(t) = \frac{1}{(2\pi)^d} \frac{\gamma}{\gamma^2 + t^2}.$$

Random Fourier features. If $k(x, y) = \varphi(x - y)$ is a spsd translation-invariant kernel on \mathbb{R}^d , such that $\varphi = \widehat{f}$, then

$$k(x, y) = \varphi(x - y) = \int_{\mathbb{R}^d} \exp(-i\langle x, \omega \rangle) \exp(i\langle y, \omega \rangle) f(\omega) d\omega, \quad (5.24)$$

where $f \geq 0$ and $f \in L^1(\mathbb{R}^d)$, with $\int_{\mathbb{R}^d} f(\omega) d\omega = \varphi(0)$. Up to rescaling of φ (and k), we can assume $\varphi(0) = 1$ and thus can interpret f as a *probability density* on \mathbb{R}^d . Let P_f denote the associated probability distribution, then the above can be rewritten as

$$k(x, y) = \mathbb{E}_{\omega \sim P_f} [\exp(-i\langle x, \omega \rangle) \exp(i\langle y, \omega \rangle)]. \quad (5.25)$$

To shorten notation we will denote $\mathbb{E}_{\omega \sim P_f}$ simply as \mathbb{E}_ω . Note that since $k(x, y)$ is real valued, this implies

$$\begin{aligned} k(x, y) &= \operatorname{Re}(\mathbb{E}_\omega [\exp(i\langle y - x, \omega \rangle)]) \\ &= \mathbb{E}_\omega [\operatorname{Re}(\exp(i\langle y - x, \omega \rangle))] \\ &= \mathbb{E}_\omega [\cos(\langle y - x, \omega \rangle)] \\ &= \mathbb{E}_\omega [\cos(\langle x, \omega \rangle) \cos(\langle y, \omega \rangle) + \sin(\langle x, \omega \rangle) \sin(\langle y, \omega \rangle)]. \end{aligned} \quad (5.26)$$

The idea of random Fourier features is to approximate the above expectation by a finite average over p randomly drawn frequency vectors $(\omega_1, \dots, \omega_p) \stackrel{\text{i.i.d.}}{\sim} P_f$. More explicitly, given p such random frequencies, define the mapping

$$\widetilde{\Phi} : \mathbb{R}^d \rightarrow \mathbb{R}^{2p} : \quad x \mapsto \frac{1}{\sqrt{p}} (\cos(\langle \omega_1, x \rangle), \sin(\langle \omega_1, x \rangle), \dots, \cos(\langle \omega_p, x \rangle), \sin(\langle \omega_p, x \rangle)). \quad (5.27)$$

Then it holds

$$\langle \widetilde{\Phi}(x), \widetilde{\Phi}(y) \rangle = \frac{1}{p} \sum_{j=1}^p \cos(\langle y - x, \omega_j \rangle) = \operatorname{Re} \left(\frac{1}{p} \sum_{j=1}^p \exp(-i\langle x, \omega_j \rangle) \exp(i\langle y, \omega_j \rangle) \right), \quad (5.28)$$

which converges to (5.25) in probability as $p \rightarrow \infty$, by the law of large numbers (and the fact that the kernel is real-valued). We can even quantify this convergence:

Proposition 5.18. *If k can be represented as (5.25) and we draw $(\omega_1, \dots, \omega_p) \stackrel{i.i.d.}{\sim} P_f$, and define $\tilde{\Phi}$ by (5.27), then for any x, y in \mathbb{R}^d , $\delta \in [0, 1)$, with probability $1 - \delta$ over the draw of these frequencies, it holds*

$$\left| k(x, y) - \langle \tilde{\Phi}(x), \tilde{\Phi}(y) \rangle \right| \leq \sqrt{\frac{2 \log(2\delta^{-1})}{p}}.$$

Proof. Direct consequence of Hoeffding's inequality (Corollary 3.9), since $|\operatorname{Re}(\exp(-i\langle x, \omega_j \rangle) \exp(i\langle y, \omega_j \rangle))| = |\cos(\langle y - x, \omega \rangle)| \leq 1$. \square

Having a control holding with high probability at any fixed point (x, y) is however not very useful, it is preferable to have a *uniform approximation* property. This is what we do next.

Theorem 5.19 (Rahimi & Recht 2007). *Let \mathcal{X} be a compact subset of \mathbb{R}^d of diameter $R \geq 1$. Assume the ssdp kernel k on \mathcal{X} can be represented as (5.24)-(5.25), with f nonnegative, integrating to 1 over \mathbb{R}^d , and such that the function $(\omega \mapsto \|\omega\|f(\omega))$ is integrable on \mathbb{R}^d .*

Let $A_f = \int \|\omega\|f(\omega)d\omega$ and assume $1 \leq A_f < \infty$.

Draw $(\omega_1, \dots, \omega_p) \stackrel{i.i.d.}{\sim} P_f$, and define $\tilde{\Phi}$ by (5.27). For $\delta, \eta \in [0, 1]$, provided

$$p \gtrsim \frac{d}{\eta^2} \log\left(\frac{A_f R}{\delta \eta}\right),$$

(where \gtrsim indicates inequality up to a numerical factor) it holds with probability at least $1 - \delta$ over the draw of $\omega_1, \dots, \omega_p$:

$$\sup_{(x, y) \in \mathcal{X}^2} \left| \langle \tilde{\Phi}(x), \tilde{\Phi}(y) \rangle - k(x, y) \right| \leq \eta.$$

The plan to prove this result is the following: we want to control the supremum norm of the function $F(x, y) = \langle \tilde{\Phi}(x), \tilde{\Phi}(y) \rangle - k(x, y)$ with high probability on \mathcal{X}^2 . Let us first simplify slightly the problem: as $k(x, y) = \varphi(x - y)$ and $\langle \tilde{\Phi}(x), \tilde{\Phi}(y) \rangle = \tilde{\varphi}(x - y)$ for some (random) function $\tilde{\varphi}$ as seen from (5.26),(5.28), we have $F(x, y) = G(x - y)$ with $G = \tilde{\varphi} - \varphi$, and we aim at controlling $|G(u)|$ uniformly over $\Delta_{\mathcal{X}} = \mathcal{X} - \mathcal{X} = \{y - x : y, x \in \mathcal{X}\} \subset \mathbb{R}^d$.

We know from Proposition 5.18 how to get a pointwise control, but we cannot directly use a union bound over $\Delta_{\mathcal{X}}$ (as in Chapter 3) because it is uncountable. Instead we will

use a **covering argument**: assume there exists a finite set of points \mathcal{C}_ε of $\Delta_{\mathcal{X}}$ that “cover” it in the sense that for any point of $\Delta_{\mathcal{X}}$ there is a point of \mathcal{C}_ε at distance at most ε . We can then control uniformly $G(u)$ on points of the finite set \mathcal{C}_ε by the union bound and then “extend” this control to close points of $\Delta_{\mathcal{X}}$ if G is Lipschitz.

To be more specific, let $\varepsilon > 0$, and \mathcal{C}_ε an ε -cover of $\Delta_{\mathcal{X}}$ for the Euclidean norm. Let L_G be the Lipschitz constant of G . For any point $u \in \Delta_{\mathcal{X}}$, let $u_0 \in \mathcal{C}_\varepsilon$ be its closed point in the cover, then $\|u - u_0\| \leq \varepsilon$, and we have

$$|G(u)| = |G(u) - G(u_0)| + |G(u_0)| \leq \varepsilon L_G + |G(u_0)|$$

so that

$$\sup_{u \in \Delta_{\mathcal{X}}} |G(u)| \leq \varepsilon L_G + \max_{u \in \mathcal{C}_\varepsilon} |G(u)|. \quad (5.29)$$

Thus the main ingredients are (a) find a suitable ε -cover of $\Delta_{\mathcal{X}}$ (and bound its cardinality in function of ε) and write a union bound over it and (b) determine if G is Lipschitz.

We start with point (a). The following result is very general.

Lemma 5.20. *Let $\|\cdot\|_*$ be a norm on \mathbb{R}^d and A a subset of \mathbb{R}^d bounded by $R > 0$ for this norm. Then for any $\varepsilon > 0$ there exists an ε -cover of A of cardinality at most $(1 + \frac{2R}{\varepsilon})^d$.*

Proof. We will consider a volume-based argument for a concept closely related to a covering, namely a “maximal packing”. An ε -packing of a compact set A is a subset \mathcal{P}_ε of A such that the balls $B(x, \varepsilon)$ of radius ε (form the norm $\|\cdot\|_*$) centered at points $x \in \mathcal{P}_\varepsilon$ are disjoint. Let \mathcal{P}_ε be an ε -packing of A of maximal cardinality. We argue that \mathcal{P}_ε must be a 2ε -covering of A . Namely, any point $x_0 \in A$ must be at distance less than 2ε from at least one point of \mathcal{P}_ε , otherwise the ball $B(x_0, \varepsilon)$ would be disjoint from all the $B(x, \varepsilon)$, $x \in \mathcal{P}_\varepsilon$ (by the triangle inequality) and we could add x_0 to \mathcal{P}_ε , contradicting maximality.

On the other hand, let V denote the Lebesgue measure of $B(0, 1)$. By the fact that $B(0, r) = rB(0, 1)$ (by homogeneity of the norm) and change of variable, we have $\text{Vol}(B(0, r)) = Vr^d$ and further by translation invariance of the Lebesgue measure $\text{Vol}(B(x, r)) = Vr^d$ as well. Since all balls $B(x, \varepsilon)$, $x \in \mathcal{P}_\varepsilon$ are disjoint, it therefore holds

$$\text{Vol}\left(\bigcup_{x \in \mathcal{P}_\varepsilon} B(x, \varepsilon)\right) = \sum_{x \in \mathcal{P}_\varepsilon} \text{Vol}(B(x, \varepsilon)) = |\mathcal{P}_\varepsilon|V\varepsilon^d.$$

On the other hand, since A is bounded by R , and $\mathcal{P}_\varepsilon \subset A$, then $\bigcup_{x \in \mathcal{P}_\varepsilon} B(x, \varepsilon) \subset B(0, R + \varepsilon)$, by the triangle inequality. Thus

$$|\mathcal{P}_\varepsilon|V\varepsilon^d = \text{Vol}\left(\bigcup_{x \in \mathcal{P}_\varepsilon} B(x, \varepsilon)\right) \leq \text{Vol}(B(0, R + \varepsilon)) = (R + \varepsilon)^d V,$$

therefore $|\mathcal{P}_\varepsilon| \leq (1 + \frac{R}{\varepsilon})^d$, and we recall \mathcal{P}_ε is a 2ε -covering; we substitute $\varepsilon/2$ for ε to reach the conclusion. \square

Since \mathcal{X} has diameter R , $\Delta_{\mathcal{X}}$ is bounded by R and we can directly apply the previous result.

Concerning point (b), we have the following lemma:

Lemma 5.21. *Under the same conditions as in the above theorem, for any $t \geq 1$ it holds with probability (over the random draw of the weights) $1 - \frac{1}{t}$ that the function $u \in \mathbb{R}^d \mapsto G(u) = \tilde{\varphi}(u) - \varphi(u)$ is globally Lipschitz on \mathbb{R}^d with constant $2A_f t$.*

Proof. We will study separately the Lipschitz property of φ and of the random function $\tilde{\varphi}$ (depending on the draw of the frequency vectors (ω_i)).

Concerning φ , due to (5.26) we have, since \cos is 1-Lipschitz:

$$\begin{aligned} |\varphi(u) - \varphi(u')| &\leq \mathbb{E}_{\omega} [|\cos(\langle u, \omega \rangle) - \cos(\langle u', \omega \rangle)|] \\ &\leq \mathbb{E}_{\omega} [|\langle u - u', \omega \rangle|] \\ &\leq \|u' - u\| \mathbb{E}_{\omega} [\|\omega\|], \end{aligned}$$

so that φ is A_f -Lipschitz (and therefore tA_f -Lipschitz since $t \geq 1$) with respect to its first variable.

