ECE 543: Statistical Learning Theory

Bruce Hajek and Maxim Raginsky

Department of Electrical and Computer Engineering
and the Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Last updated March 18, 2021

Contents

Part 1

Preliminaries

CHAPTER 1

Introduction

1.1. A simple example: coin tossing

Let us start things off with a simple illustrative example. Suppose someone hands you a coin that has an unknown probability θ of coming up heads. You wish to determine this probability (coin bias) as accurately as possible by means of experimentation. Experimentation in this case amounts to repeatedly tossing the coin (this assumes, of course, that the bias of the coin on subsequent tosses does not change, but let's say you have no reason to believe otherwise). Let us denote the two possible outcomes of a single toss by 1 (for HEADS) and 0 (for TAILS). Thus, if you toss the coin n times, then you can record the outcomes as $X_1, \ldots, X_n$, where each $X_i \in \{0,1\}$ and $\mathbf{P}(X_i = 1) = \theta$ independently of all other X_j's. More succinctly, we can write our sequence of outcomes as $X^n \in \{0,1\}^n$, which is a *random* binary n-tuple. This is our *sample*.

What would be a reasonable estimate of θ? Well, by the Law of Large Numbers we know that, in a long sequence of independent coin tosses, the relative frequency of heads will eventually approach the true coin bias with high probability. So, without further ado you go ahead and estimate θ by

$$\widehat{\theta}_n(X^n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

(recall that each $X_i \in \{0,1\}$, so the sum in the above expression simply counts the number of times the coin came up HEADS). The notation $\widehat{\theta}_n(X^n)$ indicates the fact that the above estimate depends on the *sample size* n and on the entire sample X^n.

How accurate can this estimator be? To answer this question, let us fix an *accuracy parameter* $\varepsilon \in [0,1]$. Given θ and n, we can partition the entire set $\{0,1\}^n$ into two disjoint sets:

$$\mathsf{G}_{n,\varepsilon} := \left\{ x^n \in \{0,1\}^n : \left|\widehat{\theta}_n(x^n) - \theta\right| \le \varepsilon \right\}$$
$$\mathsf{B}_{n,\varepsilon} := \left\{ x^n \in \{0,1\}^n : \left|\widehat{\theta}_n(x^n) - \theta\right| > \varepsilon \right\}.$$

As the notation suggests, the n-tuples in $\mathsf{G}_{n,\varepsilon}$ are the "good ones:" if our random sequence of tosses X^n happens to land in $\mathsf{G}_{n,\varepsilon}$, then our estimate $\widehat{\theta}_n$ will differ from the true bias θ by at most ε in either direction. On the other hand, if X^n lands in $\mathsf{B}_{n,\varepsilon}$, then we will have no such luck. Of course, since we do not know θ, we have no way of telling whether X^n is in $\mathsf{G}_{n,\varepsilon}$ or in $\mathsf{B}_{n,\varepsilon}$. The best we can do is to compute the probability of a bad sample for each possible value of θ. This can be done using the so-called *Chernoff bound* [**HR90**]

$$\mathbf{P}^n_\theta(\mathsf{B}_{n,\varepsilon}) \equiv \mathbf{P}^n_\theta\left(\left|\widehat{\theta}_n(X^n) - \theta\right| > \varepsilon\right) \le 2e^{-2n\varepsilon^2} \tag{1.1}$$

(soon you will see where this comes from). Here, $\mathbf{P}_\theta^n$ denotes the distribution of the random sample X^n when the probability of heads on each toss is θ. Now, Eq. (1.1) says two things: (1) For any desired accuracy ε, probability of getting a bad sample decreases *exponentially* with sample size n. (2) In order to guarantee that the probability of a bad sample is at most δ, you will need

$$n \ge \frac{1}{2\varepsilon^2} \log\left(\frac{2}{\delta}\right)$$

coin tosses[1]. Thus, if you toss the coin at least this many times, then, no matter what θ is, you can assert with *confidence* at least $1-\delta$ that θ is somewhere between $\widehat{\theta}_n - \varepsilon$ and $\widehat{\theta}_n + \varepsilon$. This leads to the following

OBSERVATION 1.1. *For any true value θ of the coin bias,*

$$n(\varepsilon,\delta) := \left\lceil \frac{1}{2\varepsilon^2} \log\left(\frac{2}{\delta}\right) \right\rceil$$

tosses suffice to guarantee with confidence $1-\delta$ that the estimate $\widehat{\theta}_n$ has accuracy ε.

In view of this observation, we can call the function $(\varepsilon,\delta) \mapsto n(\varepsilon,\delta)$ the *sample complexity* of coin tossing.

This simple example illustrates the essence of statistical learning theory: We wish to learn something about a phenomenon of interest, and we do so by observing random samples of some quantity pertaining to the phenomenon. There are two basic questions we can ask:

(1) How large of a sample do we need to achieve a given accuracy with a given confidence?
(2) How efficient can our learning algorithm be?

Statistical learning theory [**Vap98, Vid03**] primarily concerns itself with the first of these questions, while the second question is within the purview of *computational learning theory* [**Val84, KV94**]. However, there are some overlaps between these two fields. In particular, we can immediately classify learning problems into easy and hard ones by looking at how their sample complexity grows as a function of $1/\varepsilon$ and $1/\delta$. In general, an easy problem is one whose sample complexity is *polynomial* in $1/\varepsilon$ and *polylogarithmic* in $1/\delta$ ("polylogarithmic" means polynomial in $\log(1/\delta)$). Of course, there are other factors that affect the sample complexity, and we will pay close attention to those as well.

1.2. From estimation to prediction

The coin tossing example of Section 1.1 was concerned with *estimation*. In fact, estimation was the focus of classical statistics, with early works dating back to Gauss, Laplace and the numerous members of the Bernoulli clan (the book by Stigler [**Sti86**] is an excellent survey of the history of statistics, full of amusing anecdotes and trivia, and much historical and scientific detail besides). By contrast, much of statistical learning theory (and much of modern statistics too) focuses on *prediction* (see the book by Clarke, Fokoué and Zhang [**CFZ09**] for a comprehensive exposition of the predictive view of statistical machine learning and data mining). In a typical prediction problem, we have two jointly distributed random

[1]Unless stated otherwise, log will always denote natural logarithms (base e).

variables[2] X and Y, where only X is available for observation, and we wish to devise a means of predicting Y on the basis of this observation. Thus, a *predictor* is any well-behaved[3] function from X (the domain of X) into Y (the domain of Y). For example, in medical diagnosis, X might record the outcomes of a series of medical tests and other data for a single patient, while $Y \in \{0, 1\}$ would correspond to the patient either having or not having a particular health issue.

The basic premise of statistical learning theory is that the details of the joint distribution P of X and Y are vague (or even completely unknown), and the only information we have to go on is a sequence of n independent observations $(X_1, Y_1), \dots, (X_n, Y_n)$ drawn from P. Assuming we have a quantitative criterion by which to judge a predictor's accuracy, the same basic question presents itself: How large does the sample $\{(X_i, Y_i)\}_{i=1}^n$ have to be in order for us to be able to construct a predictor achieving a given level of accuracy and confidence?

Of course, not all learning problems involve prediction. For example, problems like clustering, density estimation, feature (or representation) learning do not. We will see later that the mathematical formalism of statistical learning theory is flexible enough to cover such problems as well. For now, though, let us focus on prediction to keep things concrete. To get a handle on the learning problem, let us first examine the ideal situation, in which the distribution P is known.

1.2.1. Binary classification. The simplest prediction problem is that of *binary classification* (also known as *pattern classification* or *pattern recognition*) [**DGL96**]. In a typical scenario, X is a subset of $\mathbb{R}^p$, the p-dimensional Euclidean space, and $\mathsf{Y} = \{0, 1\}$. A predictor (or a *classifier*) is any mapping $f : \mathsf{X} \to \{0, 1\}$. A standard way of evaluating the quality of binary classifiers is by looking at their probability of classification error. Thus, for a classifier f we define the *classification loss* (or *risk*)

$$L_P(f) := \mathbf{P}(f(X) \neq Y) \equiv \int_{\mathsf{X}\times\{0,1\}} \mathbf{1}_{\{f(x)\neq y\}} P(\mathrm{d}x, dy),$$

where $\mathbf{1}_{\{\cdot\}}$ is the *indicator function* taking the value 1 if the statement in the braces is true, and 0 otherwise. What is the best classifier for a given P? The answer is given by the following

PROPOSITION 1.1. *Given the joint distribution P on $\mathsf{X} \times \{0, 1\}$, let $\eta(x) := \mathbf{E}[Y|X = x] \equiv \mathbf{P}(Y = 1|X = x)$. Then the classifier*

$$f_P^*(x) := \begin{cases} 1, & \text{if } \eta(x) \geq 1/2 \\ 0, & \text{otherwise} \end{cases} \tag{1.2}$$

minimizes the probability of classification error over all $f : \mathsf{X} \to \{0, 1\}$, i.e.,

$$L_P(f_P^*) = \min_{f:\mathsf{X}\to\{0,1\}} L_P(f).$$

[2]Please consult Appendix A for basic definitions and notation pertaining to probability distributions and random variables.

[3]In fancy language, "well-behaved" will typically mean "measurable" with respect to appropriate σ-fields defined on X and Y. We will ignore measurability issues in this course.

Remark 1.1. Some terminology: The function η defined above is called the *regression function*, the classifier in (1.2) is called the *Bayes classifier*, and its risk

$$L_P^* := L_P(f_P^*)$$

is called the *Bayes rate*.

Proof. Consider an arbitrary classifier $f : \mathsf{X} \to \{0, 1\}$. Then

$$\begin{aligned} L_P(f) &= \int_{\mathsf{X}\times\{0,1\}} \mathbf{1}_{\{f(x)\neq y\}} P(\mathrm{d}x, \mathrm{d}y) \\ &= \int_{\mathsf{X}} P_X(\mathrm{d}x) \left\{ P_{Y|X}(1|x)\mathbf{1}_{\{f(x)\neq 1\}} + P_{Y|X}(0|x)\mathbf{1}_{\{f(x)\neq 0\}} \right\} \\ &= \int_{\mathsf{X}} P_X(\mathrm{d}x) \underbrace{\left\{ \eta(x)\mathbf{1}_{\{f(x)\neq 1\}} + (1-\eta(x))\mathbf{1}_{\{f(x)\neq 0\}} \right\}}_{:=\ell(f,x)}, \qquad (1.3) \end{aligned}$$

where we have used Eq. (1.2.1), the factorization $P = P_X \times P_{Y|X}$, and the definition of η. From the above, it is easy to see that, in order to minimize $L_P(f)$, it suffices to minimize the term $\ell(f, x)$ in (1.3) separately for each value of $x \in \mathsf{X}$. If we let $f(x) = 1$, then $\ell(f, x) = 1 - \eta(x)$, while for $f(x) = 0$ we will have $\ell(f, x) = \eta(x)$. Clearly, we should set $f(x)$ to 1 or 0, depending on whether $1 - \eta(x) \le \eta(x)$ or not. This yields the rule in (1.2). □

1.2.2. Minimum mean squared error prediction. Another prototypical example of a prediction problem is *minimum mean squared error (MMSE) prediction* [**CZ07**], where $\mathsf{X} \subseteq \mathbb{R}^p$, $\mathsf{Y} \subseteq \mathbb{R}$, and the admissible predictors are functions $f : \mathsf{X} \to \mathbb{R}$. The quality of such a predictor f is measured by the *MSE*

$$L_P(f) := \mathbf{E}(f(X) - Y)^2 \equiv \int_{\mathsf{X}\times\mathsf{Y}} (f(x) - y)^2 P(\mathrm{d}x, \mathrm{d}y).$$

The MMSE predictor is characterized by the following

Proposition 1.2. *Given the joint distribution P on $\mathsf{X} \times \mathsf{Y}$ with $\mathsf{X} \subseteq \mathbb{R}^p$ and $\mathsf{Y} \subseteq \mathbb{R}$, the regression function $f_P^*(x) := \mathbf{E}[Y|X = x]$ is the MMSE predictor. Moreover, for any other predictor f we have*

$$L_P(f) = \|f - f_P^*\|^2_{L^2(P_X)} + L_P^*,$$

where for any function $g : \mathsf{X} \to \mathbb{R}$

$$\|g\|^2_{L^2(P_X)} := \int_{\mathsf{X}} |g(x)|^2 P_X(\mathrm{d}x) \equiv \mathbf{E}|g(X)|^2$$

is the squared L^2 norm with respect to the marginal distribution P_X, and $L_P^* := L_P(f_P^*)$.

Proof. Consider an arbitrary predictor $f : \mathsf{X} \to \mathbb{R}$. Then

$$\begin{aligned} L_P(f) &= \mathbf{E}(f(X) - Y)^2 \\ &= \mathbf{E}(f(X) - f_P^*(X) + f_P^*(X) - Y)^2 \\ &= \mathbf{E}(f(X) - f_P^*(X))^2 + 2\mathbf{E}[(f(X) - f_P^*(X))(f_P^*(X) - Y)] + \mathbf{E}(f_P^*(X) - Y)^2 \\ &= \|f - f_P^*\|^2_{L^2(P_X)} + 2\mathbf{E}[(f(X) - f_P^*(X))(f_P^*(X) - Y)] + L_P^*. \end{aligned}$$

Let us analyze the second (cross) term. Using the law of iterated expectation, we have

$$\begin{aligned}\mathbf{E}[(f(X)-f_P^*(X))(f_P^*(X)-Y)] &= \mathbf{E}\big[\mathbf{E}[(f(X)-f_P^*(X))(f_P^*(X)-Y)|X]\big]\\ &= \mathbf{E}\big[(f(X)-f_P^*(X))\mathbf{E}[(f_P^*(X)-Y)|X]\big]\\ &= \mathbf{E}\big[(f(X)-f_P^*(X))(f_P^*(X)-\mathbf{E}[Y|X])\big]\\ &= 0,\end{aligned}$$

where in the last step we used the definition $f_P^*(x) := \mathbf{E}[Y|X=x]$. Thus,

$$L_P(f) = \|f - f_P^*\|^2_{L^2(P_X)} + L_P^* \geq L_P^*,$$

where equality holds if and only if $f = f_X^*$ (with P_X-probability one). □

1.2.3. A general prediction problem. In the general case, X and Y are arbitrary sets, admissible predictors are functions $f : \mathsf{X} \to \mathsf{Y}$ (or, more generally, $f : \mathsf{X} \to \mathsf{U}$ for some suitable *prediction space* U), and the quality of a predictor f on a pair $(x, y) \in \mathsf{X} \times \mathsf{Y}$ is judged in terms of some fixed *loss function* $\ell : \mathsf{U} \times \mathsf{Y} \to \mathbb{R}$ by $\ell(f(x), y)$, the loss incurred in predicting the true y by $\widehat{u} = f(x)$. The expect loss, or risk, of f is then

$$L_P(f) := \mathbf{E}[\ell(f(X), Y)] \equiv \int_{\mathsf{X}\times\mathsf{Y}} \ell(f(x), y) P(\mathrm{d}x, \mathrm{d}y).$$

This set-up covers the two previous examples:

(1) If $\mathsf{X} \subseteq \mathbb{R}^p$, $\mathsf{Y} = \mathsf{U} = \{0, 1\}$, and $\ell(u, y) := \mathbf{1}_{\{u \neq y\}}$, then we recover the binary classification problem.
(2) If $\mathsf{X} \subseteq \mathbb{R}^p$, $\mathsf{Y} \subseteq \mathbb{R} = \mathsf{U}$, and $\ell(u, y) := (u-y)^2$, then we recover the MMSE prediction problem.