Concerning $\tilde{\varphi}$, implicitly defined by (5.28), we have similarly but using an empirical expectation:

$$\begin{aligned} |\tilde{\varphi}(u) - \tilde{\varphi}(u')| &\leq \frac{1}{p} \sum_{j=1}^p |\cos(\langle u, \omega_j \rangle) - \cos(\langle u', \omega_j \rangle)| \\ &\leq \|u' - u\| \frac{1}{p} \sum_{j=1}^p \|\omega_j\|. \end{aligned}$$

Unfortunately, we can't use Hoeffding's inequality to control the last i.i.d. average with high probability, since the variables $\|\omega_j\|$ are not bounded in general. Still, since we require a single estimate for the variable $\frac{1}{p} \sum_{j=1}^p \|\omega_j\|$ resulting in a global Lipschitz property for any u , we'll just resort to Markov's inequality here: for any $t \geq 1$, it holds

$$\mathbb{P} \left[\frac{1}{p} \sum_{j=1}^p \|\omega_j\| > tA_f \right] \leq \frac{\mathbb{E} \left[\frac{1}{p} \sum_{j=1}^p \|\omega_j\| \right]}{tA_f} = \frac{1}{t},$$

so with probability $1 - 1/t$ over the frequencies' draw the function s is globally (tA_f) -Lipschitz.

Collecting the previous arguments leads to the conclusion. \square

With the two previous lemmas in hand, let us proceed to the proof of Theorem 5.19. Let us fix $\varepsilon > 0$, by Hoeffding's inequality applied to each point $u \in \mathcal{C}_{\varepsilon}$ and the union bound, for any $t > 0$ we have

$$\mathbb{P}_{\omega} \left[\max_{u \in \mathcal{C}_{\varepsilon}} |G(u)| > \frac{t}{2} \right] \leq 2|\mathcal{C}_{\varepsilon}| \exp\left(-\frac{pt^2}{8}\right) \leq 2 \left(1 + \frac{2R}{\varepsilon}\right)^d \exp\left(-\frac{pt^2}{8}\right),$$

using Lemma 5.20 for the last inequality. By Lemma 5.21, we have

$$\mathbb{P}_\omega \left[L_G > \frac{t}{2\varepsilon} \right] \leq \frac{4A_f\varepsilon}{t}.$$

Combining the last 2 inequalities with our initial argument (5.29) we get, assuming $\varepsilon \leq R$:

$$\mathbb{P}_\omega \left[\sup_{u \in \Delta_{\mathcal{X}}} |G(u)| > t \right] \leq \frac{4A_f\varepsilon}{t} + 2 \left(\frac{4R}{\varepsilon} \right)^d \exp\left(-\frac{pt^2}{8}\right) = a\varepsilon + b\varepsilon^{-d}.$$

We choose ε to equalize the two terms, that is

$$\varepsilon = (b/a)^{\frac{1}{d+1}} = \left(2(4R)^d \exp(-pt^2/8)t / (4A_f) \right)^{\frac{1}{d+1}},$$

leading to (recalling $t \leq 1 \leq A_f$ and $R \geq 1$)

$$\begin{aligned} \mathbb{P}_\omega \left[\sup_{u \in \Delta_{\mathcal{X}}} |G(u)| > t \right] &\leq 2a^{\frac{d}{d+1}} b^{\frac{1}{d+1}} = 2 \left(\frac{4A_f}{t} \right)^{\frac{d}{d+1}} 2^{\frac{1}{d+1}} (4R)^{\frac{d}{d+1}} \exp\left(-\frac{pt^2}{8(d+1)}\right) \\ &\leq 32 \left(\frac{A_f R}{t} \right) \exp\left(-\frac{pt^2}{d+1}\right), \end{aligned}$$

Choosing $p \geq C \frac{d}{t^2} \log\left(\frac{A_f R}{\delta t}\right)$ for a big enough constant C ensures that the latter quantity is less than δ (check also that with a big enough constant C this ensures $\varepsilon \leq R$ from the formula for the choice of ε ; as this condition was assumed earlier).

6 Introduction to statistical learning theory (part 3): Rademacher complexities and VC theory

6.1 Introduction, reminders

Recall that in Section 3, we studied the behavior of statistical learning methods which output a prediction function \hat{f} belonging to some class \mathcal{F} which was assumed to be *finite* or *countable*. The main mathematical tool was to control to obtain a uniform control of the form

$$\forall f \in \mathcal{F} : |\mathcal{E}(f) - \widehat{\mathcal{E}}(f)| \leq R(n, \mathcal{F}, \delta); \quad (6.1)$$

holding with probability at least $1 - \delta$ over the draw of a sample of size n . (Please note: we only consider the case of a uniform bound $R(n, \mathcal{F}, \delta)$ independent of the function f ; we do not consider bounds depending on f as for instance (3.29) in this discussion.)

When \mathcal{F} is finite, and the loss function is bounded, this was achieved as a consequence of Hoeffding's inequality (see Corollary 3.10), which gives control over a single function f , then a union bound (see Proposition 3.11).

Let us also recall briefly how a uniform bound (6.1) leads to a bound on the risk of an ERM over class \mathcal{F} :

Proposition 6.1. *Let us assume a learning setting (consisting of an observation space \mathcal{X} , a label space \mathcal{Y} , a prediction space $\tilde{\mathcal{Y}}$, and a loss function $\ell : \tilde{\mathcal{Y}} \times \mathcal{Y} \rightarrow \mathbb{R}_+$). Assume that (6.1) holds.*

Let $\eta > 0$ be fixed and \hat{f}_η denote an η -approximate ERM over the class \mathcal{F} that is, $\hat{f}_\eta \in \mathcal{F}$ and

$$\widehat{\mathcal{E}}(\hat{f}_\eta) \leq \inf_{f \in \mathcal{F}} \widehat{\mathcal{E}}(f) + \eta.$$

Then it holds (with the same probability with which (6.1) holds) that

$$\mathcal{E}(\hat{f}_\eta) \leq \mathcal{E}_{\mathcal{F}}^* + 2R(n, \delta, \mathcal{F}) + \eta.$$

Proof. This is a repetition (in a more formal setting) of the argument leading to Proposition 3.13. Let $\varepsilon > 0$ and let $f_\varepsilon \in \mathcal{F}$ be such that $\mathcal{E}(f_\varepsilon) \leq \inf_{f \in \mathcal{F}} \mathcal{E}(f) + \varepsilon$. Since $\hat{f}_\eta \in \mathcal{F}$, and putting $R = R(n, \mathcal{F}, \delta)$ for short, using twice (6.1):

$$\begin{aligned} \mathcal{E}(\hat{f}_\eta) &\leq \widehat{\mathcal{E}}(\hat{f}_\eta) + R \\ &\leq \widehat{\mathcal{E}}(f_\varepsilon) + R + \eta \\ &\leq \mathcal{E}(f_\varepsilon) + 2R + \eta \\ &\leq \mathcal{E}_{\mathcal{F}}^* + 2R + \eta + \varepsilon, \end{aligned}$$

and this holds for any $\varepsilon > 0$, hence the conclusion. \square

Observe that the proposition above is a purely deterministic one once the event (6.1) is satisfied: the only probabilistic point is to establish that this event has probability large enough (while the remainder $R(n, \mathcal{F}, \delta)$ remains hopefully “reasonable”, in particular converging to 0 as the sample size n grows). Additionally, even if the learning algorithm \hat{f} is not ERM, the bound (6.1) allows us to give a confidence bound on the (unknown) risk of \hat{f} based only on its empirical risk, whatever the algorithm used, since the bound is uniform.

In view of the above considerations, a goal of interest is to extend the uniform control (6.1) to more general classes, in particular (uncountably) infinite. Observe that (6.1) is equivalent to a probabilistic upper bound of the random variable

$$Z_{\mathcal{F}}^{|\cdot|} := \sup_{f \in \mathcal{F}} |\mathcal{E}(f) - \widehat{\mathcal{E}}(f)|, \quad (6.2)$$

holding with probability $1 - \delta$. This is what we set to do in the next sections, and we will achieve it in two steps: (a) bound the deviations of $Z_{\mathcal{F}}^{|\cdot|}$ from its expectation, with high probability; (b) bound the expectation of $Z_{\mathcal{F}}^{|\cdot|}$. We will also be interested in similar bounds for the closely related variables

$$Z_{\mathcal{F}}^+ := \sup_{f \in \mathcal{F}} (\widehat{\mathcal{E}}(f) - \mathcal{E}(f)), \text{ and } Z_{\mathcal{F}}^- := \sup_{f \in \mathcal{F}} (\mathcal{E}(f) - \widehat{\mathcal{E}}(f)). \quad (6.3)$$

Note. In complete generality, it cannot be ensured that the variable defined in (6.2) is measurable because a supremum of uncountably many measurable functions is not necessarily measurable. We will ignore that point and assume implicitly that $Z_{\mathcal{F}}^{|\cdot|}$ (and other suprema) are measurable throughout the chapter. It can for example be assumed (this is often the case) that there exists a countable subset $\tilde{\mathcal{F}} \subset \mathcal{F}$ such that the suprema over \mathcal{F} and $\tilde{\mathcal{F}}$ coincide a.s.; this is usually the case for more concrete prediction classes.

Exercise 6.1. In the case of a countably infinite set \mathcal{F} , we had derived a bound of the form (6.1) but with a bound $R(n, f, \mathcal{F}, \delta)$ also depending on f due to the influence of the “weight function”, see Proposition 3.14. Put $R(f) := R(n, f, \mathcal{F}, \delta)$ for short.

Let $\eta > 0$ be fixed and \hat{f}_η denote an η -approximate regularized ERM with regularization function $R(f)$ over class \mathcal{F} that is, $\hat{f}_\eta \in \mathcal{F}$ and

$$\widehat{\mathcal{E}}(\hat{f}_\eta) + R(\hat{f}_\eta) \leq \inf_{f \in \mathcal{F}} (\widehat{\mathcal{E}}(f) + R(f)) + \eta.$$

Prove that if (6.1) holds (but with the function $R(f)$ depending on f), then we have

$$\mathcal{E}(\hat{f}_\eta) \leq \inf_{f \in \mathcal{F}} (\mathcal{E}(f) + 2R(f)) + \eta.$$

6.2 The Azuma-McDiarmid inequality

For the first step of our programme, we will use the following concentration inequality, which is extremely useful and versatile.

**

Theorem 6.2 (Azuma-McDiarmid). *Let \mathcal{X} be a measurable space, and $f : \mathcal{X}^n \rightarrow \mathbb{R}$ a measurable function such that*

$$\forall i \in \{1, \dots, n\}, \quad \forall (x_1, \dots, x_n) \in \mathcal{X}^n, \quad \forall x'_i \in \mathcal{X} : \\ |f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)| \leq 2c_i, \quad (\text{Stab})$$

for some positive constants (c_1, \dots, c_n) .

Let (X_1, \dots, X_n) be a independent family of random variables taking values in \mathcal{X} (not necessarily identically distributed), then $f(X_1, \dots, X_n)$ is a sub-Gaussian variable with parameter $\sum_{i=1}^n c_i^2$, so that in particular

$$\mathbb{P}[f(X_1, \dots, X_n) > \mathbb{E}[f(X_1, \dots, X_n)] + t] \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n c_i^2}\right). \quad (6.4)$$

(In particular, if all constants c_i are equal to c , the bound is $\exp(-t^2/(2nc^2))$.)