Given P and ℓ, we define the minimum risk

$$L_P^* := \inf_{f:\mathsf{X}\to\mathsf{U}} \mathbf{E}[\ell(f(X), Y)], \tag{1.4}$$

where we use inf instead of min since there may not be a minimizing f (when that happens, one typically picks some small $\varepsilon > 0$ and seeks *ε-minimizers*, i.e., any $f_\varepsilon^* : \mathsf{X} \to \mathsf{U}$, such that

$$L_P(f_\varepsilon^*) \leq L_P(f) + \varepsilon \tag{1.5}$$

for all $f : \mathsf{X} \to \mathsf{U}$). We will just assume that a minimizer exists, but continue to use inf to keep things general.

Thus, an abstract prediction problem is characterized by three objects: a probability distribution P of $(X, Y) \in \mathsf{X} \times \mathsf{Y}$, a class of admissible predictors $f : \mathsf{X} \to \mathsf{U}$, and a loss function $\ell : \mathsf{U} \times \mathsf{Y} \to \mathbb{R}$. The solution to the prediction problem is any f_P^* that attains the infimum in (1.4) (or comes ε-close as in (1.5)). Once such a f_P^* is computed, we can use it to predict the *output* $Y \in \mathsf{Y}$ for any given *input* $X \in \mathsf{X}$ by $\widehat{Y} = f_P^*(X)$, where the interpretation is that the random couple $(X, Y) \sim P$ pertains to the phenomenon of interest, X corresponds to its observable aspects, and Y corresponds to some unobservable characteristic that we may want to ascertain.

1.3. Goals of learning

We will close our introduction to statistical learning theory by a rough sketch of the "goals of learning" in a random environment. Please keep in mind that this is not meant to be a definitive treatment, which will come later in the course.

So far we have discussed the "ideal" case when the distribution P of (X, Y) is known. Statistical learning theory deals with the setting where our knowledge of P is only partial (or nonexistent), but we have access to a *training sample* $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent draws from P. Formally, we say that the pairs $(X_i, Y_i), 1 \le i \le n$, are *independent and identically distributed (i.i.d.)* according to P, and we often write this as

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P, \qquad i = 1, \ldots, n.$$

To keep the notation simple, let us denote by Z the product space $\mathsf{X} \times \mathsf{Y}$ and let $Z_i = (X_i, Y_i)$ for each i. Our training sample is then $Z^n = (Z_1, \ldots, Z_n) \in \mathsf{Z}^n$. Roughly speaking, the goal of learning is to take Z^n as an input and to produce a *candidate predictor* $\widehat{f}_n : \mathsf{X} \to \mathsf{U}$ as an output. Note that since Z^n is a random variable, so is $\widehat{f}_n$. A *learning algorithm* (or a *learner*) is a procedure that can do this for any sample size n. Thus, a learning algorithm is a box for converting training samples into predictors.

Let's suppose that we have some learning algorithm to play with. Given a sample Z^n of size n, it outputs a candidate predictor $\widehat{f}_n$. How good is this predictor? Well, let's suppose that someone (say, Nature) hands us a fresh independent sample $Z = (X, Y)$ from the same distribution P that has generated the training sample Z^n. Then we can test $\widehat{f}_n$ by applying it to X and seeing how close $\widehat{U} = \widehat{f}_n(X)$ is to Y by computing the *instantaneous loss* $\ell(\widehat{U}, Y) \equiv \ell(\widehat{f}_n(X).Y)$. The *expectation* of the instantaneous loss w.r.t. the (unknown) distribution P,

$$L_P(\widehat{f}_n) \equiv \int_{\mathsf{X} \times \mathsf{Y}} \ell(\widehat{f}_n(x), y) P(\mathrm{d}x, \mathrm{d}y), \tag{1.6}$$

is called the *generalization error* of the learner at sample size n. It is crucial to note that $L_P(\widehat{f}_n)$ is a *random variable*, since $\widehat{f}_n$ is a function of the random sample Z^n. In fact, to be more precise, we should write the generalization error as the conditional expectation $\mathbf{E}[\ell(\widehat{f}_n(X), Y)|Z^n]$, but since $Z = (X, Y)$ is assumed to be independent from Z^n, we get (1.6).

Now, we will say that our learner has done a good job when its generalization error is suitably small. But how small can it be? To answer this question (or at least to point towards a possible answer), we must first agree that learning without any initial assumptions is a futile task. For example, consider fitting a curve to a training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, where both the X_i's and the Y_i's are real numbers. A simple-minded approach would be to pick any curve that precisely agreed with the entire sample – in other words, to select some $\widehat{f}_n$, such that $\widehat{f}_n(X_i) = Y_i$ for all $i = 1, \ldots, n$. We say that any such $\widehat{f}_n$ *interpolates the data*. But there is an uncountable infinity of such functions! Which one should we choose? The answer is, of course, there is no way to know, if only because we have no clue about P! We could pick a very smooth function, but it could very well happen that the optimal f^*_P tends to be smooth for some values of the input and rough for some others. Alternatively, we could choose a very wiggly and complicated curve, but then it might just be the case that f_P^* is really simple.

A way out of this dilemma is to introduce what is known in the artificial intelligence community as an *inductive bias*. We go about it by restricting the space of candidate predictors our learner is allowed to search over to some suitable family $\mathcal{H}$, which is typically called the *hypothesis space*. Thus, we stipulate that $\widehat{f}_n \in \mathcal{H}$ for any sample Z^n. Given P, let us define the minimum risk over $\mathcal{H}$:

$$L_P^*(\mathcal{H}) := \inf_{f \in \mathcal{H}} L_P(f). \tag{1.7}$$

Clearly, $L_P^*(\mathcal{H}) \geq L_P^*$, since the latter involves minimization over a larger set. However, now, provided the hypothesis space $\mathcal{H}$ is "manageable," we may actually hope to construct a learner that would guarantee that

$$L_P(\widehat{f}_n) \approx L_P^*(\mathcal{H}) \qquad \text{with high probability.} \tag{1.8}$$

Then, if we happen to be so lucky that f_P^* is actually in $\mathcal{H}$, we will have attained the Holy Grail, but even if we are not so lucky, we may still be doing pretty well. To get a rough idea of what is involved, let us look at the *excess risk* of $\widehat{f}_n$ relative to the best predictor f_P^*:

$$\mathbf{E}_P(\widehat{f}_n) := L_P(\widehat{f}_n) - L_P^* = \underbrace{L_P(\widehat{f}_n) - L_P^*(\mathcal{H})}_{\mathbf{E}_{\text{est}}} + \underbrace{L_P^*(\mathcal{H}) - L_P^*}_{\mathbf{E}_{\text{approx}}}. \tag{1.9}$$

If the learner is good in the sense of (1.8), then we will have

$$\mathbf{E}_P(\widehat{f}_n) \approx L_P^*(\mathcal{H}) - L_P^* \qquad \text{with high probability,}$$

which, in some sense, is the next best thing to the Holy Grail, especially if we can choose $\mathcal{H}$ so well that we can guarantee that the difference $L_P^*(\mathcal{H}) - L_P^*$ is small for any possible choice of P.

Recently, however, the following curious fact was pointed out [**BHM18, BRT19**]: In some situations, it is in fact possible to construct an explicit predictor $\widehat{f}_n$ that interpolates the data, i.e., $\widehat{f}_n(X_i) = Y_i$ for all $i = 1, \dots, n$, and yet performs well out of sample. This, however, requires dropping the requirement that $\widehat{f}_n$ be an element of $\mathcal{H}$ (this is referred to as *improper learning*). In the context of curve fitting, for instance, one can construct an interpolating $\widehat{f}_n$ with the property that

$$\lim_{n \to \infty} \mathbf{E}[(\widehat{f}_n(X) - \eta(X))^2] = 0$$

whenever the regression function η belongs to a certain hypothesis class $\mathcal{H}$, and thus

$$\mathbf{E}[L_P(\widehat{f}_n)] - \inf_{f \in \mathcal{H}} \mathbf{E}[L_P(f)] = \mathbf{E}[(\widehat{f}_n(X) - \eta(X))^2] \xrightarrow{n \to \infty} 0$$

where the expectation is taken with respect to both the training data $(X_1, Y_1), \dots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P$ and the fresh independent sample $(X, Y) \sim P$. This phenomenon, where there is no adverse effect of interpolation on out-of-sample performance, is relevant in the context of learning using deep neural nets.

Note the decomposition of the excess risk into two terms, denoted in (1.9) by $\mathbf{E}_{\text{est}}$ and $\mathbf{E}_{\text{approx}}$. The first term, $\mathbf{E}_{\text{est}}$, depends on the learned predictor $\widehat{f}_n$, as well as on the hypothesis class $\mathcal{H}$, and is referred to as the *estimation error* of the learner. The second term, $\mathbf{E}_{\text{approx}}$, depends only on $\mathcal{H}$ and on P, and is referred to as the *approximation error* of the hypothesis space. Most of the effort in statistical learning theory goes into analyzing and bounding the estimation error for various choices of $\mathcal{H}$. Analysis of $\mathbf{E}_{\text{approx}}$ is the natural domain of approximation theory. The overall performance of a given learning algorithm depends on the interplay between these two sources of error. The text by Cucker and Zhou [**CZ07**] does a wonderful job of treating both the estimation and the approximation aspects of learning algorithms.

1.3.1. Beyond prediction. As we had briefly pointed out earlier, not all learning problems involve prediction. Luckily, the mathematical formalism we have just introduced can be easily adapted to a more general view of learning (details are in Chapter 6). Consider a random object Z taking values in some space Z according to an unknown distribution P. Suppose that there is a very large class $\mathcal{F}$ of functions $f : \mathsf{Z} \to \mathbb{R}$, and for each $f \in \mathcal{F}$ we can define its expected loss (or risk)

$$L_P(f) := \mathbf{E}[f(Z)] = \int_{\mathsf{Z}} f(z)P(\mathrm{d}z). \tag{1.10}$$

Suppose also that $\mathcal{F}$ has the property that there exists at least one $f^*_P \in \mathcal{F}$ that achieves

$$L_P(f^*_P) = \inf_{f\in\mathcal{F}} L_P(f). \tag{1.11}$$

The class $\mathcal{F}$ may even depend on P. Let's see how we can describe some unsupervised learning problems in this way:

- **Density estimation.** Suppose that $\mathsf{Z} \subseteq \mathbb{R}^d$ for some d, and that P has a probability density function (pdf) p. We can construct a suitable class $\mathcal{F}$ as follows: pick a nonnegative function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, and let $\mathcal{F} = \mathcal{F}_{P,\ell}$ consist of all functions of the form

$$f_q(z) = \ell\big(p(z), q(z)\big), \tag{1.12}$$

as q ranges over a suitable class of pdf's q on $\mathbb{R}^d$. Then

$$L_P(f_q) = \mathbf{E}_P\left[\ell\big(p(Z), q(Z)\big)\right] = \int_{\mathbb{R}^d} p(z)\ell\big(p(z), q(z)\big)\mathrm{d}z. \tag{1.13}$$

This is fairly general. For example, if we assume that $p > 0$ on Z, then we can let $\ell(u,u') = \left|\frac{u'}{u} - 1\right|^2$, in which case we recover the L^2 criterion:

$$L_P(f_q) = \int_{\mathsf{Z}} p(z)\left|\frac{q(z)}{p(z)} - 1\right|^2 \mathrm{d}z = \|p - q\|^2_{L^2},$$

known as Chi-squared divergence, $\chi^2(q\|p)$. Or we can let $\ell(u,u') = \log(u/u')$, which gives us the *relative entropy* (also known as the *Kullback–Leibler divergence*):

$$L_P(f_q) = \int_{\mathsf{Z}} p(z)\log\frac{p(z)}{q(z)}\mathrm{d}z = D(p\|q).$$

- **Clustering.** In a basic form of the clustering problem, we seek a partition of the domain Z of interest into a fixed number, say k, of disjoint clusters $\mathsf{C}_1, \ldots, \mathsf{C}_k$, such that all points z that belong to the same cluster are somehow "similar." For example, we may define a distance function $d : \mathsf{Z} \times \mathsf{Z} \to \mathbb{R}^+$ and represent each cluster C_j, $1 \le j \le k$, by a single "representative" $v_j \in \mathsf{Z}$. A clustering $\mathbf{C}$ is then described by k pairs $\{(\mathsf{C}_j, v_j)\}_{j=1}^k$, where $\mathsf{Z} = \bigcup_{j=1}^k \mathsf{C}_j$. Consider the class $\mathcal{F} = \mathcal{F}_k$ of all functions of the form

$$f_{\mathbf{C}}(z) = \sum_{j=1}^k \mathbf{1}_{\{z \in \mathsf{C}_j\}} d(z, v_j)$$

as $\mathbf{C}$ runs over all clusterings $\{(\mathsf{C}_j, v_j)\}_{j=1}^k$. We can then evaluate the quality of our clustering $\mathbf{C}$ by looking at the expectation

$$L_P(f_{\mathbf{C}}) = \mathbf{E}_P\left[\sum_{j=1}^k \mathbf{1}_{\{Z \in \mathsf{C}_j\}} d(Z, v_j)\right]$$

- **Feature learning.** Broadly speaking, feature learning refers to constructing a representation of the original input Z that could be fed to a supervised learning algorithm further down the line. There could be multiple reasons for wanting to do this, ranging from computational considerations to a desire to capture "salient" characteristics of the data that could be relevant for prediction, while "factoring out" the irrelevant parts. Mathematically, a feature is a mapping $\varphi : \mathsf{Z} \to \widetilde{\mathsf{Z}}$ into some other representation space $\widetilde{\mathsf{Z}}$, so that each point $z \in \mathsf{Z}$ is represented $\widetilde{z} = \varphi(z)$, and it is this representation that will be used by another learning algorithm down the line. (Ideally, good feature representations should be agnostic with respect to the nature of the learning problem where they will be used.) One way to score the quality of a feature is to consider a loss function of the form $\ell : \mathsf{Z} \times \widetilde{\mathsf{Z}} \to \mathbb{R}^+$, so that $\ell(z, \widetilde{z})$ is small if z is well-represented by $\widetilde{z}$. Then, for a fixed collection Φ of candidate feature maps, we could consider a class $\mathcal{F} = \mathcal{F}_{\Phi,\ell}$ of functions of the form

$$f_\varphi(z) = \ell\big(z, \varphi(z)\big), \qquad \varphi \in \Phi.$$

  This is a very wide umbrella that can cover a wide variety of unsupervised learning tasks (e.g., clustering).

These examples show that unsupervised learning problems can also be formulated in terms of minimizing an appropriately defined expected loss. The only difference is that the loss function may sometimes depend on the underlying distribution, which is unknown. However, under suitable assumptions on the problem components, it is often possible to find an alternative hypothesis space $\mathcal{H}$ which (unlike $\mathcal{F}$) does not depend on P, such that the minimum expected loss $L_P^*(\mathcal{F})$ can be related to the minimum expected loss $L_P^*(\mathcal{H})$. Just as before, a learning algorithm is a rule for mapping an i.i.d. sample $Z^n = (Z_1, \ldots, Z_n)$ from P to an element $\widehat{f}_n \in \mathcal{H}$. The objective is also the same as before: ensure that

$$L_P(\widehat{f}_n) \approx L_P^*(\mathcal{H}) \qquad \text{with high probability.}$$

Thus, we can treat supervised learning and unsupervised learning on the same footing.

CHAPTER 2

Concentration inequalities

In the previous chapter, the following result was stated without proof. If $X_1, \ldots, X_n$ are independent Bernoulli(θ) random variables representing the outcomes of a sequence of n tosses of a coin with bias (probability of HEADS) θ, then for any $\varepsilon \in (0,1)$

$$\mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| \ge \varepsilon\right) \le 2e^{-2n\varepsilon^2} \tag{2.1}$$

where

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

is the fraction of HEADS in $X^n = (X_1, \ldots, X_n)$. Since $\theta = \mathbf{E}\widehat{\theta}_n$, (2.1) says that the *sample* (or *empirical*) *average* of the X_i's *concentrates sharply* around the statistical average $\theta = \mathbf{E}X_1$. Bounds like these are fundamental in statistical learning theory. In this chapter, we will learn the techniques needed to derive such bounds for settings much more complicated than coin tossing. This is not meant to be a complete picture; a detailed treatment can be found in the excellent recent book by Boucheron, Lugosi, and Massart [BLM13].

2.1. The basic tools

We start with *Markov's inequality*: Let $Y \in \mathbb{R}$ be a nonnegative random variable. Then for any $t > 0$ we have

$$\mathbf{P}(Y \ge t) \le \frac{\mathbf{E}Y}{t}. \tag{2.2}$$

The proof is simple:

$$\mathbf{P}(Y \ge t) = \mathbf{E}[\mathbf{1}_{\{Y \ge t\}}] \tag{2.3}$$

$$\le \frac{\mathbf{E}[Y\mathbf{1}_{\{Y \ge t\}}]}{t} \tag{2.4}$$

$$\le \frac{\mathbf{E}Y}{t}, \tag{2.5}$$

where:

- (2.3) uses the fact that the probability of an event can be expressed as the expectation of its indicator function:
$$\mathbf{P}(Y \in A) = \int_A P_Y(dy) = \int_{\mathsf{Y}} \mathbf{1}_{\{x \in A\}} P_Y(dy) = \mathbf{E}[\mathbf{1}_{\{Y \in A\}}]$$
- (2.4) uses the fact that
$$Y \ge t > 0 \qquad \Longrightarrow \qquad \frac{Y}{t} \ge 1$$

- (2.5) uses the fact that
$$Y \geq 0 \qquad \Longrightarrow \qquad Y\mathbf{1}_{\{Y \geq t\}} \leq Y,$$
so consequently $\mathbf{E}[Y\mathbf{1}_{\{Y\geq t\}}] \leq \mathbf{E}Y$.

Markov's inequality leads to our first bound on the probability that a random variable deviates from its expectation by more than a given amount: *Chebyshev's inequality*. Let X be an arbitrary real random variable. Then for any $t > 0$

$$\mathbf{P}\left(|X - \mathbf{E}X| \geq t\right) \leq \frac{\mathrm{Var}[X]}{t^2}, \tag{2.6}$$

where $\mathrm{Var}\, X := \mathbf{E}[|X - \mathbf{E}X|^2] = \mathbf{E}X^2 - (\mathbf{E}X)^2$ is the variance of X. To prove (2.6), we apply Markov's inequality (2.2) to the nonnegative random variable $Y = |X - \mathbf{E}X|^2$:

$$\mathbf{P}\left(|X - \mathbf{E}X| \geq t\right) = \mathbf{P}\left(|X - \mathbf{E}X|^2 \geq t^2\right) \tag{2.7}$$

$$\leq \frac{\mathbf{E}|X - \mathbf{E}X|^2}{t^2}, \tag{2.8}$$

where the first step uses the fact that the function $\phi(x) = x^2$ is monotonically increasing on $[0, \infty)$, so that $a \geq b \geq 0$ if and only if $a^2 \geq b^2$.

Now let's apply these tools to the problem of bounding the probability that, for a coin with bias θ, the fraction of HEADS in n trials differs from θ by more than some $\varepsilon > 0$. To that end, let us represent the outcomes of the n tosses by n independent Bernoulli(θ) random variables $X_1, \ldots, X_n \in \{0, 1\}$, where $\mathbf{P}(X_i = 1) = \theta$ for all i. Let

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i.$$

Then

$$\mathbf{E}\widehat{\theta}_n = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \underbrace{\mathbf{E}X_i}_{=\mathbf{P}(X_i=1)} = \theta$$

and

$$\mathrm{Var}[\widehat{\theta}_n] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{\theta(1-\theta)}{n},$$

where we have used the fact that the X_i's are i.i.d., so $\mathrm{Var}[X_1 + \ldots + X_n] = \sum_{i=1}^n \mathrm{Var}\, X_i = n\,\mathrm{Var}\, X_1$. Now we are in a position to apply Chebyshev's inequality:

$$\mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}[\widehat{\theta}_n]}{\varepsilon^2} = \frac{\theta(1-\theta)}{n\varepsilon^2}. \tag{2.9}$$

At the very least, (2.9) shows that the probability of getting a bad sample decreases with sample size. Unfortunately, it does not decrease fast enough. To see why, we can appeal to the Central Limit Theorem, which (roughly) states that

$$\mathbf{P}\left(\sqrt{\frac{n}{\theta(1-\theta)}}\left(\widehat{\theta}_n - \theta\right) \geq t\right) \xrightarrow{n\to\infty} 1 - \Phi(t) \leq \frac{1}{\sqrt{2\pi}}\frac{e^{-t^2/2}}{t},$$

where $\Phi(t) = (1/\sqrt{2\pi}) \int_{-\infty}^{t} e^{-x^2/2} dx$ is the standard Gaussian CDF. This would suggest something like

$$\mathbf{P}\left(\widehat{\theta}_n - \theta \geq \varepsilon\right) \approx \exp\left(-\frac{n\varepsilon^2}{2\theta(1-\theta)}\right),$$

which decays with n much faster than the right-hand side of (2.9),

2.2. The Chernoff bounding trick and Hoeffding's inequality

To fix (2.9), we will use a very powerful technique, known as the *Chernoff bounding trick* [**Che52**]. Let X be real-valued random variable. Suppose we are interested in bounding the probability $\mathbf{P}(X \geq \mathbf{E}X + t)$ for some particular $t > 0$. Observe that for any $s > 0$ we have

$$\mathbf{P}(X \geq \mathbf{E}X + t) = \mathbf{P}\left(e^{s(X-\mathbf{E}X)} \geq e^{st}\right) \leq e^{-st}\mathbf{E}\left[e^{s(X-\mathbf{E}X)}\right], \tag{2.10}$$

where the first step is by monotonicity of the function $\phi(x) = e^{sx}$ and the second step is by Markov's inequality (2.2). The Chernoff trick is to choose an $s > 0$ that would make the right-hand side of (2.10) suitably small. In fact, since (2.10) holds simultaneously for *all* $s > 0$, the optimal thing to do is to take the infimum of the bound over $s > 0$:

$$\mathbf{P}(X \geq \mathbf{E}X + t) \leq \inf_{s>0} e^{-st}\mathbf{E}\left[e^{s(X-\mathbf{E}X)}\right].$$

However, often a good upper bound on the *moment-generating function* $\mathbf{E}\left[e^{s(X-\mathbf{E}X)}\right]$ is enough. One such bound was developed by Hoeffding [**Hoe63**] for the case when X is bounded with probability one:

LEMMA 2.1 (Hoeffding). *Let X be a random variable, such tha $\mathbf{P}(a \leq X \leq b) = 1$ for some $-\infty < a \leq b < \infty$. Then for all $s > 0$*

$$\mathbf{E}\left[e^{s(X-\mathbf{E}X)}\right] \leq e^{s^2(b-a)^2/8}. \tag{2.11}$$

To prove the lemma, we first start with a useful bound on the *variance* of a bounded random variable:

LEMMA 2.2. *If U is a random variable such that $\mathbf{P}(a \leq U \leq b)$, then*

$$\mathrm{Var}[U] \leq \frac{(b-a)^2}{4}. \tag{2.12}$$

PROOF. We use the fact that, for any real-valued random variable U,

$$\mathrm{Var}[U] \leq \mathbf{E}[(U-c)^2], \qquad \forall c \in \mathbb{R}. \tag{2.13}$$

(In particular $c = \mathbf{E}U$ achieves equality in the above bound.) Now let $c = \frac{a+b}{2}$, the midpoint of the interval $[a, b]$. Then, since $a \leq U \leq b$ almost surely, we know that

$$|U - c| \leq \frac{b-a}{2}.$$

Using this c in (2.13), we obtain $\mathrm{Var}[U] \leq \mathbf{E}[(U-c)^2] \leq \frac{(b-a)^2}{4}$, as claimed. ∎

Remark 2.1. The bound of Lemma 2.2 is actually sharp: consider

$$U = \begin{cases} a, & \text{with prob. } 1/2 \\ b, & \text{with prob. } 1/2 \end{cases}.$$

Then

$$\mathrm{Var}[U] = \mathbf{E}U^2 - (\mathbf{E}U)^2 = \frac{a^2+b^2}{2} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{4}.$$

Now we can prove Hoeffding's lemma:

Proof (of Lemma 2.1). Without loss of generality, we may assume that $\mathbf{E}X = 0$. Thus, we are interested in bounding $\mathbf{E}[e^{sX}]$. Let's consider instead the *logarithmic moment-generating function*

$$\psi(s) := \log \mathbf{E}[e^{sX}].$$

Then

$$\psi'(s) = \frac{\mathbf{E}[Xe^{sX}]}{\mathbf{E}[e^{sX}]}, \qquad \psi''(s) = \frac{\mathbf{E}[X^2e^{sX}]}{\mathbf{E}[e^{sX}]} - \left[\frac{\mathbf{E}[Xe^{sX}]}{\mathbf{E}[e^{sX}]}\right]^2. \tag{2.14}$$

(we are being a bit loose here, assuming that we can interchange the order of differentiation and expectation, but in this case everything can be confirmed rigorously). Now consider another random variable U whose distribution is related to X by

$$\mathbf{E}[f(U)] = \frac{\mathbf{E}[f(X)e^{sX}]}{\mathbf{E}[e^{sX}]} \tag{2.15}$$

for any real-valued function $f : \mathbb{R} \to \mathbb{R}$. To convince ourselves that this is a legitimate construction, let's plug in an indicator function of any event A:

$$\mathbf{P}[U \in A] = \mathbf{E}[\mathbf{1}_{\{U\in A\}}] = \frac{\mathbf{E}[\mathbf{1}_{\{X\in A\}}e^{sX}]}{\mathbf{E}[e^{sX}]}. \tag{2.16}$$

It is then not hard to show that this is indeed a valid probability measure. This construction is known as the *twisting* (or *tilting*) technique or as *exponential change of measure.*

We note two things:

(1) Using (2.16) with $A = [a,b]$, we get

$$\mathbf{P}[a \le U \le b] = \frac{\mathbf{E}[\mathbf{1}_{\{a\le X\le b\}}e^{sX}]}{\mathbf{E}[e^{sX}]} = 1, \tag{2.17}$$

since $a \le X \le b$. Moreover, if A is any event in the complement of $[a,b]$, then $\mathbf{P}[U \in A] = 0$, since $\mathbf{E}[\mathbf{1}_{\{X\in A\}}e^{sX}] = 0$. That is, U is bounded between a and b with probability one, just like X.

(2) Using (2.15) first with $f(U) = U$ and then with $f(U) = U^2$, we get

$$\mathbf{E}[U] = \frac{\mathbf{E}[Xe^{sX}]}{\mathbf{E}[e^{sX}]}, \qquad \mathbf{E}[U^2] = \frac{\mathbf{E}[X^2e^{sX}]}{\mathbf{E}[e^{sX}]}. \tag{2.18}$$

Comparing the expressions in (2.18) with (2.14), we observe that $\psi''(s) = \mathrm{Var}[U]$. Now, since $a \le U \le B$, it follows from Lemma 2.2 that $\psi''(s) \le \frac{(b-a)^2}{4}$. Therefore,

$$\psi(s) = \int_0^s \int_0^t \psi''(v) \mathrm{d}v \mathrm{d}t \le \frac{s^2(b-a)^2}{8},$$

where we have used the fact that $\psi'(0) = \psi(0) = 0$. Exponentiating both sides, we are done. $\square$

We will now use the Chernoff method and the above lemma to prove the following

THEOREM 2.1 (Hoeffding's inequality). *Let $X_1, \ldots, X_n$ be independent random variables, such that $X_i \in [a_i, b_i]$ with probability one. Let $S_n := \sum_{i=1}^n X_i$. Then for any $t > 0$*

$$\mathbf{P}\left(S_n - \mathbf{E}S_n \ge t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right); \tag{2.19}$$

$$\mathbf{P}\left(S_n - \mathbf{E}S_n \le -t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{2.20}$$

Consequently,

$$\mathbf{P}\left(|S_n - \mathbf{E}S_n| \ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{2.21}$$

PROOF. By replacing each X_i with $X_i - \mathbf{E}X_i$, we may as well assume that $\mathbf{E}X_i = 0$. Then $S_n = \sum_{i=1}^n X_i$. Using Chernoff's trick, for $s > 0$ we have

$$\mathbf{P}\left(S_n \ge t\right) = \mathbf{P}\left(e^{sS_n} \ge e^{st}\right) \le e^{-st}\mathbf{E}\left[e^{sS_n}\right]. \tag{2.22}$$

Since the X_i's are independent,

$$\mathbf{E}\left[e^{sS_n}\right] = \mathbf{E}\left[e^{s(X_1+\ldots+X_n)}\right] = \mathbf{E}\left[\prod_{i=1}^n e^{sX_i}\right] = \prod_{i=1}^n \mathbf{E}\left[e^{sX_i}\right]. \tag{2.23}$$

Since $X_i \in [a_i, b_i]$, we can apply Lemma 2.1 to write $\mathbf{E}\left[e^{sX_i}\right] \le e^{s^2(b_i-a_i)^2/8}$. Substituting this into (2.23) and (2.22), we obtain

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

If we choose $s = \frac{4t}{\sum_{i=1}^n (b_i-a_i)^2}$, then we obtain (2.19). The proof of (2.20) is similar. $\square$

Now we will apply Hoeffding's inequality to improve our crude concentration bound (2.9) for the sum of n independent Bernoulli(θ) random variables, $X_1, \ldots, X_n$. Since each $X_i \in \{0, 1\}$, we can apply Theorem 2.1 to get, for any $t > 0$,

$$\mathbf{P}\left(\left|\sum_{i=1}^n X_i - n\theta\right| \ge t\right) \le 2e^{-2t^2/n}.$$

Therefore,

$$\mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) = \mathbf{P}\left(\left|\sum_{i=1}^n X_i - n\theta\right| \geq n\varepsilon\right) \leq 2e^{-2n\varepsilon^2},$$

which gives us the claimed bound (2.1).

Theorem 2.1 extends with essentially the same proof to the case that the random variables $X_1, \ldots, X_n$ are not necessarily independent, but form a martingale difference sequence, or, equivalently, the partial sums $Y_k = X_1 + \ldots + X_k$ form a martingale. A random process $(Y_n : n \geq 0)$ is a martingale with respect to a filtration of σ-algebras $\boldsymbol{\mathcal{F}} = (\mathcal{F}_n : n \geq 0)$ if $\mathbf{E}[Y_0]$ is finite, Y_n is $\mathcal{F}_n$ measurable for each $n \geq 0$, and $E[Y_{n+1}|\mathcal{F}_n] = Y_n$. A random process $(B_n : n \geq 1)$, is a predictable process for the filtration $\boldsymbol{\mathcal{F}}$ if B_n is $\mathcal{F}_{n-1}$ measurable for each $n \geq 1$.

Theorem 2.2 (Azuma-Hoeffding inequality with centering). *) Let $(Y_n : n \geq 0)$ be a martingale and $(B_n : n \geq 1)$ be a predictable process, both with respect to a filtration $\boldsymbol{\mathcal{F}} = (\mathcal{F}_n : n \geq 0)$, such that $P\{|Y_n - B_n| \leq c_n/2\} = 1$ for all $n \geq 0$. Then*

$$P\{Y_n - Y_0 \geq t\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$
$$P\{Y_n - Y_0 \leq -t\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

2.3. From bounded variables to bounded differences: McDiarmid's inequality

Hoeffding's inequality applies to sums of independent random variables. We will now develop its generalization, due to McDiarmid [McD89], to *arbitrary* real-valued functions of independent random variables that satisfy a certain condition.