To prove the above result we will first establish the following:

Theorem 6.3 (Bounded increment martingale inequality, Azuma). *Let $(M_k)_{k \geq 0}$ be a real-valued martingale with respect to a filtration $(\mathcal{F}_k)_{k \geq 0}$ (with $M_0 = 0$). Put $\Delta_k := M_k - M_{k-1}$, $k \geq 1$. Let n be a positive integer and assume $\Delta_k \in [A_k, B_k]$ holds a.s., where A_k, B_k are \mathcal{F}_{k-1} -measurable variables and $|B_k - A_k| \leq 2c_k$ a.s. for some constant c_k , for $k = 1, \dots, n$. Then M_n is a sub-Gaussian variable with parameter $\sum_{i=1}^n c_i^2$.*

Proof. Observe that $\mathbb{E}[\Delta_n | \mathcal{F}_{n-1}] = 0$ since M_n is a martingale. Furthermore, since $\Delta_n \in [A_n, B_n]$, we can apply Proposition 3.8 (Hoeffding's inequality in exponential form, for 1 variable) in conditional expectation to conclude that for any $\lambda \in \mathbb{R}$:

$$\mathbb{E}[\exp(\lambda \Delta_n) | \mathcal{F}_{n-1}] = \mathbb{E}[\exp(\lambda(\Delta_n - \mathbb{E}[\Delta_n | \mathcal{F}_{n-1}])) | \mathcal{F}_{n-1}] \leq \exp\left(\frac{\lambda^2 c_n^2}{2}\right).$$

Thus

$$\begin{aligned} \mathbb{E}[\exp(\lambda M_n)] &= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n \Delta_i\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n \Delta_i\right) \middle| \mathcal{F}_{n-1}\right]\right] \\ &= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} \Delta_i\right) \mathbb{E}\left[\exp(\lambda \Delta_n) \middle| \mathcal{F}_{n-1}\right]\right] \\ &\leq \exp\left(\frac{\lambda^2 c_n^2}{2}\right) \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} \Delta_i\right)\right], \end{aligned}$$

and we obtain the conclusion by straightforward recursion. \square

Proof of Theorem 6.2. Define the filtration $\mathcal{F}_i = \mathfrak{G}(X_1, \dots, X_i)$, $0 \leq i \leq n$, and define for $i = 1, \dots, n$ the martingale $M_i := \mathbb{E}[f(X_1, \dots, X_n) | \mathcal{F}_i] - \mathbb{E}[f(X_1, \dots, X_n)]$, and its increments $\Delta_i := \mathbb{E}[f(X_1, \dots, X_n) | \mathcal{F}_i] - \mathbb{E}[f(X_1, \dots, X_n) | \mathcal{F}_{i-1}]$. Let us prove that Δ_i satisfies the boundedness assumption of Theorem 6.3. First, because (X_1, \dots, X_n) are independent, conditional expectation conditional to (X_1, \dots, X_i) is the same as expectation with respect to (X_{i+1}, \dots, X_n) , thus

$$\mathbb{E}[f(X_1, \dots, X_n) | \mathcal{F}_i] = \int f(X_1, \dots, X_i, x_{i+1}, \dots, x_n) P(dx_{i+1}, \dots, dx_n).$$

Therefore

$$\begin{aligned} |\Delta_i| &= |\mathbb{E}[f(X_1, \dots, X_n) | \mathcal{F}_i] - \mathbb{E}[f(X_1, \dots, X_n) | \mathcal{F}_{i-1}]| \\ &\leq \int |f(X_1, \dots, X_{i-1}, X_i, x_{i+1}, \dots, x_n) \\ &\quad - f(X_1, \dots, X_{i-1}, x_i, x_{i+1}, \dots, x_n)| P(dx_i, dx_{i+1}, \dots, dx_n) \\ &\leq 2c_i, \end{aligned}$$

by assumption (Stab). We conclude that $M_n = f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]$ is a sub-Gaussian variable with parameter $\sum_{i=1}^n c_i^2$, which is the desired conclusion. \square

Let us now apply this result to the variables introduced in (6.2),(6.3):

**

Proposition 6.4. *Consider a learning setting with ℓ a bounded loss function taking values in $[0, B]$, a class \mathcal{F} of decision functions, and consider the random variables $Z_{\mathcal{F}}^{|\cdot|}, Z_{\mathcal{F}}^+, Z_{\mathcal{F}}^-$ defined by (6.2),(6.3). Denoting $Z_{\mathcal{F}}^\bullet$ either of these variables, it holds that $Z_{\mathcal{F}}^\bullet$ is sub-Gaussian with parameter $B^2/(4n)$, and thus in particular*

$$\mathbb{P}[Z_{\mathcal{F}}^\bullet \geq \mathbb{E}[Z_{\mathcal{F}}^\bullet] + t] \leq \exp\left(-\frac{2nt^2}{B^2}\right). \quad (6.5)$$

Note. If \mathcal{F} is a singleton $\{f\}$, then $\mathbb{E}[Z_{\mathcal{F}}^+] = \mathbb{E}[Z_{\mathcal{F}}^-] = 0$, and we recover Hoeffding's inequality as a particular case.

Proof. We check that the variables $Z_{\mathcal{F}}^{|\cdot|}, Z_{\mathcal{F}}^+, Z_{\mathcal{F}}^-$ satisfy the (Stab). Note that in general, if $(z_i)_{i \in I}$ and $(z'_i)_{i \in I}$ are families of real numbers in \mathbb{R}^I , for some index set I , it holds

$$\sup_{i \in I} (z_i - z'_i) \geq \sup_{i \in I} \left(z_i + \inf_{j \in I} (-z'_j) \right) = \sup_{i \in I} \left(z_i - \sup_{j \in I} z'_j \right) = \sup_{i \in I} z_i - \sup_{i \in I} z'_i. \quad (6.6)$$

Let us consider $Z_{\mathcal{F}}^+(S_n)$ as a function of the i.i.d. sample $S_n = ((X_1, Y_1), \dots, (X_n, Y_n))$. Consider a sample $\tilde{S}_n^{(i)}$ obtained with replacing (X_i, Y_i) by (X'_i, Y'_i) in S_n . Using (6.6), we obtain

$$\begin{aligned} Z_{\mathcal{F}}^+(S_n) - Z_{\mathcal{F}}^+(\tilde{S}_n^{(i)}) &= \sup_{f \in \mathcal{F}} \left(\widehat{\mathcal{E}}(f, S_n) - \mathcal{E}(f) \right) - \sup_{f \in \mathcal{F}} \left(\widehat{\mathcal{E}}(f, \tilde{S}_n^{(i)}) - \mathcal{E}(f) \right) \\ &\leq \sup_{f \in \mathcal{F}} \left(\widehat{\mathcal{E}}(f, S_n) - \widehat{\mathcal{E}}(f, \tilde{S}_n^{(i)}) \right) \\ &= \sup_{f \in \mathcal{F}} \frac{1}{n} \left(\ell(f(X_i), Y_i) - \ell(f(X'_i), Y'_i) \right) \\ &\leq \frac{B}{n}, \end{aligned}$$

so that (Stab) is satisfied with $c_i = \frac{B}{2n}$. We conclude by applying Theorem (6.2). The case of the other variables $Z_{\mathcal{F}}^{| \cdot |}, Z_{\mathcal{F}}^-$ is similar. \square

6.3 Rademacher complexity

We go to to the second step of our program: bounding the expectation of the variables $Z_{\mathcal{F}}^\bullet$.

Theorem 6.5 (Symmetrization principle). *We consider a standard learning setting, a class \mathcal{F} of decision functions, and consider the random variables $Z_{\mathcal{F}}^{| \cdot |}, Z_{\mathcal{F}}^+, Z_{\mathcal{F}}^-$ defined by (6.2),(6.3) based on an i.i.d. sample S_n of size n .*

Let $\sigma_1, \dots, \sigma_n$ be i.i.d. variables with values in $\{-1, 1\}$, independent from S_n , and such that $\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = 1/2$ (so-called “Rademacher variables”). Then it holds that

$$\mathbb{E}_{S_n} [Z_{\mathcal{F}}^+] = \mathbb{E}_{S_n} \left[\sup_{f \in \mathcal{F}} (\mathcal{E}(f) - \widehat{\mathcal{E}}(f, S_n)) \right] \leq \frac{2}{n} \mathbb{E}_{S_n, (\sigma_i)_{1 \leq i \leq n}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i \ell(f(X_i), Y_i) \right]. \quad (6.7)$$

The same inequality as above holds for $\mathbb{E}_{S_n} [Z_{\mathcal{F}}^+]$, while

$$\mathbb{E}_{S_n} [Z_{\mathcal{F}}^{| \cdot |}] = \mathbb{E}_{S_n} \left[\sup_{f \in \mathcal{F}} |\mathcal{E}(f) - \widehat{\mathcal{E}}(f, S_n)| \right] \leq \frac{2}{n} \mathbb{E}_{S_n, (\sigma_i)_{1 \leq i \leq n}} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i \ell(f(X_i), Y_i) \right| \right]. \quad (6.8)$$

**

Definition 6.6. Let \mathcal{W} be a measurable space and \mathcal{G} a set of measurable functions $\mathcal{W} \rightarrow \mathbb{R}$; S_n an i.i.d. sample of variables $(W_i)_{1 \leq i \leq n}$ of distribution P , and $\sigma :=$

$(\sigma_1, \dots, \sigma_n)$ a family of i.i.d. Rademacher variables, independent of S_n . Then the quantity

$$\mathcal{R}_{P,n}(\mathcal{G}) := \mathbb{E}_{S_n, \sigma} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^n \sigma_i g(W_i) \right] \quad (6.9)$$

is called *Rademacher complexity* of the class \mathcal{G} .

We introduce similarly

$$\mathcal{R}_{P,n}^{|\cdot|}(\mathcal{G}) := \mathbb{E}_{S_n, \sigma} \left[\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n \sigma_i g(W_i) \right| \right]. \quad (6.10)$$

◇

Note: geometrical interpretation. For a fixed sample $\mathbf{W} = (W_1, \dots, W_n) \in \mathcal{W}^n$, denote $\mathcal{G}(\mathbf{W}) = \{(g(W_1), \dots, g(W_n)), g \in \mathcal{G}\} \subseteq \mathbb{R}^n$. Then $\sup_{g \in \mathcal{G}} (\sum_{i=1}^n \sigma_i g(W_i)) = \sup_{u \in \mathcal{G}(\mathbf{W})} \langle \sigma, u \rangle$, and

$$\begin{aligned} \frac{2}{\sqrt{n}} \mathbb{E}_{\sigma} \left[\sup_{g \in \mathcal{G}} \left(\sum_{i=1}^n \sigma_i g(W_i) \right) \right] &= \frac{2}{\sqrt{n}} \mathbb{E}_{\sigma} \left[\sup_{u \in \mathcal{G}(\mathbf{W})} \langle \sigma, u \rangle \right] \\ &= \frac{1}{\sqrt{n}} \left(\mathbb{E}_{\sigma} \left[\sup_{u \in \mathcal{G}(\mathbf{W})} \langle \sigma, u \rangle \right] + \mathbb{E}_{\sigma} \left[\sup_{u \in \mathcal{G}(\mathbf{W})} \langle -\sigma, u \rangle \right] \right) \\ &= \mathbb{E}_{\sigma} \left[\sup_{u \in \mathcal{G}(\mathbf{W})} \frac{\langle \sigma, u \rangle}{\|\sigma\|} - \inf_{u \in \mathcal{G}(\mathbf{W})} \frac{\langle \sigma, u \rangle}{\|\sigma\|} \right]. \end{aligned}$$

This can be interpreted as the averaged maximal “width” of the set $\mathcal{G}(\mathbf{W})$ projected in the direction of the random Rademacher vector σ . Hence the above quantities are also known as *Rademacher widths*, which play an important role in high-dimensional geometry.