Let X be some set, and consider a function $g : \mathsf{X}^n \to \mathbb{R}$. We say that g has *bounded differences* if there exist nonnegative numbers $c_1, \ldots, c_n$, such that

$$\sup_{x \in \mathsf{X}} g(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) - \inf_{x \in \mathsf{X}} g(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \leq c_i \tag{2.24}$$

for all $i = 1, \ldots, n$ and all $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \in \mathsf{X}$. In words, if we change the ith variable while keeping all the others fixed, the value of g will not change by more than c_i.

Theorem 2.3 (McDiarmid's inequality [McD89]). *Let $X^n = (X_1, \ldots, X_n) \in \mathsf{X}^n$ be an n-tuple of independent X-valued random variables. If a function $g : \mathsf{X}^n \to \mathbb{R}$ has bounded differences, as in (2.24), then, for all $t > 0$,*

$$\mathbf{P}\left(g(X^n) - \mathbf{E}g(X^n) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right); \tag{2.25}$$

$$\mathbf{P}\left(\mathbf{E}g(X^n) - g(X^n) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right). \tag{2.26}$$

Proof. Let us first sketch the general idea behind the proof. Let $Z = g(X^n)$ and $V = Z - \mathbf{E}Z$. The first step will be to write V as a sum $\sum_{i=1}^n V_i$, where the terms V_i are constructed so that:

(1) V_i is a function only of $X^i = (X_1, \ldots, X_i)$, and $\mathbf{E}[V_i|X^{i-1}] = 0$.

(2) There exist functions $A_i, B_i : \mathsf{X}^{i-1} \to \mathbb{R}$ such that, conditionally on X^{i-1},

$$A_i(X^{i-1}) \le V_i \le B_i(X^{i-1}),$$

and, moreover, $B_i(X^{i-1}) - A_i(X^{i-1}) \le c_i$.

Provided we can arrange things in this way, we can apply Lemma 2.1 to V_i conditionally on X^{i-1}:

$$\mathbf{E}\Big[e^{sV_i}\Big|X^{i-1}\Big] \le e^{s^2c_i^2/8}. \tag{2.27}$$

Then, using Chernoff's method, we have

$$\begin{aligned}
\mathbf{P}\left(Z - \mathbf{E}Z \ge t\right) &= \mathbf{P}(V \ge t) \\
&\le e^{-st}\mathbf{E}\left[e^{sV}\right] \\
&= e^{-st}\mathbf{E}\left[e^{s\sum_{i=1}^n V_i}\right] \\
&= e^{-st}\mathbf{E}\left[e^{s\sum_{i=1}^{n-1} V_i}e^{sV_n}\right] \\
&= e^{-st}\mathbf{E}\left[e^{s\sum_{i=1}^{n-1} V_i}\mathbf{E}\Big[e^{sV_n}\Big|X^{n-1}\Big]\right] \\
&\le e^{-st}e^{s^2c_n^2/8}\mathbf{E}\left[e^{s\sum_{i=1}^{n-1} V_i}\right],
\end{aligned}$$

where in the next-to-last step we used the fact that $V_1, \ldots, V_{n-1}$ depend only on X^{n-1}, and in the last step we used (2.27) with $i = n$. If we continue peeling off the terms involving $V_{n-1}, V_{n-2}, \ldots, V_1$, we will get

$$\mathbf{P}\left(Z - \mathbf{E}Z \ge t\right) \le \exp\left(-st + \frac{s^2}{8}\sum_{i=1}^n c_i^2\right).$$

Taking $s = 4t/\sum_{i=1}^n c_i^2$, we end up with (2.25).

It remains to construct the V_i's with the desired properties. To that end, let

$$V_i = \mathbf{E}[Z|X^i] - \mathbf{E}[Z|X^{i-1}],$$

where $\mathbf{E}[Z|X^0] = \mathbf{E}Z$, and, by telescoping,

$$\sum_{i=1}^n V_i = \sum_{i=1}^n \left\{\mathbf{E}[Z|X^i] - \mathbf{E}[Z|X^{i-1}]\right\} = \mathbf{E}[Z|X^n] - \mathbf{E}Z = Z - \mathbf{E}Z = V.$$

Note that V_i depends only on X^i by construction, and that

$$\begin{aligned}
\mathbf{E}[V_i|X^{i-1}] &= \mathbf{E}\Big[\mathbf{E}[Z|X^i] - \mathbf{E}[Z|X^{i-1}]\Big|X^{i-1}\Big] \\
&= \mathbf{E}\Big[\mathbf{E}[Z|X^{i-1}, X_i]\Big|X^{i-1}\Big] - \mathbf{E}[Z|X^{i-1}] \\
&= \mathbf{E}[Z|X^{i-1}] - \mathbf{E}[Z|X^{i-1}] \\
&= 0,
\end{aligned}$$

where we have used the law of iterated expectation in the conditional form $\mathbf{E}[\mathbf{E}[U|V,W]|V] = \mathbf{E}[U|V]$. Moreover, let

$$\begin{aligned}
A_i(X^{i-1}) &= \inf_{x \in \mathsf{X}} \mathbf{E}[g(X^{i-1}, x, X_{i+1}^n) - g(X^n)|X^{i-1}] \\
B_i(X^{i-1}) &= \sup_{x \in \mathsf{X}} \mathbf{E}[g(X^{i-1}, x, X_{i+1}^n) - g(X^n)|X^{i-1}],
\end{aligned}$$

where we have used the fact that the X_i's are independent, and where $X_{i+1}^n := (X_{i+1}, \ldots, X_n)$. Then evidently $A_i(X^{-1}) \le V_i \le B_i(X^{i-1})$, and

$$\begin{aligned}
B_i(X^{i-1}) - A_i(X^{i-1}) &= \sup_{x \in \mathsf{X}} \sup_{x' \in \mathsf{X}} \mathbf{E}[g(X^{i-1}, x, X_{i+1}^n) - g(X^{i-1}, x', X_{i+1}^n)|X^{i-1}] \\
&= \sup_{x \in \mathsf{X}} \sup_{x' \in \mathsf{X}} \left(\int \left[g(X^{i-1}, x, x_{i+1}^n) - g(X^{i-1}, x', x_{i+1}^n) \right] P(\mathrm{d}x_{i+1}^n) \right) \\
&\le \int \sup_{x \in \mathsf{X}} \sup_{x' \in \mathsf{X}'} \left| g(X^{i-1}, x, x_{i+1}^n) - g(X^{i-1}, x', x_{i+1}^n) \right| P(\mathrm{d}x_{i+1}^n) \\
&\le c_i,
\end{aligned}$$

where the last step follows from the bounded difference property of g. $\square$

2.4. McDiarmid's inequality in action

McDiarmid's inequality is an extremely powerful and often used tool in statistical learning theory. We will now discuss several examples of its use. To that end, we will first introduce some notation and definitions.

Let X be some (measurable) space. If Q is a probability distribution of an X-valued random variable X, then we can compute the expectation of any (measurable) function $f : \mathsf{X} \to \mathbb{R}$ w.r.t. Q. So far, we have denoted this expectation by $\mathbf{E}f(X)$ or by $\mathbf{E}_Q f(X)$. We will often find it convenient to use an alternative notation, $Q(f)$.

Let $X^n = (X_1, \ldots, X_n)$ be n independent identically distributed (i.i.d.) X-valued random variables with common distribution P. The main object of interest to us is the *empirical distribution* induced by X^n, which we will denote by P_n. The empirical distribution assigns the probability $1/n$ to each X_i, i.e.,

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}.$$

Here, δ_x denotes a unit mass concentrated at a point $x \in \mathsf{X}$, i.e., the probability distribution on X that assigns each event A the probability

$$\delta_x(A) = \mathbf{1}_{\{x \in A\}}, \qquad \forall \text{ measurable } A \subseteq \mathsf{X}.$$

We note the following important facts about P_n:

(1) Being a function of the sample X^n, P_n is a *random variable* taking values in the space of probability distributions over X.
(2) The probability of a set $A \subseteq \mathsf{X}$ under P_n,

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

is the *empirical frequency* of the set A on the sample X^n. The expectation of $P_n(A)$ is equal to $P(A)$, the P-probability of A. Indeed,

$$\mathbf{E}P_n(A) = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i \in A\}}\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[\mathbf{1}_{\{X_i \in A\}}] = \frac{1}{n}\sum_{i=1}^n \mathbf{P}(X_i \in A) = P(A).$$

(Think back to our coin-tossing example – this is a generalization of that idea, where we approximate actual probabilities of events by their relative frequencies in a series of independent trials.)

(3) Given a function $f : \mathsf{X} \to \mathbb{R}$, we can compute its expectation w.r.t. P_n:

$$P_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i),$$

which is just the sample mean of f on X^n. It is also referred to as the *empirical expectation* of f on X^n. We have

$$\mathbf{E}P_n(f) = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n f(X_i)\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}f(X_i) = \mathbf{E}f(X) = P(f).$$

We can now proceed to our examples.

2.4.1. Sums of bounded random variables. In the special case when $\mathsf{X} = \mathbb{R}$, P is a probability distribution supported on a finite interval, and $g(X^n)$ is the sum

$$g(X^n) = \sum_{i=1}^n X_i,$$

McDiarmid's inequality simply reduces to Hoeffding's. Indeed, for any $x^n \in [a,b]^n$ and $x'_i \in [a,b]$ we have

$$g(x^{i-1}, x_i, x^n_{i+1}) - g(x^{i-1}, x'_i, x^n_{i+1}) = x_i - x'_i \le b - a.$$

Interchanging the roles of x'_i and x_i, we get

$$g(x^{i-1}, x'_i, x^n_{i+1}) - g(x^{i-1}, x_i, x^n_{i+1}) = x'_i - x_i \le b - a.$$

Hence, we may apply Theorem 2.3 with $c_i = b - a$ for all i to get

$$\mathbf{P}\left(|g(X^n) - \mathbf{E}g(X^n)| \ge t\right) \le 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right).$$

2.4.2. Uniform deviations. Let $X_1, \dots, X_n$ be n i.i.d. X-valued random variables with common distribution P. By the Law of Large Numbers, for any $A \subseteq \mathsf{X}$ and any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}\left(|P_n(A) - P(A)| \ge \varepsilon\right) = 0.$$

In fact, we can use Hoeffding's inequality to show that

$$\mathbf{P}\left(|P_n(A) - P(A)| \ge \varepsilon\right) \le 2e^{-2n\varepsilon^2}.$$

This probability bound holds for each A *separately*. However, in learning theory we are often interested in the deviation of empirical frequencies from true probabilities simultaneously over some *collection* of subsets of X. To that end, let $\mathcal{A}$ be such a collection and consider the function

$$g(X^n) := \sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \,. \tag{2.28}$$

Later in the course we will see that, for certain choices of $\mathcal{A}$, $\mathbf{E}g(X^n) = O(1/\sqrt{n})$. However, regardless of what $\mathcal{A}$ is, it is easy to see that, by changing only one X_i, the value of $g(X^n)$ can change at most by $1/n$. Let $x^n = (x_1, \ldots, x_n)$, choose some other $x'_i \in \mathsf{X}$, and let $x^n_{(i)}$ denote x^n with x_i replaced by x'_i:

$$x^n = (x^{i-1}, x_i, x^n_{i+1}), \qquad x^n_{(i)} = (x^{i-1}, x'_i, x^n_{i+1}).$$

Then

$$\begin{aligned}
g(x^n) - g(x^n_{(i)}) &= \sup_{A \in \mathcal{A}} \Big| P_{x^n}(A) - P(A) \Big| - \sup_{A' \in \mathcal{A}} \Big| P_{x^n_{(i)}}(A') - P(A') \Big| \\
&= \sup_{A \in \mathcal{A}} \inf_{A' \in \mathcal{A}'} \left\{ \Big| P_{x^n}(A) - P(A) \Big| - \Big| P_{x^n_{(i)}}(A') - P(A') \Big| \right\} \\
&\le \sup_{A \in \mathcal{A}} \left\{ \Big| P_{x^n}(A) - P(A) \Big| - \Big| P_{x^n_{(i)}}(A) - P(A) \Big| \right\} \\
&\le \sup_{A \in \mathcal{A}} \Big| P_{x^n}(A) - P_{x^n_{(i)}}(A) \Big| \\
&= \frac{1}{n} \sup_{A \in \mathcal{A}} \left| \mathbf{1}_{\{x_i \in A\}} - \mathbf{1}_{\{x'_i \in A\}} \right| \\
&\le \frac{1}{n}.
\end{aligned}$$

Interchanging the roles of x^n and $x^n_{(i)}$, we obtain

$$g(x^n_{(i)}) - g(x^n) \le \frac{1}{n}.$$

Thus,

$$\left| g(x^n) - g(x^n_{(i)}) \right| \le \frac{1}{n}.$$

Note that this bound holds for all i and all choices of x^n and $x^n_{(i)}$. This means that the function g defined in (2.28) has bounded differences with $c_1 = \ldots = c_n = 1/n$. Consequently, we can use Theorem 2.3 to get

$$\mathbf{P}\left(|g(X^n) - \mathbf{E}g(X^n)| \ge \varepsilon \right) \le 2e^{-2n\varepsilon^2}.$$

This shows that the *uniform deviation* $g(X^n)$ concentrates sharply around its mean $\mathbf{E}g(X^n)$.

2.4.3. Uniform deviations continued. The same idea applies to arbitrary real-valued functions over X. Let $X^n = (X_1, \ldots, X_n)$ be as in the previous example. Given any function $f : \mathsf{X} \to [0, 1]$, Hoeffding's inequality tells us that

$$\mathbf{P}\left(|P_n(f) - \mathbf{E}f(X)| \ge \varepsilon \right) \le 2e^{-2n\varepsilon^2}.$$

However, just as in the previous example, in learning theory we are primarily interested in controlling the deviations of empirical means from true means simultaneously over whole classes of functions. To that end, let $\mathcal{F}$ be such a class consisting of functions $f : \mathsf{X} \to [0,1]$ and consider the *uniform deviation*

$$g(X^n) := \sup_{f \in \mathcal{F}} |P_n(f) - P(f)|.$$

An argument entirely similar to the one in the previous example[1] shows that this g has bounded differences with $c_1 = \ldots = c_n = 1/n$. Therefore, applying McDiarmid's inequality, we obtain

$$\mathbf{P}\left(|g(X^n) - \mathbf{E}g(X^n)| \ge \varepsilon\right) \le 2e^{-2n\varepsilon^2}.$$

We will see later that, for certain function classes $\mathcal{F}$, we will have $\mathbf{E}g(X^n) = O(1/\sqrt{n})$.

2.4.4. Kernel density estimation. For our final example, let $X^n = (X_1, \ldots, X_n)$ be an n-tuple of i.i.d. real-valued random variables whose common distribution P has a probability density function (pdf) f, i.e.,

$$P(A) = \int_A f(x) \mathrm{d}x$$

for any measurable set $A \subseteq \mathbb{R}$. We wish to estimate f from the sample X^n. A popular method is to use a *kernel estimate* (the book by Devroye and Lugosi [**DL01**] has plenty of material on density estimation, including kernel methods, from the viewpoint of statistical learning theory). To that end, we pick a nonnegative function $K : \mathbb{R} \to \mathbb{R}$ that integrates to one, $\int K(x)\mathrm{d}x = 1$ (such a function is called a *kernel*), as well as a positive *bandwidth* (or *smoothing constant*) $h > 0$ and form the estimate

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

It is not hard to verify[2] that $\widehat{f}_n$ is a valid pdf, i.e., that it is nonnegative and integrates to one. A common way of quantifying the performance of a density estimator is to use the L^1 distance to the true density f:

$$\|\widehat{f}_n - f\|_{L^1} = \int_{\mathbb{R}} \left|\widehat{f}_n(x) - f(x)\right| \mathrm{d}x.$$

Note that $\|\widehat{f}_n - f\|_{L^1}$ is a random variable since it depends on the random sample X^n. Thus, we can write it as a function $g(X^n)$ of the sample X^n. Leaving aside the problem of actually bounding $\mathbf{E}g(X^n)$, we can easily establish a concentration bound for it using McDiarmid's inequality. To do that, we need to check that g has bounded differences. Choosing x^n and

[1] Exercise: verify this!