With this definition and notation, we can rewrite (6.7) and (6.8) as:

$$\mathbb{E}_{S_n} [Z_{\mathcal{F}}^{\pm}] = \mathbb{E}_{S_n} \left[\sup_{f \in \mathcal{F}} (\mathcal{E}(f) - \widehat{\mathcal{E}}(f, S_n)) \right] \leq \frac{2}{n} \mathcal{R}_{P,n}(\ell \circ \mathcal{F}), \quad (6.11)$$

$$\mathbb{E}_{S_n} [Z_{\mathcal{F}}^{|\cdot|}] = \mathbb{E}_{S_n} \left[\sup_{f \in \mathcal{F}} |\mathcal{E}(f) - \widehat{\mathcal{E}}(f, S_n)| \right] \leq \frac{2}{n} \mathcal{R}_{P,n}^{|\cdot|}(\ell \circ \mathcal{F}), \quad (6.12)$$

where

$$\ell \circ \mathcal{F} := \{g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}, (x, y) \mapsto \ell(f(x), y), f \in \mathcal{F}\}. \quad (6.13)$$

Proof of Theorem 6.5. The first step of the symmetrization principle is to use the fact that for any *fixed* decision function f , it holds $\mathbb{E}_{S_n} [\widehat{\mathcal{E}}(f, S_n)] = \mathcal{E}(f)$. We now introduce

a (virtual) second sample $S'_n = ((X'_i, Y'_i))_{1 \leq i \leq n}$ independent of S_n and having the same distribution as S_n (sometimes called “ghost sample” or “independent copy of S_n ”), and replace $\mathcal{E}(f) = \mathbb{E}_{S_n}[\widehat{\mathcal{E}}(f, S_n)]$:

$$\begin{aligned}
\mathbb{E}_{S_n} \left[\sup_{f \in \mathcal{F}} \left(\widehat{\mathcal{E}}(f, S_n) - \mathcal{E}(f) \right) \right] &= \mathbb{E}_{S_n} \left[\sup_{f \in \mathcal{F}} \left(\widehat{\mathcal{E}}(f, S_n) - \mathbb{E}_{S'_n} \left[\widehat{\mathcal{E}}(f, S'_n) \right] \right) \right] \\
&= \mathbb{E}_{S_n} \left[\sup_{f \in \mathcal{F}} \mathbb{E}_{S'_n} \left[\left(\widehat{\mathcal{E}}(f, S_n) - \widehat{\mathcal{E}}(f, S'_n) \right) \right] \right] \\
&\leq \mathbb{E}_{S_n, S'_n} \left[\sup_{f \in \mathcal{F}} \left(\widehat{\mathcal{E}}(f, S_n) - \widehat{\mathcal{E}}(f, S'_n) \right) \right] \\
&= \mathbb{E}_{S_n, S'_n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left(\ell(f(X_i), Y_i) - \ell(f(X'_i), Y'_i) \right) \right] \\
&= \frac{1}{n} \mathbb{E}_{S_n, S'_n} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^n \left(g(W_i) - g(W'_i) \right) \right],
\end{aligned}$$

where for the inequality we have used $\sup_{t \in T} \mathbb{E}[U_t] \leq \mathbb{E}[\sup_{t \in T} U_t]$ for a family of real-valued variables $(U_t)_{t \in T}$; and we have used the notation $\mathcal{G} := \ell \circ \mathcal{F}$ as defined in (6.13) and $W_i := (X_i, Y_i), W'_i = (X'_i, Y'_i)$ for shortening.

The second step is based on the observation that the distribution of S_n, S'_n is unchanged if we swap W_i and W'_i between the two samples. Hence the above double expectation remains unchanged by this operation, which flips the sign of the i -th term in the sum inside the expectation. Now, given arbitrary *fixed* signs $(\sigma_i)_{1 \leq i \leq n} = \boldsymbol{\sigma} \in \{-1, 1\}^n$, consider swapping W_i and W'_i if $\sigma_i = -1$ and leave them alone if $\sigma_i = 1$, again the expectation is unchanged while the sign of the i -th term in the sum inside of the expectation is multiplied by σ_i . Thus

$$\forall \boldsymbol{\sigma} \in \{-1, 1\}^n : \mathbb{E}_{S_n, S'_n} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^n \left(g(W_i) - g(W'_i) \right) \right] = \mathbb{E}_{S_n, S'_n} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^n \sigma_i \left(g(W_i) - g(W'_i) \right) \right].$$

Hence, the above quantity also remains the sum if we take an expectation over *random* signs $(\sigma_1, \dots, \sigma_n)$ having *any* joint distribution; it turns out that it is most fruitful to take i.i.d. Rademacher variables. Finally, we notice that

$$\begin{aligned}
\mathbb{E}_{S_n, S'_n, \boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^n \sigma_i \left(g(W_i) - g(W'_i) \right) \right] &\leq \mathbb{E}_{S_n, S'_n, \boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^n \sigma_i g(W_i) + \sup_{g \in \mathcal{G}} \sum_{i=1}^n -\sigma_i g(W'_i) \right] \\
&= \mathbb{E}_{S_n, \boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^n \sigma_i g(W_i) \right] + \mathbb{E}_{S'_n, \boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^n \sigma_i g(W'_i) \right] \\
&= 2\mathcal{R}_{P, n}(\mathcal{G}),
\end{aligned}$$

collecting the above inequalities yields the conclusion (the argument when introducing absolute values is entirely similar). \square

Taking stock, combining the results of Proposition 6.4 and (6.11)-(6.12) we obtain the following corollary:

**

Corollary 6.7. *Consider a learning setting with ℓ a bounded loss function taking values in $[0, B]$, a class \mathcal{F} of decision functions, and an i.i.d. sample S_n of size n .*

For any fixed $\delta \in (0, 1)$, each of the following inequalities holds with probability at least $1 - \delta$ over the draw of S_n :

$$\sup_{f \in \mathcal{F}} (\widehat{\mathcal{E}}(f) - \mathcal{E}(f)) \leq \frac{2}{n} \mathcal{R}_{P,n}(\ell \circ \mathcal{F}) + B \sqrt{\frac{\log \delta^{-1}}{2n}}; \quad (6.14)$$

$$\sup_{f \in \mathcal{F}} (\mathcal{E}(f) - \widehat{\mathcal{E}}(f)) \leq \frac{2}{n} \mathcal{R}_{P,n}(\ell \circ \mathcal{F}) + B \sqrt{\frac{\log \delta^{-1}}{2n}}; \quad (6.15)$$

$$\sup_{f \in \mathcal{F}} |\widehat{\mathcal{E}}(f) - \mathcal{E}(f)| \leq \frac{2}{n} \mathcal{R}_{P,n}^{|\cdot|}(\ell \circ \mathcal{F}) + B \sqrt{\frac{\log \delta^{-1}}{2n}}. \quad (6.16)$$

6.4 Properties of the Rademacher complexity

The following proposition gathers some simple but fundamental properties of the Rademacher complexity.

Proposition 6.8. *Let \mathcal{W} a measurable space and P a probability distribution on \mathcal{W} . In what follows, \mathcal{F} and \mathcal{G} are sets of measurable functions $\mathcal{W} \rightarrow \mathbb{R}$.*

(a) $\mathcal{R}_{P,n}^{|\cdot|}(\mathcal{F}) = \mathcal{R}_{P,n}(\mathcal{F} \cup (-\mathcal{F}))$. In particular if \mathcal{F} is symmetric around 0 then $\mathcal{R}_{P,n}^{|\cdot|}(\mathcal{F}) = \mathcal{R}_{P,n}(\mathcal{F})$.

(b) If h is a fixed measurable function $\mathcal{W} \rightarrow \mathbb{R}$, then $\mathcal{R}_{P,n}(\{h\}) = 0$ while $\mathcal{R}_{P,n}^{|\cdot|}(\{h\}) \leq \sqrt{n} \|h\|_{L^2(P)}$.

(c) Let $a \in \mathbb{R}$ be fixed. Then $\mathcal{R}_{P,n}(a\mathcal{F}) = |a| \mathcal{R}_{P,n}(\mathcal{F})$.

(d) $\mathcal{R}_{P,n}(\mathcal{F} + \mathcal{G}) = \mathcal{R}_{P,n}(\mathcal{F}) + \mathcal{R}_{P,n}(\mathcal{G})$.

(e) $\mathcal{R}_{P,n}(\text{Conv}(\mathcal{F})) = \mathcal{R}_{P,n}(\mathcal{F})$, where $\text{Conv}(\mathcal{F})$ is the set of finite convex combinations of elements of \mathcal{F} .

Proof. Point (a) is straightforward and left as an exercise, as is the first point of point (b).

For the second part of point (b), by Jensen's inequality:

$$\begin{aligned}
\mathcal{R}_{P,n}^{|\cdot|}(\{h\}) &= \mathbb{E}_{S_n, \sigma} \left[\left| \sum_{i=1}^n \sigma_i h(W_i) \right| \right] \\
&\leq \mathbb{E}_{S_n, \sigma} \left[\left(\sum_{i=1}^n \sigma_i h(W_i) \right)^2 \right]^{\frac{1}{2}} \\
&\leq \mathbb{E}_{S_n, \sigma} \left[\sum_{i=1}^n \sigma_i^2 h(W_i)^2 \right]^{\frac{1}{2}} \\
&= \sqrt{n} \|h\|_{L^2(P)}.
\end{aligned}$$

For point (c) we have by symmetry of the distribution of the vector or random signs σ :

$$\begin{aligned}
\mathcal{R}_{P,n}(a\mathcal{F}) &= \mathbb{E}_{S_n, \sigma} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i a f(W_i) \right] \\
&= \mathbb{E}_{S_n, \sigma} \left[\sup_{f \in \mathcal{F}} |a| \sum_{i=1}^n \sigma_i f(W_i) \right] \\
&= |a| \mathcal{R}_{P,n}(a\mathcal{F}).
\end{aligned}$$

For point (d):

$$\begin{aligned}
\mathcal{R}_{P,n}(\mathcal{F} + \mathcal{G}) &= \mathbb{E}_{S_n, \sigma} \left[\sup_{f \in \mathcal{F}, g \in \mathcal{G}} \sum_{i=1}^n \sigma_i (f(W_i) + g(W_i)) \right] \\
&= \mathbb{E}_{S_n, \sigma} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i f(W_i) + \sup_{g \in \mathcal{G}} \sum_{i=1}^n \sigma_i g(W_i) \right] \\
&= \mathcal{R}_{P,n}(\mathcal{F}) + \mathcal{R}_{P,n}(\mathcal{G}).
\end{aligned}$$

For point (e): let us denote $\text{Conv}_2(\mathcal{F}) := \{\lambda f + (1 - \lambda)g; \quad f, g \in \mathcal{F}; \lambda \in [0, 1]\}$ the set of 2-points convex combinations of elements of \mathcal{F} . It holds

$$\begin{aligned}
\mathcal{R}_{P,n}(\text{Conv}_2(\mathcal{F})) &= \mathbb{E}_{S_n, \sigma} \left[\sup_{\lambda \in [0,1]} \sup_{f, g \in \mathcal{F}} \sum_{i=1}^n \sigma_i (\lambda f(W_i) + (1 - \lambda)g(W_i)) \right] \\
&= \mathbb{E}_{S_n, \sigma} \left[\sup_{\lambda \in [0,1]} \left(\lambda \sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i f(W_i) + (1 - \lambda) \sup_{g \in \mathcal{F}} \sum_{i=1}^n \sigma_i g(W_i) \right) \right] \\
&= \mathbb{E}_{S_n, \sigma} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i \lambda f(W_i) \right] \\
&= \mathcal{R}_{P,n}(\mathcal{F}).
\end{aligned}$$

By straightforward recursion it holds $\mathcal{R}_{P,n}(\text{Conv}_{2^k}(\mathcal{F})) = \mathcal{R}_{P,n}(\mathcal{F})$ for any integer $k \geq 0$, and finally since $\text{Conv}(\mathcal{F}) = \bigcup_{k \geq 0} \text{Conv}_{2^k}(\mathcal{F})$ we obtain the result by monotone convergence. \square

The following property is extremely useful in learning theory.

**

Proposition 6.9 (Lipschitz comparison principle). *Consider a standard learning setting with $\tilde{\mathcal{Y}} \subseteq \mathbb{R}$ and a loss function $\ell : \tilde{\mathcal{Y}} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ such that for any fixed $y \in \mathcal{Y}$, the function $\ell(\cdot, y) : \tilde{\mathcal{Y}} \rightarrow \mathbb{R}_+$ is L -Lipschitz. Then it holds for any set \mathcal{F} of measurable functions from \mathcal{X} to \mathbb{R} :*

$$\mathcal{R}_{P,n}(\ell \circ \mathcal{F}) \leq L \mathcal{R}_{P,n}(\mathcal{F}). \quad (6.17)$$

The proof hinges on the following lemma:

Lemma 6.10. *Let $(A_t)_{t \in I}$, $(B_t)_{t \in I}$ be families of real numbers indexed by a countable set I , and $\gamma : \mathbb{R} \rightarrow \mathbb{R}$ a L -Lipschitz function. Let σ be a single Rademacher variable (random sign). Then*

$$\mathbb{E} \left[\sup_{t \in I} (A_t + \sigma \gamma(B_t)) \right] \leq \mathbb{E} \left[\sup_{t \in I} (A_t + L B_t) \right]. \quad (6.18)$$

Proof. We have

$$\begin{aligned} \mathbb{E} \left[\sup_{t \in I} (A_t + \sigma \gamma(B_t)) \right] &= \frac{1}{2} \sup_{t \in I} (A_t + \gamma(B_t)) + \sup_{t \in I} (A_t - \gamma(B_t)) \\ &= \frac{1}{2} \sup_{t, t' \in I} (A_t + A_{t'} + \gamma(B_t) - \gamma(B_{t'})) \\ &\leq \frac{1}{2} \sup_{t, t' \in I} (A_t + A_{t'} + L|B_t - B_{t'}|) \\ &= \frac{1}{2} \sup_{t, t' \in I} (A_t + A_{t'} + L(B_t - B_{t'})) \\ &= \frac{1}{2} \sup_{t \in I} (A_t + L B_t) + \sup_{t \in I} (A_t - L B_t) \\ &= \mathbb{E} \left[\sup_{t \in I} (A_t + L \sigma B_t) \right]. \end{aligned}$$

Note that the “magic” happens in the equality just after the Lipschitz inequality. By symmetry between t, t' we can remove the absolute value! It is worth pausing to think about it. \square

Proof of Proposition 6.9. Let n be fixed: we will prove by recursion the following property (for $m \leq n$):

$$H(m) : \mathcal{R}_{P,n}(\ell \circ \mathcal{F}) \leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left(L \sum_{i=1}^m \sigma_i f(X_i) + \sum_{i=m+1}^n \sigma_i \ell(f(X_i), Y_i) \right) \right],$$

where the expectation is over S_n and σ . Observe that $H(0)$ is obvious (it is the definition of $\mathcal{R}_{P,n}(\ell \circ \mathcal{F})$) and $H(n)$ is what we want to prove. Assuming $H(m-1)$ holds for $1 \leq m \leq n$, put $A_f := L \sum_{i=1}^{m-1} \sigma_i f(X_i) + \sum_{i=m+1}^n \sigma_i \ell(f(X_i), Y_i)$ and $B_f := f(X_m)$, then $H(m-1)$ reads

$$\mathcal{R}_{P,n}(\ell \circ \mathcal{F}) \leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} (A_f + \sigma_m \ell(B_f, Y_m)) \right].$$

Performing the expectation over σ_m first, conditionally to the other random variables (namely S_n and the signs $(\sigma_i)_{i \neq m}$), since A_f and B_f do not depend on σ_m they can be considered constants in this conditional expectation, and by independence from the rest σ_m is still a random sign conditionally to the rest. We can therefore apply Lemma 6.10, (with $\gamma(\cdot) = \ell(\cdot, Y_m)$ considered as fixed since we argue conditionally to Y_m), which yields the conclusion. \square

6.5 Application to kernel methods

We start with the following important result for the Rademacher complexity of a ball in a rkhs.

Proposition 6.11. *Let k be a spsd kernel on a nonempty set \mathcal{X} , \mathcal{H} the associated rkhs and for $R \geq 0$,*

$$B_{\mathcal{H}}(R) := \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq R\}$$

the closed ball of radius R in \mathcal{H} centered at the origin. Then

$$\mathcal{R}_{P,n}^{|\cdot|}(B_{\mathcal{H}}(R)) \leq \sqrt{n} R \sqrt{\mathbb{E}_{X \sim P}[k(X, X)]}. \quad (6.19)$$

In particular, if $\sup_{x \in \mathcal{X}} k(x, x) \leq M^2 < \infty$, then for any distribution P :

$$\mathcal{R}_{P,n}^{|\cdot|}(B_{\mathcal{H}}(R)) \leq \sqrt{n} M R. \quad (6.20)$$

Proof. It holds, using the Cauchy-Schwarz then Jensen's inequality:

$$\begin{aligned}
\mathcal{R}_{P,n}^{| \cdot |}(B_{\mathcal{H}}(R)) &= \mathbb{E} \left[\sup_{f \in B_{\mathcal{H}}R} \left| \sum_{i=1}^n \sigma_i f(X_i) \right| \right] \\
&= \mathbb{E} \left[\sup_{f \in B_{\mathcal{H}}R} \left| \sum_{i=1}^n \sigma_i \langle f, k_{X_i} \rangle_{\mathcal{H}} \right| \right] \\
&= \mathbb{E} \left[\sup_{f \in B_{\mathcal{H}}R} \left| \left\langle f, \sum_{i=1}^n \sigma_i k_{X_i} \right\rangle_{\mathcal{H}} \right| \right] \\
&\leq \mathbb{E} \left[\sup_{f \in B_{\mathcal{H}}R} \|f\|_{\mathcal{H}} \left\| \sum_{i=1}^n \sigma_i k_{X_i} \right\|_{\mathcal{H}} \right] \\
&\leq R \mathbb{E} \left[\left\| \sum_{i=1}^n \sigma_i k_{X_i} \right\|_{\mathcal{H}}^2 \right]^{\frac{1}{2}} \\
&= R \mathbb{E}_{S_n} \mathbb{E}_{\sigma} \left[\sum_{i,j=1}^n \sigma_i \sigma_j k(X_i, X_j) \right]^{\frac{1}{2}} \\
&= R(n \mathbb{E}_{X \sim P} [k(X, X)])^{\frac{1}{2}}.
\end{aligned}$$

□

Notice in particular the following interesting fact: the Rademacher complexity of a rkhs ball depends on its radius, but not on the dimensionality (which might be infinite) *provided the kernel is bounded*. This has a number of interesting consequences.

* **Proposition 6.12.** *Let k be a spsd kernel on a nonempty set \mathcal{X} , with $\sup_{x \in \mathcal{X}} k(x, x) \leq M^2$; let \mathcal{H} the associated rkhs and $R > 0$ be fixed. Assume ℓ is a loss function (with prediction space $\tilde{\mathcal{Y}} = \mathbb{R}$) such that*

(a) *for all $y \in \mathcal{Y}$ and $\tilde{y} \in \mathbb{R}$ with $|\tilde{y}| \leq MR$ it holds: $0 \leq \ell(\tilde{y}, y) \leq B$.*

(b) *for all $y \in \mathcal{Y}$ and $(\tilde{y}, \tilde{y}') \in [-MR, MR]^2$, it holds: $|\ell(\tilde{y}, y) - \ell(\tilde{y}', y)| \leq L|\tilde{y} - \tilde{y}'|$.*

Let \hat{f}_n be an estimator acting on a sample S_n of size n and such that $\hat{f}_n \in B_{\mathcal{H}}(R)$ a.s. Then for any $\delta \in (0, 1)$, with probability larger than $1 - \delta$ over the draw of the i.i.d. sample S_n it holds:

$$\left| \mathcal{E}(\hat{f}_n) - \hat{\mathcal{E}}(\hat{f}_n) \right| \leq \left| \sup_{f \in B_{\mathcal{H}}(R)} \left(\mathcal{E}(f) - \hat{\mathcal{E}}(f) \right) \right| \leq \frac{1}{\sqrt{n}} \left(2LRM + B \sqrt{\frac{\log \delta^{-1}}{2}} \right). \quad (6.21)$$

Proof. Start by noticing that by the usual argument based on the reproducing property, any function $f \in B_{\mathcal{H}}(R)$ satisfies $|f(x)| = |\langle f, k_x \rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|k_x\|_{\mathcal{H}} \leq RM$. For this reason,

we may consider that the prediction space $\tilde{\mathcal{Y}}$ is $[-MR, MR]$. We then have

$$\begin{aligned}
|\mathcal{E}(\hat{f}_n) - \hat{\mathcal{E}}(\hat{f}_n)| &\leq \left| \sup_{f \in B_{\mathcal{H}}(R)} \left(\mathcal{E}(f) - \hat{\mathcal{E}}(f) \right) \right| \\
&\leq \frac{2}{n} \mathcal{R}_{P,n}^{\|\cdot\|}(\ell \circ B_{\mathcal{H}}(R)) + B \sqrt{\frac{\log \delta^{-1}}{2n}} \quad \text{with probability } \geq 1 - \delta, \\
&\quad \text{(Corollary 6.7), eq. (6.16)} \\
&\leq \frac{2L}{n} \mathcal{R}_{P,n}(B_{\mathcal{H}}(R)) + B \sqrt{\frac{\log \delta^{-1}}{2n}} \quad \text{(Proposition 6.9)} \\
&\leq \frac{2LRM}{\sqrt{n}} + B \sqrt{\frac{\log \delta^{-1}}{2n}} \quad \text{(Proposition 6.11)}.
\end{aligned}$$

□

* **Corollary 6.13.** *Consider the same setting as in Proposition 6.12, with the squared loss function $\ell(\tilde{y}, y) = (\tilde{y} - y)^2$, and $\mathcal{Y} = [-A, A]$ for some $A > 0$ (bounded regression: we assume that the label is always bounded by A in absolute value). Then for any $\delta \in (0, \frac{1}{2}]$, with probability larger than $1 - \delta$ it holds:*

$$|\mathcal{E}(\hat{f}_n) - \hat{\mathcal{E}}(\hat{f}_n)| \leq \left| \sup_{f \in B_{\mathcal{H}}(R)} \left(\mathcal{E}(f) - \hat{\mathcal{E}}(f) \right) \right| \leq \frac{C \sqrt{\log \delta^{-1}}}{\sqrt{n}}, \quad (6.22)$$

where $C := 6(A + RM)^2$.