[2] Another exercise!

$x^n_{(i)}$ as before, we have

$$\begin{aligned}
&g(x^n) - g(x^n_{(i)}) \\
&= \int_{\mathbb{R}} \left| \frac{1}{nh} \sum_{j=1}^{i-1} K\left(\frac{x-x_j}{h}\right) + \frac{1}{nh} K\left(\frac{x-x_i}{h}\right) + \frac{1}{nh} \sum_{j=i+1}^{n} K\left(\frac{x-x_j}{h}\right) - f(x) \right| \mathrm{d}x \\
&\qquad - \int_{\mathbb{R}} \left| \frac{1}{nh} \sum_{j=1}^{i-1} K\left(\frac{x-x_j}{h}\right) + \frac{1}{nh} K\left(\frac{x-x_i'}{h}\right) + \frac{1}{nh} \sum_{j=i+1}^{n} K\left(\frac{x-x_j}{h}\right) - f(x) \right| \mathrm{d}x \\
&\le \frac{1}{nh} \int_{\mathbb{R}} \left| K\left(\frac{x-x_i}{h}\right) - K\left(\frac{x-x_i'}{h}\right) \right| \mathrm{d}x \\
&\le \frac{2}{nh} \int_{\mathbb{R}} K\left(\frac{x}{h}\right) \mathrm{d}x \\
&= \frac{2}{n}.
\end{aligned}$$

Thus, we see that $g(X^n)$ has the bounded differences property with $c_1 = \ldots = c_n = 2/n$, so that

$$\mathbf{P}\left(|g(X^n) - \mathbf{E}g(X^n)| \ge \varepsilon\right) \le 2e^{-n\varepsilon^2/2}.$$

2.5. Subgaussian random variables

Often bounds proven for collections of Gaussian random variables can be readily extended to collections of random variables with similar exponential bounds, defined as follows.

Definition 2.1. *A random variable X is said to be* subgaussian with scale parameter ν, *if X has a finite mean and $\mathbf{E}\left[\mathrm{e}^{s\{X-\mathbf{E}[X]\}}\right] \le \mathrm{e}^{\frac{s^2\nu^2}{2}}$ for all $s \in \mathbb{R}$.*

Sometimes ν^2 is called the proxy variance because a Gaussian random variable with variance σ^2 is subgaussian for $\nu^2 = \sigma^2$. Also, if X is subgaussian with scale parameter ν, then the variance of X, σ^2, satisfies $\sigma^2 \le \nu^2$.

The definition of subgaussian random variable meshes very well with Hoeffding's bounds. Hoeffding's lemma, Lemma 2.1, can be restated as follows. If U is a random variable such that for some parameters a, b, $\mathbf{P}\{U \in [a,b]\} = 1$, U is subgaussian with scale parameter $\nu = \frac{b-a}{2}$. If $S_n = X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are independent random variables, such that X_i is subgaussian with scale parameter ν_i, then S_n is subgaussian with proxy variance given by $\nu^2 = \nu_1^2 + \cdots + \nu_n^2$. The methodology used to derive the Chernoff inequality shows that if X is subgaussian with scale parameter ν, then

$$\mathbf{P}\{X - \mathbf{E}[X] \ge \nu t\} \le \mathrm{e}^{-\frac{t^2}{2}} \tag{2.29}$$

for all $t \ge 0$. The bound (2.29) together with the equation $\nu^2 = \nu_1^2 + \cdots + \nu_n^2$ discussed above for sums of indepedent subgaussian random variables, is essentially a restatement of Hoeffding's inequality, Theorem 2.1.

The following lemma addresses the distribution of the maximum of a collection of subgaussian random variables.

Lemma 2.3 (Maximal lemma for subgaussian random variables). *Suppose $X_1, \ldots, X_n$ are mean zero, and each is subgaussian with scale parameter ν. (The variables are not assumed to be independent.) Then*

$$\mathbf{E}\left[\max_i X_i\right] \leq \nu\sqrt{2\log n} \tag{2.30}$$

$$\mathbf{P}\left\{\max_i X_i \geq \nu(\sqrt{2\log n}+t)\right\} \leq e^{-t\sqrt{2\log n}-\frac{t^2}{2}} \qquad \textit{for } t \geq 0 \tag{2.31}$$

Proof. Starting with Jensen's inequality, for any $s \geq 0$,

$$\begin{aligned}
\mathrm{e}^{s\mathbf{E}[\max_i X_i]} &\leq \mathbf{E}\left[\mathrm{e}^{s\max_i X_i}\right] = \mathbf{E}\left[\max_i \mathrm{e}^{sX_i}\right] \\
&\leq \mathbf{E}\left[\sum_i \mathrm{e}^{sX_i}\right] = \sum_i \mathbf{E}\left[\mathrm{e}^{sX_i}\right] \leq n\mathrm{e}^{s^2\nu^2/2}.
\end{aligned}$$

Taking the logarithm of each side yields $\mathbf{E}[\max_i X_i] \leq \frac{\log n}{s} + s\nu^2/2$. Letting $s = \frac{\sqrt{2\log n}}{\nu}$ yields (2.30).

Note that $\{\max_i X_i \geq c\} = \cup_i\{X_i \geq c\}$ so by the union bound, $\mathbf{P}\{\max_i X_i \geq c\} \leq \sum_i \mathbf{P}\{X_i \geq c\}$. By the assumptions and the bound (2.29), $\mathbf{P}\left\{X_i \geq \nu(\sqrt{2\log n}+t)\right\} \leq \exp(-(\sqrt{2\log n}+t)^2/2)$ for each i. Assembling with $c = \nu(\sqrt{2\log n}+t)$ yields

$$\mathbf{P}\left\{\max_i X_i \geq \nu(\sqrt{2\log n}+t)\right\} \leq n\exp(-(\sqrt{2\log n}+t)^2/2) = e^{-t\sqrt{2\log n}-\frac{t^2}{2}},$$

and (2.31) is proved. □

CHAPTER 3

Minima, convexity, strong convexity, and smoothness of functions

3.1. The minima of a function

Suppose f is a real-valued function with domain S. A point $x^* \in S$ is a *minimizer* of f if $f(x^*) \leq f(x)$ for all $x \in S$. The set of all minimizers of f over S is denoted by $\arg\min_{x\in S} f(x)$. It is possible there are no minimizers, but f must have an infimum, where the *infimum* of f is the maximum value $V \in \mathbb{R} \cup \{-\infty\}$ such that $f(x) \geq V$ for all $x \in S$. The infimum of f is denoted by $\inf_{y\in S} f(y)$. The set of minimizers of f is denoted by $\arg\min_{x\in S} f(x) = \{x \in S : f(x) = \inf_{y\in S} f(y)\}$. Maximizers of f are similarly related to the *supremum* of f, which satisfies $\sup_{y\in S} f(y) = -\inf_{y\in S} -f(y)$.

THEOREM 3.1. *(Weierstrass extreme value theorem) Suppose $f : S \to \mathbb{R}$ is a continuous function and the domain S is a sequentially compact set. (For example, S could be a closed, bounded subset of $\mathbb{R}^m$ for some m.) Then there exists a minimizer of f. In other words,* $\arg\min_{x\in S} f(x) \neq \varnothing$.

PROOF. Let $V = \inf_{x\in S} f(x)$. Note that $V \geq -\infty$. Let (x_n) denote a sequence of points in S such that $\lim_{n\to\infty} f(x_n) = V$. By the compactness of S, there is a subsequence (x_{n_k}) of the points that is convergent to some point $x^* \in S$. In other words, $\lim_{k\to\infty} x_{n_k} = x^*$. By the continuity of f, $f(x^*) = \lim_{k\to\infty} f(x_{n_k})$, and also the subsequence of values has the same limit as the entire sequence of values, so $\lim_{k\to\infty} f(x_{n_k}) = V$. Thus, $f(x^*) = V$, which implies the conclusion of the theorem. □

EXAMPLE 3.1. *(a) If $S = [0, 1)$ or $S = \mathbb{R}$ and $f(x) = \frac{1}{1+x^2}$, there is no minimizer. Theorem 3.1 doesn't apply because S is not compact. (b) If $S = [0, 1]$ and $f(x) = 1$ for $0 \leq x \leq 0.5$ and $f(x) = x$ for $0.5 < x \leq 1$ then there is no minimizer. Theorem 3.1 doesn't apply because f is not continuous.*

3.2. Derivatives of functions of several variables

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$. We say that f is differentiable at a point x if f is well enough approximated in a neighborhood of x by a linear approximation. Specifically, an $m \times n$ matrix $J(x)$ is the Jacobian of f at x if

$$\lim_{a\to x} \frac{\|f(a) - f(x) - J(x)(a - x)\|}{\|a - x\|} = 0$$

The Jacobian is also denoted by $\frac{\partial f}{\partial x}$ and if f is differentiable at x the Jacobian is given by a matrix of partial derivatives:

$$\frac{\partial f}{\partial x} = J = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

Moreover, according to the multidimensional differentiability theorem, a sufficient condition for f to be differentiable at x is for the partial derivatives $\frac{\partial f_i}{\partial x_j}$ to exist and be continuous in a neighborhood of x. In the special case $m = 1$ the gradient is the transpose of the derivative:

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}.$$

A function $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable at x if there is an $n \times n$ matrix $H(x)$, called the Hessian matrix, such that

$$\lim_{a \to x} \frac{\|f(a) - f(x) - J(x) \cdot (a - x) - \frac{1}{2}(a - x)^T H(x)(a - x)\|}{\|a - x\|^2} = 0.$$

The matrix $H(x)$ is also denoted by $\frac{\partial^2 f}{(\partial x)^2}(x)$, and is given by a matrix of second order partial derivatives:

$$\frac{\partial^2 f}{(\partial x)^2} = H = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}.$$

The function f is twice differentiable at x if both the first partial derivatives $\frac{\partial f}{\partial x_i}$ and second order partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$ exist and are continuous in a neighborhood of x.

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and if $x, \alpha \in \mathbb{R}^n$, then we can find the first and second derivatives of the function $t \mapsto f(x + \alpha t)$ from $\mathbb{R} \to \mathbb{R}$:

$$\begin{aligned} \frac{\partial f(x + \alpha t)}{\partial t} &= \sum_i \frac{\partial f}{\partial x_i}\bigg|_{x+\alpha t} \alpha_i = \alpha^T \nabla f(x + \alpha t). \\ \frac{\partial^2 f(x + \alpha t)}{(\partial t)^2} &= \sum_i \sum_j \frac{\partial^2 f}{\partial x_i \partial x_j}\bigg|_{x+\alpha t} \alpha_i \alpha_j \\ &= \alpha^T H(x + \alpha t)\alpha. \end{aligned}$$

If $H(y)$ is positive semidefinite for all y, in other words $\alpha^T H(y)\alpha \geq 0$ for all $\alpha \in \mathbb{R}^n$ and all y, then f is a convex function.

3.3. Convex sets and convex functions

Let $\mathcal{H}$ be a Hilbert space (defined in Section 4.1). For example, $\mathcal{H}$ could be Euclidean space $\mathbb{R}^d$ for some $d \geq 1$. A subset $\mathcal{F} \subseteq \mathcal{H}$ is *convex* if

$$f_1, f_2 \in \mathcal{F} \quad \Longrightarrow \quad \lambda f_1 + (1 - \lambda) f_2 \in \mathcal{F},\ \forall \lambda \in [0, 1].$$

A function $\varphi : \mathcal{F} \to \mathbb{R}$ is convex if

$$\varphi(\lambda f_1 + (1-\lambda) f_2) \leq \lambda \varphi(f_1) + (1-\lambda)\varphi(f_2), \qquad \forall f_1, f_2 \in \mathcal{F},\ \lambda \in [0,1].$$

An element $g \in \mathcal{H}$ is a *subgradient* of a convex function φ at $f \in \mathcal{F}$ if

$$\varphi(f') \geq \varphi(f) + \langle g, f' - f \rangle, \qquad \forall f' \in \mathcal{F}.$$

The set of all subgradients of φ at f is denoted by $\partial\varphi(f)$ and is referred to as the *subdifferential* of φ at f. We say that φ is *subdifferentiable* at f if $\partial\varphi(f) \neq \varnothing$. In particular, it can be shown that $\partial\varphi(f) \neq \varnothing$ for every f in the interior of $\mathcal{F}$. We say that φ is differentiable at f if $\partial\varphi(f)$ has only one element, in which case we refer to this element as the *gradient* of φ at f and denote it by $\nabla\varphi(f)$.

Given a convex function φ on $\mathcal{F}$, it is often of interest to find a minimizer of φ on $\mathcal{F}$. We have the following basic result:

Lemma 3.1 (First-order optimality condition). *Let $\varphi : \mathcal{F} \to \mathbb{R}$ be a differentiable convex function. The point $f^* \in \mathcal{F}$ is a minimizer of φ on $\mathcal{F}$ if and only if*

$$\langle \nabla\varphi(f^*), f - f^* \rangle \geq 0, \qquad \forall f \in \mathcal{F}. \tag{3.1}$$

Proof. To prove sufficiency, note that, by definition of the subgradient,

$$\varphi(f) \geq \varphi(f^*) + \langle g^*, f - f^* \rangle$$

for any $g^* \in \partial\varphi(f^*)$. If (3.1) holds, then $\varphi(f) \geq \varphi(f^*)$ for all $f \in \mathcal{F}$. (Note that here we do not require differentiability of φ.)

To prove necessity, let f^* be a minimizer of φ on $\mathcal{F}$, and suppose that (3.1) does not hold. That is, there exists some $f \in \mathcal{F}$, such that $\langle \nabla\varphi(f^*), f - f^* \rangle < 0$. By convexity of $\mathcal{F}$, $f^* + t(f - f^*) \in \mathcal{F}$ for all sufficiently small $t > 0$. Consider the function $F(t) := \varphi(f^* + t(f - f^*))$. Since φ is differentiable, so is F. By the chain rule, which holds in a Hilbert space, we have

$$F'(0) = \langle \nabla\varphi(f^* + t(f - f^*)), f - f^* \rangle \Big|_{t=0} = \langle \nabla\varphi(f^*), f - f^* \rangle < 0.$$

But this means $F(t) < F(0)$ for all small $t > 0$, which contradicts the optimality of f^*. $\square$

3.4. Strongly convex functions

Definition 3.1. *A function $\varphi : \mathcal{F} \to \mathbb{R}$ is m-*strongly convex *for some $m > 0$ if φ is convex, subdifferentiable, and*

$$\varphi(f') \geq \varphi(f) + \langle g, f' - f \rangle + \frac{m}{2}\|f - f'\|^2 \tag{3.2}$$

for all $f, f' \in \mathcal{F}$ and all $g \in \partial\varphi(f)$.

It is not hard to see that φ is m-strongly convex if and only if the function $\varphi(f) - \frac{m}{2}\|f\|^2$ is convex.