Proof. We check that the assumptions on the loss function of Proposition 6.12 are satisfied with appropriate constants. For any \tilde{y}, y with $|\tilde{y}| \leq MR$, $|y| \leq A$ it holds $\ell(\tilde{y}, y) = (\tilde{y} - y)^2 \leq (A + MR)^2 =: B$ (satisfying loss boundedness assumption (a)), and additionally for any \tilde{y}' with $|\tilde{y}'| \leq MR$:

$$|\ell(\tilde{y}, y) - \ell(\tilde{y}', y)| = |(\tilde{y} - y)^2 - (\tilde{y}' - y)^2| = |\tilde{y} + \tilde{y}' - 2y| |\tilde{y} - \tilde{y}'| \leq 2(A + RM) |\tilde{y} - \tilde{y}'|,$$

so Lipschitz assumption (b) is satisfied with $L := 2\sqrt{B}$. We therefore have the high probability bound (6.21), we can further upper bound $2LRM$ by $4B$, and finally use that $4 \leq 8\sqrt{(\log \delta^{-1})/2}$, since $\delta \leq 1/2$ by assumption. □

We now consider the analysis of kernel ridge regression (regularized least squares ERM).

* **Proposition 6.14** (Oracle-type inequality for krr). *We consider the same assumptions as in Corollary 6.13: squared loss, bounded regression with labels bounded by $A > 0$ in absolute value, kernel bounded by M^2 . For $\lambda \in [0, M^2]$ define the kernel ridge regression (krr) estimator, based on sample S_n of size n , as*

$$\hat{f}_\lambda \in \underset{f \in \mathcal{H}}{\text{Arg Min}} \left(\hat{\mathcal{E}}(f) + \lambda \|f\|_{\mathcal{H}}^2 \right). \quad (6.23)$$

Then for any $\delta \in (0, \frac{1}{2}]$, with probability larger than $1 - \delta$ it holds:

$$\left(\mathcal{E}(\hat{f}_\lambda) + \lambda \|\hat{f}_\lambda\|_{\mathcal{H}}^2 \right) \leq \min_{f \in \mathcal{H}} \left(\mathcal{E}(f) + \lambda \|f\|_{\mathcal{H}}^2 \right) + c \frac{A^2 M^2}{\lambda \sqrt{n}} \sqrt{\log \delta^{-1}}, \quad (6.24)$$

where c is a numerical constant.

Proof. We start with noticing that the norm of \widehat{f}_λ must be bounded. Namely, by the definition (6.23) of the estimator, it must correspond to a lower objective function than the constant 0 function (denoted $\bar{0}$), thus

$$\begin{aligned}\|\widehat{f}_\lambda\|_{\mathcal{H}}^2 &\leq \lambda^{-1} \left(\widehat{\mathcal{E}}(f) + \lambda \|f\|_{\mathcal{H}}^2 \right) \\ &\leq \lambda^{-1} \left(\widehat{\mathcal{E}}(\bar{0}) + \lambda \|\bar{0}\|_{\mathcal{H}}^2 \right) \\ &= \frac{1}{\lambda n} \sum_{i=1}^n (Y_i - 0)^2 \\ &\leq \frac{A^2}{\lambda}.\end{aligned}$$

Therefore, $\widehat{f}_\lambda \in B_{\mathcal{H}}(R)$ with $R = \frac{A}{\sqrt{\lambda}}$. Applying Corollary 6.13, we get that (6.22) is satisfied with probability at least $1 - \delta$, in particular

$$\mathcal{E}(\widehat{f}_\lambda) \leq \widehat{\mathcal{E}}(\widehat{f}_\lambda) + \frac{C\sqrt{\log \delta^{-1}}}{\sqrt{n}}, \quad (6.25)$$

with $C = 6(A + RM)^2 = 6A^2(1 + M/\sqrt{\lambda})^2 \leq 24A^2M^2/\lambda$, since $M/\lambda \geq 1$ by assumption.

Let now $f_\lambda^* \in \text{Arg Min}_{f \in \mathcal{H}} \left(\mathcal{E}(f) + \lambda \|f\|_{\mathcal{H}}^2 \right)$. By an argument similar to above, it must hold $f_\lambda^* \in B_{\mathcal{H}}(R)$, and, provided (6.22) is satisfied:

$$\widehat{\mathcal{E}}(f_\lambda^*) \leq \mathcal{E}(f_\lambda^*) + \frac{C\sqrt{\log \delta^{-1}}}{\sqrt{n}}. \quad (6.26)$$

Now, using (6.25) and (6.26) as well as the definitions of \widehat{f}_λ , f_λ^* , we get (with probability $1 - \delta$ of the event (6.22) being satisfied):

$$\begin{aligned}\mathcal{E}(\widehat{f}_\lambda) + \lambda \|\widehat{f}_\lambda\|_{\mathcal{H}}^2 &\leq \widehat{\mathcal{E}}(\widehat{f}_\lambda) + \lambda \|\widehat{f}_\lambda\|_{\mathcal{H}}^2 + \frac{C\sqrt{\log \delta^{-1}}}{\sqrt{n}} \\ &\leq \widehat{\mathcal{E}}(f_\lambda^*) + \lambda \|f_\lambda^*\|_{\mathcal{H}}^2 + \frac{C\sqrt{\log \delta^{-1}}}{\sqrt{n}} \\ &\leq \mathcal{E}(f_\lambda^*) + \lambda \|f_\lambda^*\|_{\mathcal{H}}^2 + 2\frac{C\sqrt{\log \delta^{-1}}}{\sqrt{n}} \\ &\leq \mathcal{E}(\widehat{f}_\lambda) + \lambda \|\widehat{f}_\lambda\|_{\mathcal{H}}^2 + 48\frac{A^2M^2}{\lambda\sqrt{n}} \sqrt{\log \delta^{-1}}.\end{aligned}$$

□

We have as a consequence the following *universal consistency* result:

**

Corollary 6.15. *Under the same assumptions as for Corollary 6.14, assume additionally that \mathcal{X} is a compact topological space and that the kernel k is universal on \mathcal{X} . Let $(\lambda_n)_{n \geq 1}$ be a sequence of regularization parameters such that $\lambda_n \rightarrow 0$ and $\lambda_n \sqrt{n/\log n} \rightarrow \infty$, as $n \rightarrow \infty$. Then for any distribution P of the data, if $(X_i, Y_i)_{i \geq 1}$ is an i.i.d. sequence from P , and $\widehat{f}_{\lambda_n}^{(n)}$ denotes the krr estimator trained using sample $S_n = ((X_1, Y_1), \dots, (X_n, Y_n))$, with regularization parameter λ_n , it holds that*

$$\mathcal{E}(\widehat{f}_{\lambda_n}^{(n)}) \rightarrow \mathcal{E}^*,$$

as $n \rightarrow \infty$, a.s. and in probability.

Proof. Let $\delta_n = \frac{1}{n^2}$, and A_n denote the event (6.24) for the sample S_n and estimator $\widehat{f}_{\lambda_n}^{(n)}$. Since $\sum_{i=1}^n \mathbb{P}[A_n^c] \leq \sum_{n \geq 1} n^{-2} < \infty$, by the Borel-Cantelli lemma $P(\bigcap_{k \geq 1} \bigcup_{n \geq k} A_n^c) = 0$, i.e., for any $\omega \in \Omega$ there exists a (random) integer $n_0(\omega)$ such that $\omega \in A_n$ for all $n \geq n_0(\omega)$ – in other words the events A_n are satisfied for all $n \geq n_0(\omega)$.

Next, let $\varepsilon > 0$ be fixed; we establish that there exists $f_\varepsilon \in \mathcal{H}$ such that $\mathcal{E}(f_\varepsilon) \leq \mathcal{E}^* + \varepsilon$. Namely, since $\mathcal{Y} = [-A, A]$, $f^*(x) = \mathbb{E}[Y|X=x]$ (for regression with quadratic loss) also takes values in $[-A, A]$ and thus belongs to $L^2(X, P)$. Furthermore, we now that $\mathcal{E}(f) - \mathcal{E}^* = \mathbb{E}[(f(X) - f^*(X))^2] = \|f - f^*\|_{L^2(\mathcal{X}, P)}^2$. Since k is universal, \mathcal{H} is dense in $\mathcal{C}(X)$, which itself is dense in $L^2(\mathcal{X}, P)$. Hence we can find such an f_ε .

We now have for any $\omega \in \Omega$, $n \geq n_0(\omega)$, for any $\varepsilon > 0$ using (6.24)

$$\begin{aligned} \mathcal{E}(\widehat{f}_{\lambda_n}^{(n)}) &\leq \min_{f \in \mathcal{H}} \left(\mathcal{E}(f) + \lambda_n \|f\|_{\mathcal{H}}^2 \right) + c \frac{A^2 M^2}{\lambda_n \sqrt{n}} \sqrt{\log \delta_n^{-1}} \\ &\leq \mathcal{E}(f_\varepsilon) + \lambda_n \|f_\varepsilon\|_{\mathcal{H}}^2 + c \frac{A^2 M^2}{\lambda_n \sqrt{n}} \sqrt{\log \delta_n^{-1}} \\ &\leq \mathcal{E}^* + \varepsilon + \lambda_n \|f_\varepsilon\|_{\mathcal{H}}^2 + c \frac{A^2 M^2}{\lambda_n} \sqrt{\frac{2 \log n}{n}}, \end{aligned}$$

by the assumptions on λ_n we deduce $\limsup_n \mathcal{E}(\widehat{f}_{\lambda_n}^{(n)}) \leq \mathcal{E}^* + \varepsilon$ a.s. for any $\varepsilon > 0$, and get the conclusion. \square

6.6 Vapnik-Chervonenkis theory

In this section we deal specifically with the classification setting $\mathcal{Y} = \widetilde{\mathcal{Y}} = \{-1, 1\}$, and the 0-1 loss $\ell(\widetilde{y}, y) = \mathbf{1}\{\widetilde{y} \neq y\}$. If we use this loss function combined with a real-valued score function $f(x)$ and predict $\text{sign}(f(x))$, we see that the resulting loss $\ell(f, y) = \mathbf{1}\{\text{sign}(f) \neq y\}$ is not Lipschitz in f , and therefore the Lipschitz comparison principle (Proposition 6.10) does not apply in order to bound the Rademacher complexity. We need to find other means

to upper bound the complexity (observe that Corollary 6.7 still holds, as the loss function is bounded by $B = 1$).

**

Theorem 6.16. Consider a standard binary classification setting ($\mathcal{Y} = \tilde{\mathcal{Y}} = \{-1, 1\}$) with 0-1 loss. Let \mathcal{F} be a set of measurable classification functions $\mathcal{X} \rightarrow \{-1, 1\}$. For a fixed integer $n > 0$, introduce the n -evaluation functional

$$G : \mathcal{X}^n \times \mathcal{F} \rightarrow \{-1, 1\}^n; \quad ((x_1, \dots, x_n), f) \mapsto (f(x_1), \dots, f(x_n)). \quad (6.27)$$

For fixed $s_n^x = (x_1, \dots, x_n) \in \mathcal{X}^n$, denote $G(s_n^x, \mathcal{F}) := \{G(s_n^x, f), f \in \mathcal{F}\}$. Then it holds for any distribution P on \mathcal{X} , for S_n and i.i.d. sample from P :

$$\mathcal{R}_{p,n}(\ell \circ \mathcal{F}) \leq \sqrt{2n} \mathbb{E}_{S_n^x \sim P_X^{\otimes n}} \left[\sqrt{\log |G(S_n^x, \mathcal{F})|} \right]. \quad (6.28)$$

We start with the following important lemma.

Lemma 6.17. Let ξ, \dots, ξ_K be K sub-Gaussian variables with common parameter σ^2 , centered ($\mathbb{E}[\xi_i] = 0, i = 1, \dots, n$), not necessarily independent. Then

$$\mathbb{E} \left[\sup_{i=1, \dots, K} \xi_i \right] \leq \sigma \sqrt{2 \log K}.$$

Proof. Put $\gamma := \mathbb{E}[\sup_{i=1, \dots, K} \xi_i]$. Then for any $\lambda > 0$:

$$\begin{aligned} \exp(\lambda \gamma) &\leq \mathbb{E} \left[\exp(\lambda \sup_{i=1, \dots, K} \xi_i) \right] && \text{(Jensen)} \\ &= \mathbb{E} \left[\sup_{i=1, \dots, K} \exp(\lambda \xi_i) \right] \\ &\leq \sum_{i=1}^n \mathbb{E}[\exp(\lambda \xi_i)] \\ &\leq K \exp\left(\frac{\sigma^2 \lambda^2}{2}\right) && \text{(Sub-Gaussianity)}. \end{aligned}$$

We deduce $\gamma \leq \frac{\log K}{\lambda} + \lambda \sigma^2 / 2$, which gives the claim when choosing $\lambda = \sqrt{2 \log K} / \sigma$. \square

Proof of Theorem 6.16. For a sample $S_n = ((X_i, Y_i))_{1 \leq i \leq n}$ denote

$$\tilde{G}(S_n, f) = (\mathbf{1}\{f(X_1) \neq Y_1\}, \dots, \mathbf{1}\{f(X_n) \neq Y_n\}) \in \{0, 1\}^n,$$

and define $\tilde{G}(S_n, \mathcal{F}) := \left\{ \tilde{G}(S_n, f), f \in \mathcal{F} \right\}$. Observe that $|\tilde{G}(S_n, \mathcal{F})| = |G(S_n^x, \mathcal{F})|$. With this notation we have

$$\mathcal{R}_{p,n}(\ell \circ \mathcal{F}) = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X_i) \neq Y_i\} \right] = \mathbb{E} \left[\sup_{u \in \tilde{G}(S_n, \mathcal{F})} \sum_{i=1}^n \sigma_i u^{(i)} \right].$$

For a fixed $u \in \{0, 1\}^n$ let $\xi_u(\boldsymbol{\sigma}) := \sum_{i=1}^n \sigma_i u^{(i)}$. By Hoeffding's inequality (more precisely Proposition 3.8 and properties of sub-Gaussian variables), $\xi_u(\boldsymbol{\sigma})$ is centered and sub-Gaussian with parameter $\sigma^2 = \|u\|_1 \leq n$. Hence by Lemma 6.17, taking expectation with respect to $\boldsymbol{\sigma}$ first:

$$\begin{aligned} \mathcal{R}_{p,n}(\ell \circ \mathcal{F}) &= \mathbb{E}_{S_n, \boldsymbol{\sigma}} \left[\sup_{u \in \tilde{G}(S_n, \mathcal{F})} \xi_u(\boldsymbol{\sigma}) \right] \leq \mathbb{E}_{S_n} \left[\sqrt{2n \log |\tilde{G}(S_n, \mathcal{F})|} \right] \\ &= \mathbb{E}_{S_n} \left[\sqrt{2n \log |G(S_n^x, \mathcal{F})|} \right]. \end{aligned}$$

□

While the bound (6.28) is nice, it turns out that in most interesting cases we can upper bound $|G(s_n^x, \mathcal{F})|$ uniformly for any sample s_n^x , as a function of \mathcal{F} , and thus bound the expectation in (6.28) independently of the distribution P !

**

Definition 6.18 (Growth function). Let \mathcal{F} be a set of functions from \mathcal{X} to \mathcal{Y} . Define the *growth function* of \mathcal{F} , using the notation for the evaluation functional introduced in (6.27), as

$$\Gamma(\mathcal{F}, n) = \sup_{S_n \in \mathcal{X}^n} |G(S_n, \mathcal{F})|. \quad (6.29)$$

Observe that $\Gamma(\mathcal{F}, n) \leq 2^n$ always holds.

**

Theorem 6.19 (Vapnik/Sauer). Let \mathcal{F} be a set of functions from \mathcal{X} to $\{0, 1\}$. Define

$$d = \max\{n > 0 : \Gamma(\mathcal{F}, n) = 2^n\}. \quad (6.30)$$

Then it holds for any n :

$$\Gamma(\mathcal{F}, n) \leq \sum_{i=0}^{\min(d,n)} \binom{n}{i} \leq (n+1)^d. \quad (6.31)$$

The number d is called Vapnik-Chervonenkis (VC) dimension of the class of prediction functions \mathcal{F} . (Observe that since \mathcal{F} is a set of indicator functions, equivalently one can define the VC dimension of a family of subsets of \mathcal{X} .)

Proof. For a subset $A \subseteq \{0, 1\}^n$, and $I \subseteq \llbracket n \rrbracket := \{1, \dots, n\}$, let $(i_1, \dots, i_{|I|})$ be the ordered elements of I and denote $A_I \subseteq \{0, 1\}^{|I|}$ the projection of A on the coordinates of indices $(i_1, \dots, i_{|I|})$. For coherence, if $I = \emptyset$, we define $A_\emptyset = \emptyset$. Let us call a subset of indices I (possibly empty) *shattered* by A if $A_I = \{0, 1\}^{|I|}$ (we define $\{0, 1\}^0 = \emptyset$).

We will establish the following: for any $A \subseteq \{0, 1\}^n$,

$$|A| \leq \left| \left\{ I \subseteq \llbracket n \rrbracket : A_I = \{0, 1\}^{|I|} \right\} \right| \quad (6.32)$$

in words, the cardinality of A is upper bounded by the number of indices sets I shattered by A .

We prove this by recursion on n . For $n = 1$ it is true since $I = \emptyset$ is always shattered, and if $A = \{0, 1\}$ then $I = \{1\}$ is also shattered.

Assume the property is true for some $n \geq 1$, and let $A \subseteq \{0, 1\}^{n+1}$. Let $\tilde{A} \subseteq \{0, 1\}^n$ the sets of elements $\tilde{a} \in \{0, 1\}^n$ such that both $(\tilde{a}, 0)$ and $(\tilde{a}, 1)$ belong to A . Let also denote $A' := A_{\llbracket n \rrbracket}$. Then it holds that

$$|A| = |A'| + |\tilde{A}|. \quad (6.33)$$

By recursion, both A' and \tilde{A} satisfy (6.32). For $I \subseteq \llbracket n \rrbracket$, it holds $A'_I = A_I$. Hence

$$\begin{aligned} |A'| &\leq \left| \left\{ I \subseteq \llbracket n \rrbracket : A_I = \{0, 1\}^{|I|} \right\} \right| \\ &= \left| \left\{ I \subseteq \llbracket n+1 \rrbracket \text{ s.t. } (n+1) \notin I, A_I = \{0, 1\}^{|I|} \right\} \right|. \end{aligned} \quad (6.34)$$

On the other hand, if $I \subseteq \llbracket n \rrbracket$ is shattered by \tilde{A} , then by construction $I \cup \{n+1\}$ is shattered by A . Hence

$$\begin{aligned} |\tilde{A}| &\leq \left| \left\{ I \subseteq \llbracket n \rrbracket : A_{I \cup \{n+1\}} = \{0, 1\}^{|I|+1} \right\} \right| \\ &= \left| \left\{ I \subseteq \llbracket n+1 \rrbracket \text{ s.t. } (n+1) \in I, A_I = \{0, 1\}^{|I|+1} \right\} \right|. \end{aligned} \quad (6.35)$$

Noticing that the set of indices concerned in (6.34) and (6.35) are disjoint and putting back in (6.33), we obtain the property (6.32) for $(n+1)$.

We now apply property (6.32) for the set $A = G(S_n, \mathcal{F})$ where $S_n \in \mathcal{X}^n$ is arbitrary. By assumption (6.30) the largest possible cardinality of a shattered index set I (which determines a sub-sample of size $|I|$) is d . Hence

$$|G(S_n, \mathcal{F})| \leq |\{I \subset \llbracket n \rrbracket : |I| \leq d\}| = \sum_{i=0}^d \binom{n}{i}.$$

Taking a supremum over all possible $S_n \in \mathcal{X}^n$ yields the first inequality in (6.31).

Finally, note that $\binom{n}{i} \leq \binom{d}{i} n^i$ and use the binomial formula for the second inequality. \square

It is possible to get an exact or upper bound of the VC dimension (6.30), or possibly directly of the growth function (6.29) in certain cases. A first fundamental fact is the following bound on linear discrimination function classes.

**

Proposition 6.20. *Let $X = \mathbb{R}^d$ and $\mathcal{F} = \{x \mapsto \mathbf{1}\{\langle x, w \rangle > 0\}, w \in \mathbb{R}^d\}$ be the set of linear classifiers without offset (i.e. indicators of half-spaces whose boundary contain the origin). Then the VC dimension of \mathcal{F} is equal to d .*

Proof. First, we prove that the VC-dimension of the class \mathcal{F} of half-spaces with boundary going through the origin is at least d . For this, simply consider the d -uple S_d of points of \mathbb{R}^d formed by the canonical basis ($x_i = e_i, i = 1, \dots, d$). Let $I \subseteq \llbracket n \rrbracket$ be any index set, denote $I^c := \{1, \dots, n\} \setminus I$. Define the vector w_I as $w_I^{(i)} = 1$ if $i \in I$, and $w_I^{(i)} = -1$ if $i \notin I$. Then obviously, if $f_w(x) := \mathbf{1}\{\langle x, w \rangle > 0\}$, we have $f_{w_I}(S_d) = (\mathbf{1}\{i \in I\})_{1 \leq i \leq d}$. Since this works for any I , we have $G(S_d, \mathcal{F}) = \{0, 1\}^d$ and $|G(S_d, \mathcal{F})| = 2^d$.

Conversely, we prove that any family $S_{d+1} = (x_1, \dots, x_{d+1})$ of $(d+1)$ vectors in \mathbb{R}^d cannot be “shattered” by \mathcal{F} (i.e. $|G(S_{d+1}, \mathcal{F})| < 2^{d+1}$). Since we are in dimension d , there is a nontrivial linear combination of these vectors that vanishes: $\exists \boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{d+1}) \in \mathbb{R}^{d+1}$ such that $\boldsymbol{\lambda} \neq \mathbf{0}$ and $\sum_{i=1}^{d+1} \lambda_i x_i = 0$. Let $I := \{i \in \llbracket n \rrbracket : \lambda_i > 0\}$. Without loss of generality we can assume $I \neq \emptyset$ (otherwise replace $\boldsymbol{\lambda}$ by $-\boldsymbol{\lambda}$). Let w be a vector such that $f_w(x_i) > 0$ for all $i \in I$. Then

$$0 < \sum_{i \in I} \lambda_i f_w(x_i) = \left\langle w, \sum_{i \in I} \lambda_i x_i \right\rangle = \left\langle w, - \sum_{i \in I^c} \lambda_i x_i \right\rangle = \sum_{i \in I^c} (-\lambda_i) f_w(x_i).$$

Therefore there is at least one $i \in I^c$ such that $f_w(x_i) > 0$. It means that $(\mathbf{1}\{i \in I\})_{1 \leq i \leq d} \notin G(S_{d+1}, \mathcal{F})$, therefore $|G(S_{d+1}, \mathcal{F})| < 2^{d+1}$. \square

**

Corollary 6.21. *Let \mathcal{G} be a linear space of dimension d of real-valued functions on an arbitrary set \mathcal{X} . Then the VC-dimension of*

$$\mathcal{F} := \{x \mapsto \mathbf{1}\{g(x) > 0\}, g \in \mathcal{G}\}$$

is equal to d .