Lemma 3.2. *Suppose $\mathcal{F}$ is a nonempty, closed, convex subset of a Hilbert space $\mathcal{H}$, and $\varphi : \mathcal{F} \to \mathbb{R}$ is m-strongly convex for some $m > 0$. Then the following hold:*

(1) For any $f, f' \in \mathcal{F}$ and $0 \leq \lambda \leq 1$,

$$\varphi(\lambda f + (1-\lambda) f')) \leq \lambda\varphi(f) + (1-\lambda)\varphi(f') - \frac{\lambda(1-\lambda)m}{2}\|f - f'\|^2 \tag{3.3}$$

(2) *There exists a unique $f^* \in \mathcal{F}$ that minimizes φ. (Hence we write $f^* = \arg\min_{f\in\mathcal{F}} \varphi(f)$.)*
(3) *For any $f \in \mathcal{F}$, $\varphi(f) - \varphi(f^*) \geq \frac{m}{2}\|f - f^*\|^2$, where $f^* = \arg\min_{f\in\mathcal{F}} \varphi(f)$.*
(4) *(Stability of minimizers under Lipschitz perturbations) Suppose B is an L-Lipschitz function on $\mathcal{F}$, i.e.,*

$$|B(f) - B(f')| \leq L\|f - f'\|, \qquad \forall f, f' \in \mathcal{F}$$

and suppose $\widetilde{f}^*$ is a minimizer of $\varphi + B$ over $\mathcal{F}$. Then $\|f^* - \widetilde{f}^*\| \leq \frac{L}{m}$.
(5) *For any $f \in \mathcal{F}$, $\varphi(f) - \varphi(f^*) \leq \frac{1}{2m}\|\nabla\varphi(f)\|^2$.*
(6) *For any $f, f' \in \mathcal{F}$,*

$$\langle \nabla\varphi(f) - \nabla\varphi(f'), f - f'\rangle \geq m\|f - f'\|^2. \tag{3.4}$$

Proof. For part 1, let g be any choice of subgradient in $\partial\varphi(\lambda f + (1-\lambda)f')$. By the definition of strong convexity,

$$\varphi(f) \geq \varphi(\lambda f + (1-\lambda)f') + (1-\lambda)\langle g, f - f'\rangle + \frac{(1-\lambda)^2 m}{2}\|f - f'\|^2 \tag{3.5}$$

$$\varphi(f') \geq \varphi(\lambda f + (1-\lambda)f') \ - \ \lambda\langle g, f - f'\rangle \ + \ \frac{\lambda^2 m}{2}\|f - f'\|^2 \tag{3.6}$$

Multiply both sides of (3.5) by λ and both sides of (3.6) by $1-\lambda$ and then add the equations together to obtain (3.3). Parts 2 and 4 are proved in homework. For Part 3, note that the 0 element of $\mathcal{H}$ is a subgradient of φ at f^*, so Part 3 follows from (3.2) with $f = f^*$ and $g = 0$. For Part 5, the definition (3.2) yields for any f,

$$\begin{aligned}
\varphi(f^*) - \varphi(f) &\geq \langle \nabla\varphi(f), f^* - f\rangle + \frac{m}{2}\|f - f^*\|^2 \\
&\geq \min_g \left\{ \langle \nabla\varphi(f), g\rangle + \frac{m}{2}\|g\|^2 \right\} \\
&= -\frac{1}{2m}\|\nabla\varphi(f)\|^2.
\end{aligned}$$

For Part 6, we have

$$\begin{aligned}
\varphi(f') &\geq \varphi(f) + \langle \nabla\varphi(f), f' - f\rangle + \frac{m}{2}\|f - f'\|^2 \\
\varphi(f) &\geq \varphi(f') + \langle \nabla\varphi(f'), f - f'\rangle + \frac{m}{2}\|f - f'\|^2.
\end{aligned}$$

Adding these two inequalities and rearranging, we get (3.4). □

Example 3.2. *Let $\varphi(f) = \frac{mf^2}{2}$ and $\widetilde{\varphi}(f) = \frac{mf^2}{2} - Lf$ for $f \in \mathbb{R}$. Then $|f^* - \widetilde{f}^*| = \left|0 - \frac{L}{m}\right| = \frac{L}{m}$. In this case, the bound in (3.2) part 4 holds with equality.*

3.5. Smooth convex functions

Let $\mathcal{F}$ denote a Hilbert space. We say that a differentiable (not necessarily convex) function $\varphi : \mathcal{F} \to \mathbb{R}$ is *M-smooth*, for some $M \geq 0$, if the gradient mapping $f \mapsto \nabla\varphi(f)$ is M-Lipschitz:

$$\|\nabla\varphi(f) - \nabla\varphi(f')\| \leq M\|f - f'\|, \qquad \forall f, f' \in \mathcal{F}. \tag{3.7}$$

We give some properties of smooth convex functions. Smoothness of a convex function is very helpful in case a gradient descent algorithm is used to minimize the function.

LEMMA 3.3. *(Smooth functions in finite dimensions) Suppose $\mathcal{F}$ is a d-dimensional Euclidean space, and φ is twice continuously differentiable. (By convexity, the Hessian satisfies, $\nabla^2\phi(f) \succeq 0$ for all f, where $A \succeq B$ for symmetric matrices A and B means $A-B$ is positive semidefinite.) If φ is convex, then φ is M-smooth if $\nabla^2\phi(f) \preceq MI$ for all f.*

PROOF. Note that $\nabla\varphi(f') - \nabla\varphi(f) = \int_0^1 \frac{d}{dt}(\nabla\varphi(f_t))\,dt$ where $f_t = (1-t)f + tf'$. Taking the elements of $\mathcal{F}$ to be column vectors, the definition of gradient and the chain rule imply:

$$\frac{d}{dt}(\nabla\varphi(f_t))_i = \frac{d}{dt}\left(\frac{\partial\varphi}{\partial f_i}(f_t)\right) = \sum_j \frac{\partial^2\varphi(f_t)}{\partial f_i \partial f_j}(f'_j - f_j)$$

Thus,

$$\nabla\varphi(f') - \nabla\varphi(f) = \int_0^1 \left(\nabla^2\varphi(f_t)\right)(f'-f)dt = H(f'-f)$$

where H is the $d \times d$ matrix $H = \int_0^1 \nabla^2\varphi(f_t)dt$. Since $\nabla^2\varphi(f_t) \preceq MI$ for each t it follows readily that $H \preceq MI$. So the spectral (i.e. operator) norm of H satisfies $\|H\| \le M$. Hence $\|\nabla\varphi(f') - \nabla\varphi(f)\| = \|H(f'-f)\| \le \|H\|\|f-f'\| \le M\|f-f'\|$. So φ is M-smooth. $\square$

LEMMA 3.4. *Suppose φ is an M-smooth convex function, and given $\alpha > 0$ define the gradient descent iteration map $G_{\varphi,\alpha}$ by $G_{\varphi,\alpha}(f) = f - \alpha\nabla\varphi(f)$. Then the following hold.*

(a) For any $f, f' \in \mathcal{F}$, $\varphi(f') - \varphi(f) \le \langle \nabla\varphi(f), f'-f\rangle + \frac{M}{2}\|f'-f\|^2$.

(b) For any $\alpha > 0$ and $f \in \mathcal{F}$,

$$\varphi(G_{\varphi,\alpha}(f)) \le \varphi(f) - \alpha\left(1 - \frac{\alpha M}{2}\right)\|\nabla\varphi(f)\|^2.$$

In particular,

$$\begin{aligned} \varphi(G_{\varphi,\alpha}(f)) &\le \varphi(f) && \text{if } 0 \le \alpha \le \tfrac{2}{M} \\ \varphi(G_{\varphi,\frac{1}{M}}(f)) &\le \varphi(f) - \tfrac{1}{2M}\|\nabla\varphi(f)\|^2 && \text{(special case } \alpha = \tfrac{1}{M}\text{)} \\ \varphi(f^*) &\le \varphi(f) - \tfrac{1}{2M}\|\nabla\varphi(f)\|^2 && \text{if } f^* \text{ is a global minimizer of } \varphi. \end{aligned}$$

(c) $\varphi(f') - \varphi(f) \ge \langle \nabla\varphi(f), f'-f\rangle + \frac{1}{2M}\|\nabla\varphi(f') - \nabla\varphi(f)\|^2$ for any $f, f' \in \mathcal{F}$.

(d) (Co-coercive *property of the gradient of an M-smooth convex function):*

$$\langle \nabla\varphi(f') - \nabla\varphi(f), f'-f\rangle \ge \frac{1}{M}\|\nabla\varphi(f') - \nabla\varphi(f)\|^2.$$

(e) (Contraction property of gradient descent map) If $0 \le \alpha < 2/M$, then

$$\|G_{\varphi,\alpha}(f) - G_{\varphi,\alpha}(f')\| \le \|f-f'\| \qquad \forall f, f' \in \mathcal{F}.$$

*(f) (Strict contraction property of gradient descent map) (As all results in this chapter, this result is classical [**Pol87**].) If $0 \le \alpha < 2/M$ and φ is also m-strongly convex for some $m > 0$, then*

$$\|G_{\varphi,\alpha}(f) - G_{\varphi,\alpha}(f')\| \le \eta\|f-f'\| \qquad \forall f, f' \in \mathcal{F},$$

where $\eta^2 = 1 - \alpha m(2-\alpha M) < 1$. In particular, if f^* is the minimizer of φ, $\|G_{\varphi,\alpha}(f) - f^*\| \le \eta\|f-f^*\|$ for all $f \in \mathcal{F}$. Also, if $0 < \alpha \le \frac{1}{M}$, then $\eta \le \sqrt{1-\alpha m} \le 1 - \frac{\alpha m}{2}$.)

PROOF. (a) Note that $\varphi(f') - \varphi(f) = \int_0^1 \frac{d\varphi(f_t)}{dt} dt$, where $f_t = (1-t)f + tf'$, and by the chain rule, $\frac{d\varphi(f_t)}{dt} = \langle \nabla\varphi(f_t), f' - f \rangle$. So

$$\begin{aligned}
\varphi(f') - \varphi(f) - \langle \nabla\varphi(f), f' - f \rangle &= \int_0^1 \langle \nabla\varphi(f_t) - \nabla\varphi(f), f' - f \rangle dt \\
&\leq \int_0^1 \|\nabla\varphi(f_t) - \nabla\varphi(f)\| \|f' - f\| dt \\
&\leq \int_0^1 M \|f_t - f\| \|f' - f\| dt \\
&= \int_0^1 Mt \|f' - f\|^2 dt = \frac{M}{2} \|f' - f\|^2
\end{aligned}$$

(b) Follows from (a) by letting $f' = G_{\varphi,\alpha}(f)$ and using $f' - f = -\alpha\nabla\varphi(f)$.

(c) Fix $f, f' \in \mathcal{F}$. For any g, the inequality holds for a function φ if and only if it holds for $\widetilde{\varphi}(f) \triangleq \varphi(f) - \langle g, f \rangle$, because the contributions due to g are the same on each side of the inequality. And $\widetilde{\varphi}$ is also M-smooth. Letting $g = \nabla\varphi(f)$ makes $\nabla\widetilde{\varphi}(f) = 0$, so that f is a global minimizer of $\widetilde{\varphi}$. The inequality for $\widetilde{\varphi}$ is true because it reduces to the last inequality of part (b).

(d) By part (c), for any $f, f' \in \mathcal{F}$,

$$\begin{aligned}
\varphi(f') - \varphi(f) &\geq \langle \nabla\varphi(f), f' - f \rangle + \frac{1}{2M} \|\nabla\varphi(f') - \nabla\varphi(f)\|^2 \\
\varphi(f) - \varphi(f') &\geq \langle \nabla\varphi(f'), f - f' \rangle + \frac{1}{2M} \|\nabla\varphi(f') - \nabla\varphi(f)\|^2
\end{aligned}$$

Adding the respective sides of these equations and rearranging yields the desired inequality.

(e-f) We can prove (e) and (f) together, because taking $m = 0$ in (f) corresponds to (e). Observe that

$$\begin{aligned}
\|G_{\varphi,\alpha}(f) - G_{\varphi,\alpha}(f')\|^2 &= \|f - f' - \alpha\left(\nabla\varphi(f) - \nabla\varphi(f')\right)\| 2 \\
&= \|f - f'\|^2 - 2\alpha\langle f - f', \nabla\varphi(f) - \nabla\varphi(f') \rangle + \alpha^2 \|\nabla\varphi(f) - \nabla\varphi(f')\|^2 \\
&\overset{(a)}{\leq} \|f - f'\|^2 - \alpha(2 - \alpha M)\langle \nabla\varphi(f) - \nabla\varphi(f'), f - f' \rangle \\
&\overset{(b)}{\leq} \eta^2 \|f - f'\|^2,
\end{aligned}$$

where (a) follows from the co-coercive property of Lemma 3.4(d) and (b) follows from Lemma 3.2(6), which holds in the special case $m = 0$ by the convexity of φ. □

REMARK 3.1. *The properties of strong convexity and smoothness of convex functions are strongly related. Roughly speaking, the first gives a lower bound on the curvature of a function and the second an upper bound. Moreover, the properties are dual properties for the Legendre-Fenchel transform. If φ is a closed (means sets of the form $\varphi \leq t$ are closed), convex function and $\varphi^*(y) = \sup_{x \in \mathsf{X}} \langle x, y \rangle - \varphi(x)$, then φ is m-strongly convex if and only if φ^* is $1/m$-smooth.*

CHAPTER 4

Function spaces determined by kernels

A powerful way of building complicated classifiers is to use linear combinations of simple functions. The number of simple functions used is potentially infinite, so it is natural to consider an infinite dimensional generalization of ordinary finite dimensional Euclidean space, known as a Hilbert space. Some particular Hilbert spaces of functions are naturally specified in terms of a *kernel*. Kernel methods are popular in machine learning for a variety of reasons, not the least of which is that any algorithm that operates in a Euclidean space and relies only on the computation of inner products between feature vectors can be modified to work with any suitably well-behaved kernel. (See the representer theorem in Section 8.6.)

4.1. The basics of Hilbert spaces

Hilbert spaces are generalizations of the usual finite-dimensional Euclidean spaces. While Hilbert spaces can be infinite dimensional, they retain many of the important key properties of finite-dimensional Euclidean space. As long as an inner product is defined for pairs of elements in the space, a notion of *angle* and, consequently, orthogonality, can be defined for two elements in the space. Moreover, a Hilbert space has certain favorable convergence properties, yielding (unique) linear projections of their elements onto closed linear subspaces, or, more generally, unique nonlinear projections onto closed convex sets.

Definition 4.1. *A real vector space* V *is an* inner product space *if there exists a function* $\langle \cdot, \cdot \rangle_{\mathsf{V}} : \mathsf{V} \times \mathsf{V} \to \mathbb{R}$*, which is:*

(1) Symmetric: $\langle v, v' \rangle_{\mathsf{V}} = \langle v', v \rangle_{\mathsf{V}}$ *for all* $v, v' \in \mathsf{V}$
(2) Bilinear: $\langle \alpha v_1 + \beta v_2, v' \rangle_{\mathsf{V}} = \alpha \langle v_1, v' \rangle_{\mathsf{V}} + \beta \langle v_2, v' \rangle_{\mathsf{V}}$ *for* $\alpha, \beta \in \mathbb{R}$ *and* $v_1, v_2, v' \in \mathsf{V}$
(3) Positive definite: $\langle v, v \rangle_{\mathsf{V}} \geq 0$ *for all* $v \in \mathsf{V}$*, and* $\langle v, v \rangle_{\mathsf{V}} = 0$ *if and only if* $v = 0$

Let $(\mathsf{V}, \langle \cdot, \cdot \rangle_{\mathsf{V}})$ be an inner product space. Then we can define a *norm* on V via

$$\|v\|_{\mathsf{V}} := \sqrt{\langle v, v \rangle_{\mathsf{V}}}.$$

It is easy to check that this is, indeed, a norm —

(1) It is homogeneous: for any $v \in \mathsf{V}$ and any $\alpha \in \mathbb{R}$,

$$\|\alpha v\|_{\mathsf{V}} = \sqrt{\langle \alpha v, \alpha v \rangle_{\mathsf{V}}} = \sqrt{\alpha^2 \langle v, v \rangle_{\mathsf{V}}} = |\alpha| \sqrt{\langle v, v \rangle_{\mathsf{V}}} = |\alpha| \cdot \|v\|_{\mathsf{V}}$$

(2) It satisfies the triangle inequality: for any $v, v' \in \mathsf{V}$,

$$\|v + v'\|_{\mathsf{V}} \leq \|v\|_{\mathsf{V}} + \|v'\|_{\mathsf{V}}. \tag{4.1}$$

To prove this, we first need to establish another key property of $\| \cdot \|_{\mathsf{V}}$: the *Cauchy–Schwarz inequality*, which generalizes its classical Euclidean counterpart and says that

$$|\langle v, v' \rangle_{\mathsf{V}}| \leq \|v\|_{\mathsf{V}} \|v'\|_{\mathsf{V}}. \tag{4.2}$$

To prove (4.2), start with the observation $\|v - \lambda v'\|_{\mathsf{V}}^2 = \langle v - \lambda v', v - \lambda v' \rangle_{\mathsf{V}} \geq 0$ for any $\lambda \in \mathbb{R}$. Expanding this yields

$$\langle v - \lambda v', v - \lambda v' \rangle_{\mathsf{V}} = \lambda^2 \|v'\|_{\mathsf{V}}^2 - 2\lambda \langle v, v' \rangle_{\mathsf{V}} + \|v\|_{\mathsf{V}}^2 \geq 0.$$

This is a quadratic function of λ, and the above implies that its graph does not cross the horizontal axis. Therefore,

$$4|\langle v, v' \rangle_{\mathsf{V}}|^2 \leq 4\|v\|_{\mathsf{V}}^2 \|v'\|_{\mathsf{V}}^2 \qquad \Longleftrightarrow \qquad |\langle v, v' \rangle_{\mathsf{V}}| \leq \|v\|_{\mathsf{V}} \|v'\|_{\mathsf{V}},$$

and the Cauchy-Schwarz inequality (4.2) is proved. Thus,

$$\begin{aligned} \|v + v'\|_{\mathsf{V}}^2 &\equiv \langle v + v', v + v' \rangle_{\mathsf{V}} \\ &= \langle v, v \rangle_{\mathsf{V}} + \langle v, v' \rangle_{\mathsf{V}} + \langle v', v \rangle_{\mathsf{V}} + \langle v', v' \rangle_{\mathsf{V}} \\ &= \|v\|_{\mathsf{V}}^2 + 2\langle v, v' \rangle_{\mathsf{V}} + \|v'\|_{\mathsf{V}}^2 \\ &\leq \|v\|_{\mathsf{V}}^2 + 2\|v\|_{\mathsf{V}}^2 \|v'\|_{\mathsf{V}}^2 + \|v'\|_{\mathsf{V}}^2 = (\|v\|_{\mathsf{V}} + \|v'\|_{\mathsf{V}})^2 \end{aligned}$$

where we've used the definition of norm, the bilinear and symmetry properties of the inner product, and the Cauchy–Schwarz inequality. Since all norms are nonnegative, we can take square roots of both sides to get the triangle inequality.

(3) Finally, $\|v\|_{\mathsf{V}} \geq 0$, and $\|v\|_{\mathsf{V}} = 0$ if and only if $v = 0$ – this is obvious from definitions.

Thus, an inner product space can be equipped with a norm that has certain special properties (mainly, the Cauchy–Schwarz inequality, since a lot of useful things follow from it alone). Now that we have a norm, we can talk about *convergence* of sequences in V:

Definition 4.2. *A sequence of elements of* V, $\{v_n\}_{n=1}^{\infty}$, converges *to* $v \in \mathsf{V}$ *if*

$$\lim_{n \to \infty} \|v_n - v\|_{\mathsf{V}} = 0. \tag{4.3}$$

Any norm-convergent sequence has the property that, as n gets larger, its elements get closer and closer to one another. Specifically, suppose that $\{v_n\}$ converges to v. Then (4.3) implies that for any $\varepsilon > 0$ we can choose N large enough, so that $\|v_n - v\|_{\mathsf{V}} < \varepsilon/2$ for all $n \geq N$. But the triangle inequality gives

$$\|v_n - v_m\|_{\mathsf{V}} \leq \|v_n - v\|_{\mathsf{V}} + \|v_m - v\|_{\mathsf{V}} < \varepsilon, \qquad \forall \quad m, n \geq N.$$

In other words,

$$\lim_{\min(m,n) \to \infty} \|v_n - v_m\| = 0. \tag{4.4}$$

Any sequence $\{v_n\}$ that has the property (4.4) is called a *Cauchy sequence*. We have just proved that any convergent sequence is Cauchy. However, the converse is not necessarily true: a Cauchy sequence does not have to be convergent. This motivates the following definition:

Definition 4.3. *A normed space* $(\mathsf{V}, \|\cdot\|_{\mathsf{V}})$ *is* complete *if any Cauchy sequence* $\{v_n\}$ *of its elements is convergent. If the norm* $\|\cdot\|_{\mathsf{V}}$ *is induced by an inner product and if it is complete, then we say that* V *is a* Hilbert space*.*

For an example of an inner product space that is not complete, consider the space of sequences of the form $x = (x_1, x_2, \ldots)$ with $x_i \in \mathbb{R}$ such that only finitely many of the x_i's are nonzero, with the inner product $\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i$. There is a standard procedure of starting with an inner product and the corresponding normed space and then *completing* it by adding the limits of all Cauchy sequences. We will not worry too much about this procedure. Here are a few standard examples of Hilbert spaces:

(1) The Euclidean space $\mathsf{V} = \mathbb{R}^d$ with the usual inner product

$$\langle v, v' \rangle = \sum_{j=1}^{d} v_j v'_j.$$

The corresponding norm is the familiar ℓ_2 norm, $\|v\| = \sqrt{\langle v, v \rangle}$.

(2) More generally, if A is a positive definite $d \times d$ matrix, then the inner product

$$\langle v, v' \rangle_A := \langle v, Av' \rangle$$

induces the A-weighted norm $\|v\|_A := \sqrt{\langle v, v \rangle_A} = \sqrt{\langle v, Av \rangle}$, which makes $\mathbb{R}^d$ into a Hilbert space. The preceding example is a special case with $A = I_d$, the $d \times d$ identity matrix.

(3) The space $L^2(\mathbb{R}^d)$ of all *square-integrable* functions $f : \mathbb{R}^d \to \mathbb{R}$, i.e.,

$$\int_{\mathbb{R}^d} f^2(x)\mathrm{d}x < \infty,$$

is a Hilbert space with the inner product

$$\langle f, g \rangle_{L^2(\mathbb{R}^d)} := \int_{\mathbb{R}^d} f(x)g(x)\mathrm{d}x$$

and the corresponding norm

$$\|f\|_{L^2(\mathbb{R}^d)} := \sqrt{\int_{\mathbb{R}^d} f^2(x)\mathrm{d}x}.$$

(4) Let $(\Omega, \mathcal{B}, P)$ be a probability space. Then the space $L^2(P)$ of all real-valued random variables $X : \Omega \to \mathbb{R}$ with finite second moment, i.e.,

$$\mathbf{E}X^2 = \int_{\Omega} X^2(\omega)P(\mathrm{d}\omega) < +\infty,$$

is a Hilbert space with the inner product

$$\langle X, X' \rangle_{L^2(P)} := \mathbf{E}[XX'] = \int_{\Omega} X(\omega)X'(\omega)P(\mathrm{d}\omega)$$

and the corresponding norm

$$\|X\|_{L^2(P)} := \sqrt{\int_{\Omega} |X(\omega)|^2 P(\mathrm{d}\omega)} \equiv \sqrt{\mathbf{E}X^2}.$$

From now on, we will denote a typical Hilbert space by $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$; the induced norm will be denoted by $\|\cdot\|_{\mathcal{H}}$.

An enormous advantage of working with Hilbert spaces is the availability of the notion of *orthogonality* and *orthogonal projection*. Two elements h, g of a Hilbert space $\mathcal{H}$ are said to be *orthogonal* if $\langle h, g\rangle_{\mathcal{H}} = 0$.

A subset $\mathcal{H}_1$ of $\mathcal{H}$ is defined to be closed if the limit of any convergent sequence $\{h_n\}$ of elements of $\mathcal{H}_1$ is also contained in $\mathcal{H}_1$. A subset $\mathcal{H}_1$ of $\mathcal{H}$ is a linear subspace if $v_1 + v_2 \in \mathcal{H}_1$ whenever $v_1, v_2 \in \mathcal{H}_1$. We have the following basic facts:

PROPOSITION 4.1 (Projection onto a subspace of a Hilbert space). *Let $\mathcal{H}_1$ be a closed linear subspace of a Hilbert space $\mathcal{H}$. For any $g \in \mathcal{H}$ there is a* unique *solution of the optimization problem*

$$\text{minimize } \|h-g\| \text{ subject to } h \in \mathcal{H}_1.$$

The solution is denoted by Πg and is called the projection *of g onto $\mathcal{H}_1$. We can write $\Pi g = \arg\min_{h\in\mathcal{H}_1}\|h-g\|$. The projection mapping Π has the following properties:*

(1) It is a linear operator: $\Pi(ag+bg') = a\Pi g + b\Pi g'$ for $a, b \in \mathbb{R}$, $g, g' \in \mathcal{H}$.
(2) $\Pi^2 = \Pi$, i.e., $\Pi(\Pi g) = \Pi g$ for any $g \in \mathcal{H}$.
(3) If $g \in \mathcal{H}_1$, then $\Pi g = g$.
(4) For any $g \in \mathcal{H}$ and any $h \in \mathcal{H}_1$, $\langle \Pi g, h\rangle_{\mathcal{H}} = \langle g, h\rangle_{\mathcal{H}}$.

Let $\mathcal{H}_1^\perp$ be the set of all $h^\perp \in \mathcal{H}$, such that $\langle g, h^\perp\rangle_{\mathcal{H}} = 0$ for all $g \in \mathcal{H}_1$. Then:

(1) $\mathcal{H}_1^\perp$ is also a closed linear subspace of $\mathcal{H}$.
(2) Any element g of $\mathcal{H}$ can be uniquely decomposed as $g = h + h^\perp$, where $h \in \mathcal{H}_1$ and $h^\perp \in \mathcal{H}_1^\perp$. Furthermore, $h = \Pi(g)$.

REMARK 4.1. It is important for $\mathcal{H}_1$ to be a *closed* linear subspace of $\mathcal{H}$ for the above results to hold.

PROPOSITION 4.2 (Projections onto a closed convex subset of a Hilbert space). *Let $\mathcal{F}$ be a nonempty, closed, convex subset of $\mathcal{H}$. Define the projection operator $\Pi : \mathcal{H} \to \mathcal{F}$ as*

$$\Pi(g) := \arg\min_{f\in\mathcal{F}} \|f-g\|^2. \tag{4.5}$$

Then:

(a) For any g, $\Pi(g)$ exists and is unique.
(b) The projection operator is nonexpansive, i.e., for any $g, g' \in \mathcal{H}$,

$$\|\Pi(g) - \Pi(g')\| \le \|g - g'\|. \tag{4.6}$$

PROOF. (a) The function $\varphi(f) := \frac{1}{2}\|f-g\|^2$ is strongly convex. Hence, by Lemma 3.2, it has a unique minimizer in $\mathcal{F}$.

(b) Note that $\nabla\varphi(f) = f - g$. By the first-order optimality condition, for any $f' \in \mathcal{F}$,

$$\langle \nabla\varphi(\Pi(g)), f' - \Pi(g)\rangle = \langle \Pi(g) - g, f' - \Pi(g)\rangle \ge 0. \tag{4.7}$$

In particular, using this with $f' = \Pi(g')$, we get

$$\langle \Pi(g) - g, \Pi(g) - \Pi(g')\rangle \le 0. \tag{4.8}$$

Similarly,

$$\langle g' - \Pi(g'), \Pi(g) - \Pi(g')\rangle \le 0. \tag{4.9}$$

Adding the inequalities (4.8) and (4.9) and rearranging, we obtain

$$\|\Pi(g) - \Pi(g')\|^2 \le \langle g - g', \Pi(g) - \Pi(g')\rangle. \tag{4.10}$$

Applying Cauchy–Schwarz to the right-hand side and canceling the factor of $\|\Pi(g) - \Pi(g')\|$ from both sides gives (4.6). □

4.2. Reproducing kernel Hilbert spaces

A reproducing kernel Hilbert space (RKHS) is a family of functions on some set X that forms a Hilbert space, with an associated kernel, as we describe later. Often in practice, the norm of a function is an appropriate measure of the complexity of the function. Such classes of functions are well suited to statistical learning theory because it makes sense to adjust the complexity of the predictors/classifiers based on the availability of the data. To start with, let us define what we mean by a kernel. We will stick to Euclidean feature spaces X, although everything works out if X is an arbitrary separable metric space.

Definition 4.4. *Let X be a closed subset of $\mathbb{R}^d$. A real-valued function $K : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ is called a* Mercer kernel *provided the following conditions are met:*

(1) It is symmetric, *i.e., $K(x, x') = K(x', x)$ for any $x, x' \in \mathsf{X}$.*
(2) It is continuous, *i.e., if $\{x_n\}$ is a sequence of points in X converging to a point x, then*

$$\lim_{n\to\infty} K(x_n, x') = K(x, x'), \qquad \forall x' \in \mathsf{X}.$$

(3) It is positive semidefinite, *i.e., for all $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and all $x_1, \dots, x_n \in \mathsf{X}$,*

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \ge 0. \tag{4.11}$$

Remark 4.2. Another way to interpret the positive semidefiniteness condition is as follows. For any n-tuple $x^n = (x_1, \dots, x_n) \in \mathsf{X}^n$, define the $n \times n$ *kernel Gram matrix*

$$G_K(x^n) := [K(x_i, x_j)]_{i,j=1}^{n}.$$

Then (4.11) is equivalent to saying that $G_K(x^n)$ is positive semidefinite in the usual sense, i.e., for any vector $v \in \mathbb{R}^n$ we have

$$\langle v, G_K(x^n)v\rangle \ge 0.$$

Remark 4.3. From now on, we will just say "kernel," but always mean "Mercer kernel."

Here are some examples of kernels. Additional examples are provided in the next section.

(1) With $\mathsf{X} = \mathbb{R}^d$, $K(x, x') = \langle x, x'\rangle$, the usual Euclidean inner product.
(2) A more general class of kernels based on the Euclidean inner product can be constructed as follows. Let $\mathsf{X} = \{x \in \mathbb{R}^d : \|x\| \le R\}$; choose any sequence $\{a_j\}_{j=0}^{\infty}$ of nonnegative reals such that

$$\sum_{j=0}^{\infty} a_j R^{2j} < \infty.$$

Then

$$K(x, x') = \sum_{j=0}^{\infty} a_j \langle x, x' \rangle^j$$

is a kernel.

(3) Let $\mathsf{X} = \mathbb{R}^d$, and let $k : \mathbb{R}^d \to \mathbb{R}$ be a continuous function, which is *reflection-symmetric*, i.e., $k(-x) = k(x)$ for all x. Then $K(x, x') := k(x - x')$ is a kernel provided the Fourier transform of k,

$$\widehat{k}(\xi) := \int_{\mathbb{R}^d} e^{-i\langle \xi, x \rangle} k(x) \mathrm{d}x,$$

is nonnegative. A prime example is the *Gaussian kernel*, induced by the function $k(x) = e^{-\gamma \|x\|^2}$.