Proof. Let (g_1, \dots, g_d) be a basis of \mathcal{G} . Define the mapping $A(x) := (g_1(x), \dots, g_d(x)) \in \mathbb{R}^d$. Since $\mathcal{G} = \text{Span}\{g_1, \dots, g_d\}$, it holds that $\mathcal{F} := \{x \mapsto \mathbf{1}\{\langle w, A(x) \rangle > 0\}, w \in \mathbb{R}^d\}$. Therefore, any family of points of $S_k = (x_1, \dots, x_k)$ that is “shattered” by \mathcal{F} (i.e. $|G(S_k, \mathcal{F})| < 2^k$) implies that the family $(A(x_1), \dots, A(x_k))$ of elements of \mathbb{R}^d is shattered by linear classifiers without offset, hence from Proposition 6.20 it must be the case that $k \leq d$.

On the other hand, notice that $\text{Span}(A(\mathcal{X})) = \mathbb{R}^d$. If this was not the case, there would exist $u \in \mathbb{R}^d$, $u \neq 0$, such that $\langle u, g(x) \rangle = \sum_{i=1}^n u_i g_i(x) = 0$ for all $x \in \mathcal{X}$, contradicting that (g_1, \dots, g_d) are independent. We can therefore find $S_d = (x_1, \dots, x_d) \in \mathcal{X}^d$ such that $(A(x_1), \dots, A(x_d))$ are linearly independent vectors in \mathbb{R}^d . A family of independent vectors in \mathbb{R}^d is shattered by linear classifiers (repeat the argument of Proposition 6.20 after change of basis, i.e. choose $\tilde{w}_I = (M^t)^{-1}(\sum_{i \in I} e_i - \sum_{j \in I^c} e_j)$ for any subset $I \subseteq \llbracket n \rrbracket$, where M is the matrix of $(A(x_1), \dots, A(x_d))$). Hence S_d is shattered by \mathcal{F} . \square

An example of application of the previous corollary is the set of binary classifiers defined from polynomial score functions of degree up to k on an open set of \mathbb{R}^d has VC-dimension equal to $(k + 1)^d$.

Exercise 6.2. Let $X = \mathbb{R}^d$, f be a fixed real-valued function on \mathcal{X} and

$$\mathcal{F} = \{x \mapsto \mathbf{1}\{\langle x, w \rangle + f(x) > 0\}, w \in \mathbb{R}^d\}.$$

Prove that the VC dimension of \mathcal{F} is equal to d . *Hint: recycle the proof of Proposition 6.20 with appropriate changes.* Deduce that if \mathcal{G} is a linear space of dimension d of real-valued functions on an arbitrary set \mathcal{X} , and f a fixed real-valued function on \mathcal{X} , then the VC-dimension of

$$\mathcal{F} := \{x \mapsto \mathbf{1}\{g(x) + f(x) > 0\}, g \in \mathcal{G}\}$$

is equal to d .

Exercise 6.3. Let $X = \mathbb{R}^d$, f be a fixed real-valued function on \mathcal{X} and

$$\mathcal{F} = \{x \mapsto \mathbf{1}\{\langle x, w \rangle + a > 0\}, w \in \mathbb{R}^d, a \in \mathbb{R}\}$$

the set of linear classifiers with offset (i.e. indicators of affine half-spaces). Prove that the VC dimension of \mathcal{F} is equal to $d + 1$.

6.7 Application to artificial neural networks

Artificial feedforward neural networks (ANNs) can be described in the following way. The basic unit or “artificial neuron” consists of a function $f_w : \mathbb{R}^d \rightarrow \mathbb{R} : x \mapsto \varphi(\langle w, x \rangle)$, for some input dimension d , where w is called “activation weight vector” and φ is fixed *a priori* and called an “activation function”. The activation function is nonlinear and acts as a thresholding function, classical choices involve $\varphi(t) = \tanh(t)$ (generally called sigmoid function), $\varphi(t) = (t)_+$ (Rectified Linear Unit or “ReLU” function), and $\varphi(t) = \text{sign}(t)$ (in this case the function f_w is just a linear classifier).

A *layer* of the ANN consists in several parallel ANs with different parameters acting on the same input space. The number of the ANs in the k -th layer is n_k and the input of the k -th layer is the output of the $(k - 1)$ -th, so that the k -th layer is a (nonlinear) mapping $\mathbb{R}^{n_{k-1}} \rightarrow \mathbb{R}^{n_k}$. The input of the first layer is the x -data belonging to \mathbb{R}^d ($n_0 = d$). The output $x^{(k)}$ in \mathbb{R}^{n_k} of the k th layer is thus computed by the formula

$$x_i^{(k)} = \varphi(\langle w_i^{(k)}, x^{(k)} \rangle), \quad i = 1, \dots, n_k.$$

The final layer (say K) consists of a single neuron ($n_K = 1$) and outputs the prediction of the network.

Thus, an ANN is parametrized by $\sum_{i=1}^K n_i n_{i-1}$ real parameters corresponding to the weight vectors of each AN. It is possible to reduce this dimensionality by specifying that the activation weight vector of a given AN is restricted to a specific support of reduced size k (i.e. the weights outside of the support are 0). It means that this AN can only use the output of a specific (given) subset of size k of the ANs of the previous layer.

In this section, we will not explain how to construct ANNs from data, but give a rough analysis of their statistical complexity using a simplified model. We will consider the sign activation function, and thus assume that the k -th layer is in fact acting on $\{-1, 1\}^{n_{k-1}}$, we will also assume that the input space is binary, i.e. $\{-1, 1\}^d$ for simplicity. Finally we also assume that a constant output neuron (equal to 1) is added in each layer (thus providing the means to add an offset to the linear part of each AN), including on the input data.

Proposition 6.22. *Given a input space $\{-1, 1\}^d$, and using the sign activation function, it is possible to find a weight vector implementing the “OR” and “AND” functions on a specific subset I , that is, there exists (w_I, a) and (w'_I, a') in \mathbb{R}^{d+1} (the extra parameter is because we add a constant coordinate to the input which we treat as an offset) such that*

$$f_{w_I}(x) = \bigwedge_{i \in I} x_i; \quad f_{w'_I}(x) = \bigvee_{i \in I} x_i.$$

Proof. We take $w_I = w'_I$ with i -th coordinate equal to 1 if $i \in I$ and 0 else, and take $a = |I| - 1/2$, $a' = 1/2$. \square

Proposition 6.23. *Any boolean function $F : \{-1, 1\}^d \rightarrow \{-1, 1\}$ can be realized with a 2-layer ANN with sufficiently large first layer.*

Proof. Let $\xi_F := \{x \in \{-1, 1\}^d : F(x) = 1\}$. For each $x \in \xi_F$ we construct an AN in the first layer with weight $w_x = x$ and offset $a = d - \frac{1}{2}$. It can be checked that $f_{x,a}(t) = 1$ if $t = x$, and is -1 otherwise. The neuron in the second layer is just an OR of all the neurons of the first layer. \square

Proposition 6.23 states that a 2-layer ANN can already represent any boolean function, however the first layer can be of size up to 2^d to achieve this, which can be prohibitive. On the other hand, using Proposition 6.22 we see that we can implement a “boolean circuit” function (which comprises of applications of the elementary AND, OR and NOT gates along an evaluation tree) with an ANN having as many (nonconstant) neurons as gates in the boolean circuit, and with the number of layers equal to the depth of the evaluation tree.

To summarize, we get the following qualitative understanding:

- A “shallow” (with few layers) ANN with sign activation function can realize any boolean function, but their complexity (size) must be very large to do so (there are similar results for approximation of continuous functions using other activation functions).

- A “deep” network (with many layers) can approximate more efficiently (i.e. using less ANs) boolean functions that can be represented in short form as a composition of elementary boolean gates.

To make the second point more quantitative, we will analyze the complexity (in the sense of VC theory) of an ANN with architectural connection constraints, namely if each AN is restricted to use only a (fixed) subset of the previous layer as an input (this corresponds to its activation weight vector w being restricted to a fixed support of reduced size, see beginning of the section). We can represent such connection constraints abstractly as a graph $G = (V, E)$: each vertex of V of the graph represents either a single ANs or an individual coordinate of the input data in $\{-1, 1\}^d$, and the (directed) edges E represent the nonzero activation weights from one individual neuron of a given layer to the next.

Proposition 6.24. *Let $G = (V, E)$ be a graph representing the structure and connection constraints of an ANN, and \mathcal{F}_G be the set of boolean functions on $\{-1, 1\}^d$ that can be represented by an ANN following these constraints.*

Then it holds for $n \geq 1$:

$$\Gamma(\mathcal{F}_G, n) \leq (n + 1)^{|E|}, \quad (6.36)$$

and it follows that the VC dimension of \mathcal{F}_G is bounded by $c|E| \log |E|$, for c a numerical constant.

Proof. We will use the notion of growth function (6.29) when the output set of functions can be larger than $\{0, 1\}$ (the definition is unchanged). Let K be the number of layers in the ANN. For $k \leq K$, consider the set of functions $\mathcal{F}_{G,k} \subseteq \mathcal{F}(\{-1, 1\}^{n_{k-1}}, \{-1, 1\}^{n_k})$ obtained by only considering the k -th layer of the ANN under the structural constraints given by G .

Note that the input resp. output space of $\mathcal{F}_{G,k}$ is $\{-1, 1\}^{n_{k-1}}$ resp. $\{-1, 1\}^{n_k}$, given by the values of the number n_{k-1} resp n_k of ANs in the $(k - 1)$ -th resp, k -th layer. For AN number i of the k -th layer, $1 \leq i \leq n_k$, denote $\ell_{k,i}$ the number of incoming edges from the previous layer. To this AN is associated a linear classifier of weight $w_{k,i}$ of dimension $\ell_{k,i}$. Therefore from Proposition 6.20 and Theorem 6.30, if $\mathcal{F}_{k,i}$ is the set of linear classifiers that can be represented by this AN, we have

$$\Gamma(\mathcal{F}_{k,i}, n) \leq (n + 1)^{\ell_{k,i}}.$$

Since

$$\mathcal{F}_k = \{x \in \{-1, 1\}^{n_{k-1}} \mapsto (f_{k,i}(x))_{1 \leq i \leq n_k}, f_{k,i} \in \mathcal{F}_{k,i}\},$$

we deduce

$$\Gamma(\mathcal{F}_k, n) \leq \prod_{i=1}^{n_k} (n + 1)^{\ell_{k,i}} = (n + 1)^{\sum_{i=1}^{n_k} \ell_{k,i}}.$$

Finally it is easy to check that we have in general the “composition rule” for $\mathcal{F} \circ \mathcal{G} = f \circ g, f \in \mathcal{F}, g \in \mathcal{G}$:

$$\Gamma(\mathcal{F} \circ \mathcal{G}, n) \leq \Gamma(\mathcal{F}, n) \circ \Gamma(\mathcal{G}, n);$$

namely, for any n -uple S_n in the input space of \mathcal{G} , it holds

$$\begin{aligned}
|G(S_n, \mathcal{F} \circ \mathcal{G})| &= |\cup_{S'_n \in G(S_n, \mathcal{G})} G(S'_n, \mathcal{F})| \\
&\leq |G(S_n, \mathcal{G})| \max_{S'_n \in G(S_n, \mathcal{G})} |G(S'_n, \mathcal{F})| \\
&\leq \Gamma(\mathcal{G}, n) \Gamma(\mathcal{F}, n).
\end{aligned}$$

Since we have $\mathcal{F}_G = \mathcal{F}_K \circ \mathcal{F}_{K-1} \circ \dots \circ \mathcal{F}_1$, we get

$$\Gamma(\mathcal{F}_G, n) \leq \prod_{k=1}^K \Gamma(\mathcal{F}_k, n) \leq (n+1)^{\sum_{k=1}^K \sum_{i=1}^{n_k} \ell_{k,i}} = (n+1)^{|E|}.$$

□