In all of the above cases, the first two properties of a Mercer kernel are easy to check. The third, i.e., positive semidefiniteness, requires a bit more work. For details, consult Section 2.5 of the book by Cucker and Zhou [**CZ07**].

Suppose we have a fixed kernel K on our feature space X (which we assume to be a closed subset of $\mathbb{R}^d$). Let $\mathcal{L}_K(\mathsf{X})$ be the *linear span* of the set $\{K(x', \cdot) : x' \in \mathsf{X}\}$, i.e., the set of all functions $f : \mathsf{X} \to \mathbb{R}$ of the form

$$f(x) = \sum_{j=1}^{N} c_j K(x_j, x) \tag{4.12}$$

for all possible choices of $N \in \mathbb{N}$, $c_1, \ldots, c_N \in \mathbb{R}$, and $x_1, \ldots, x_N \in \mathsf{X}$. It is easy to see that $\mathcal{L}_K(\mathsf{X})$ is a *vector space*: for any two functions f, f' of the form (4.12), their sum is also of that form; if we multiply any $f \in \mathcal{L}_K(\mathsf{X})$ by a scalar $c \in \mathbb{R}$, we will get another element of $\mathcal{L}_K(\mathsf{X})$; and the zero function is clearly in $\mathcal{L}_K(\mathsf{X})$. It turns out that, for any (Mercer) kernel K, we can *complete* $\mathcal{L}_K(\mathsf{X})$ into a *Hilbert space* of functions that can potentially represent *any* continuous function from X into $\mathbb{R}$, provided K is chosen appropriately.

The following result is essential (for the proof, see Section 2.4 of Cucker and Zhou [**CZ07**]):

THEOREM 4.1. *Let X be a closed subset of $\mathbb{R}^d$, and let $K : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ be a Mercer kernel. Then there exists a unique Hilbert space $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$ of real-valued functions on X with the following properties:*

(1) For all $x \in \mathsf{X}$, the function $K_x(\cdot) := K(x, \cdot)$ is an element of $\mathcal{H}_K$, and $\langle K_x, K_{x'} \rangle_K = K(x, x')$ for all $x, x' \in \mathsf{X}$.

(2) The linear space $\mathcal{L}_K(\mathsf{X})$ is dense *in $\mathcal{H}_K$, i.e., for any $f \in \mathcal{H}_K$ and any $\varepsilon > 0$ there exist some $N \in \mathbb{N}$, $c_1, \ldots, c_N \in \mathbb{R}$, and $x_1, \ldots, x_N \in \mathsf{X}$, such that*

$$\left\| f - \sum_{j=1}^{N} c_j K_{x_j} \right\|_K < \varepsilon.$$

(3) For all $f \in \mathcal{H}_K$ and all $x \in \mathsf{X}$,

$$f(x) = \langle K_x, f \rangle_K. \tag{4.13}$$

Moreover, the functions in $\mathcal{H}_K$ are continuous. The Hilbert space $\mathcal{H}_K$ is called the Reproducing Kernel Hilbert Space (RKHS) *associated with K; the property (4.13) is referred to as the* reproducing kernel property.

REMARK 4.4. The reproducing kernel property essentially states that the value of any function $f \in \mathcal{H}_K$ at any point $x \in \mathsf{X}$ can be extracted by taking the inner product of f and the function $K_x(\cdot) = K(x, \cdot)$, i.e., a copy of the kernel K centered at the point x. It is easy to prove when $f \in \mathcal{L}_K(\mathsf{X})$. Indeed, if f has the form (4.12), then

$$\begin{aligned}\langle f, K_x \rangle_K &= \left\langle \sum_{j=1}^{N} c_j K_{x_j}, K_x \right\rangle_K \\ &= \sum_{j=1}^{N} c_j \langle K_{x_j}, K_x \rangle_K \\ &= \sum_{j=1}^{N} c_j K(x_j, x) \\ &= f(x).\end{aligned}$$

Since any $f \in \mathcal{H}_K$ can be expressed as a limit of functions from $\mathcal{L}_K(\mathsf{X})$, the proof of (4.13) for a general f follows by continuity.

For any function $f : \mathsf{X} \to \mathbb{R}$, define the *sup norm* by $\|f\|_\infty := \sup_{x \in \mathsf{X}} |f(x)|$.

LEMMA 4.1. *Let $\mathcal{H}_K$ be the RHKS generated by a kernel K and let $C_K = \sup_{x \in \mathsf{X}} \sqrt{K(x,x)}$. If $C_K < \infty$, the values of any function $f \in \mathcal{H}_K$ can be bounded in terms of the kernel as follows:*

$$\|f\|_\infty \le C_K \|f\|_K. \tag{4.14}$$

PROOF. For any $f \in \mathcal{H}_K$ and $x \in \mathsf{X}$,

$$|f(x)| = |\langle f, K_x \rangle_K| \le \|f\|_K \|K_x\|_K = \|f\|_K \sqrt{K(x,x)},$$

where the first step is by the reproducing kernel property, and the second step is by Cauchy–Schwarz. Taking the supremum of both sides over X yields (4.14). □

4.3. Kernels and weighted inner products

This section gives a fuller explanation of how to understand the use of an RKHS norm for the purpose of regularization. For a given learning application it makes sense to select the norm so that predictors that we might expect to naturally arise would be the ones with smaller norm. It is similar to the selection of a prior probability distribution in the Bayesian estimation framework.

Let X denote a closed subset of $\mathbb{R}^d$ for some $d \ge 1$. Suppose we wish to consider classifiers of the form $g(x) = \operatorname{sgn}(\sum_i c_i \psi_i(x))$ for some set of functions $\psi_1, \ldots, \psi_n$ on X. For a given learning problem these functions would typically represent features that we expect to be useful for classification. One way to view the kernel method is as a way to introduce an inner product, and hence a norm, for linear combinations of the ψ's. See the final two paragraphs of this section.

To explore this view we need to understand how the choice of kernel affects the norms of functions. This is addressed by the following two propositions and some example kernels. For simplicity we begin by considering the case of finitely many basis functions (i.e. a finite dictionary of features). At the end of the section we return to the application to regularization.

Proposition 4.3. *Let $\psi_1, \ldots, \psi_n$ be linearly independent continuous functions on a set* X, *assumed to be a closed subset of a finite dimensional Euclidean space. Let K be the Mercer kernel defined by $K(x, x') = \sum_{i=1}^n \psi_i(x)\psi_i(x')$. Then $\psi_1, \ldots, \psi_n$ is a complete orthonormal basis for the RKHS $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$. That is, $\mathcal{H}_K$ is the set of linear combinations of $\psi_1, \ldots, \psi_n$, and $\langle \psi_i, \psi_j \rangle_K = \mathbf{1}_{\{i=j\}}$.*

Proof. Let $\mathcal{H}_\psi$ be the linear span of $\psi_1, \ldots, \psi_n$, which is a complete vector space. Each $f \in \mathcal{H}_\psi$ has a representation of the form $f = \sum_{i=1}^n c_i \psi_i$, and the representation is unique because of the assumed linear independence of $\psi_1, \ldots, \psi_n$. We can thus define an inner product $\langle \cdot, \cdot \rangle_\psi$ on $\mathcal{H}_\psi$ by

$$\left\langle \sum_{i=1}^n c_i \psi_i, \sum_{j=1}^n c'_j \psi_j \right\rangle_\psi = \sum_{i=1}^n c_i c'_i.$$

Then $(\mathcal{H}_\psi, \langle \cdot, \cdot \rangle_\psi)$ is a finite dimensional Hilbert space and $\psi_1, \ldots, \psi_n$ forms a complete orthonormal basis for it. To complete the proof we show that this space equals $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$.

The assumed form of K implies that for any fixed $x \in \mathsf{X}$, $K_x \in \mathcal{H}_\psi$, so as vector spaces, $\mathcal{H}_K \subset \mathcal{H}_\psi$. Also,

$$\begin{aligned}
\langle K_x, K_{x'} \rangle_\psi &= \left\langle \sum_{i=1}^n \psi_i(x)\psi_i(\cdot), \sum_{j=1}^n \psi_j(x')\psi_j(\cdot) \right\rangle_\psi \\
&= \sum_{i=1}^n \sum_{j=1}^n \psi_i(x_i)\psi_j(x'_j)\mathbf{1}_{\{i=j\}} \\
&= \sum_{i=1}^n \psi_i(x_i)\psi_i(x'_i) = K(x, x') \\
&= \langle K_x, K_{x'} \rangle_K
\end{aligned}$$

So $\langle \cdot, \cdot \rangle_\psi$ restricted to the subspace $\mathcal{H}_K$ agrees with $\langle \cdot, \cdot \rangle_K$. That is, the Hilbert space $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$ is a subspace of $(\mathcal{H}_\varphi, \langle \cdot, \cdot \rangle_\varphi)$. To show that the spaces are equal it suffices to show that the orthogonal complement of $\mathcal{H}_K$ within $\mathcal{H}_\psi$, $(\mathcal{H}_K)^\perp$, contains only the zero vector. Let $g \in (\mathcal{H}_K)^\perp$. Since $g \in \mathcal{H}_\psi$ there is a vector c so that $g = \sum_{j=1}^n c_j \psi_j$, and since $g \perp \mathcal{H}_K$,

$$0 = \langle K_x, g \rangle_\psi = \left\langle \sum_{i=1}^n \psi_i(x)\psi_i, \sum_{j=1}^n c_j \psi_j \right\rangle_\psi = \sum_i c_i \psi_i(x)$$

for all x. By the assumed linear independence of $\psi_1, \ldots, \psi_n$, it follows that $c = 0$ so that $g = 0$, as claimed. $\square$

The following proposition generalizes Proposition 4.3 to the case of a countably infinite basis. The assumption of linear independence is replaced by a stronger assumption. Let

$\ell^2 = \{c \in \mathbb{R}^{\mathbb{N}} : \sum_{i=1}^{\infty} c_i^2 < \infty\}$. With the inner product $\langle c, c' \rangle = \sum_{i=1}^{\infty} c_i c_i'$, ℓ^2 is a Hilbert space.

PROPOSITION 4.4. *Suppose K is a continuous kernel on $\mathsf{X} \times \mathsf{X}$ that has a representation of the form*

$$K(x, x') = \sum_{i=1}^{\infty} \psi_i(x) \psi_i(x'), \tag{4.15}$$

such that the functions ψ_i are continuous functions on X, and the following condition holds:

$$c \in \ell^2 \text{ and } \sum_{i=1}^{\infty} c_i \psi_i \equiv 0 \implies c = 0. \tag{4.16}$$

Then $\psi_1, \psi_2, \ldots$ forms a complete orthonormal basis for the RKHS $\mathcal{H}_K$.

PROOF. Let $\mathcal{H}_\psi = \left\{\sum_{i=1}^{\infty} c_i \psi_i : c \in \ell^2\right\}$. By the assumption (4.16), each $f \in \mathcal{H}_\psi$ is represented by a unique $c \in \ell^2$. We can thus define an inner product $\langle \cdot, \cdot \rangle_\psi$ on $\mathcal{H}_\psi$ by

$$\left\langle \sum_{i=1}^{\infty} c_i \psi_i, \sum_{j=1}^{\infty} c_j' \psi_j \right\rangle_\psi = \sum_{i=1}^{n} c_i c_i'.$$

Moreover, the mapping $c \to \sum_i c_i \psi_i$ from ℓ^2 to $\mathcal{H}_\psi$ is a Hilbert space isomorphism. The rest of the proof is the same as the proof of Proposition 4.3. $\square$

REMARK 4.5. *If K and the ψ_i's are continuous and satisfy* (4.15)*, then the convergence in* (4.15) *is absolute and uniform on compact subsets of X. (Details in problem set 4, 2017).*

Some examples popular in machine learning are given in the previous section. Additional examples, along with their expansions, are given here.

EXAMPLE 4.1. *(Bilinear kernels) Suppose $K(x, x') = 1 + \langle x, x' \rangle$ for $x, x' \in \mathbb{R}^d$. Then $K(x, x') = \sum_{i=0}^{n} \psi_i(x) \psi_i(x')$ with $\psi_0 \equiv 1$ and $\psi_i(x) = x_i$ for $1 \le i \le d$. Thus $\mathcal{H}_K$ consists of functions of the form $f_{b,w}(x) = b + \langle w, x \rangle$ for $b \in \mathbb{R}$ and $w \in \mathbb{R}^d$, and $\|f_{b,w}\|_K^2 = b^2 + \|w\|^2$.*

EXAMPLE 4.2. *(Polynomial kernels) Suppose $K(x, x') = (1 + \langle x, x' \rangle)^k$ for $x, x' \in \mathbb{R}^d$. Then*

$$\begin{aligned} K(x, x') &= (1 + x_1 x_1' + x_2 x_2' + \cdots + x_d x_d')^k \\ &= \sum_{(j_0, \ldots, j_d) \in \mathbb{Z}_+^{d+1} : j_0 + \ldots + j_d = k} \binom{k}{j_0 \; j_1 \; \cdots \; j_d} \prod_{i=1}^{d} x_i^{j_i} (x_i')^{j_i}, \end{aligned}$$

where $\binom{k}{j_0 \; j_1 \; \cdots \; j_d}$ is the multinomial coefficient. Thus, $\mathcal{H}_K$ has a complete orthonormal basis consisting of all functions of the form

$$\sqrt{\binom{k}{j_0 \; j_1 \; \cdots \; j_d}} \prod_{i=1}^{d} x_i^{j_i}$$

for $(j_0, \ldots, j_d) \in \mathbb{Z}_+^{d+1} : j_0 + \ldots + j_d = k$. *The functions in* $\mathcal{H}_K$ *consist of the multivariate polynomials on* $\mathbb{R}^d$ *with degree less than or equal to* k.

EXAMPLE 4.3. *(Gaussian kernel in one dimension) Suppose* $K(x, x') = \exp\left(-\frac{(x-x')^2}{2}\right)$ *for* $x, x' \in \mathbb{R}$. *Then*

$$\begin{aligned} K(x, x') &= \exp\left(-\frac{x^2 + (x')^2}{2}\right) \exp(xx') \\ &= \exp\left(-\frac{x^2 + (x')^2}{2}\right) \sum_{i=0}^{\infty} \frac{(xx')^i}{i!} \\ &= \sum_{i=1}^{\infty} \psi_i(x)\psi_i(x') \end{aligned}$$

where $\psi_i(x) = \exp\left(-\frac{x^2}{2}\right) \frac{x^i}{\sqrt{i!}}$.

To check condition (4.16), *suppose* $c \in \ell$ *is such that* $\sum_{i=1}^{\infty} c_i \psi_i(x) \equiv 0$, *or equivalently,* $h(x) \triangleq \sum_{i=1}^{\infty} c_i \frac{x^i}{\sqrt{i!}} \equiv 0$. *The series defining* h *and its derivatives of all orders are absolutely convergent on compact subsets, so by the dominated convergence theorem* h *can be repeatedly differentiated term by term. Thus, the* k^{th} *derivative of* h *evaluated at* $x = 0$ *is given by:*

$$D^k h(0) = \sum_{i=1}^{\infty} c_i D^k \left(\frac{x^i}{\sqrt{i!}}\right)\bigg|_{x=0} = c_k \sqrt{k!}$$

The assumption $h \equiv 0$ *therefore implies that* $c_k = 0$ *for all* k, *or* $c = 0$. *Thus, condition* (4.16) *holds. Therefore, by Proposition 4.4,* $\psi_1, \psi_2, \ldots$ *forms a complete orthonormal basis for the RKHS* $\mathcal{H}_K$.

EXAMPLE 4.4. *(Radial basis functions) Suppose* $K(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2}\right)$ *for* $x, x' \in \mathbb{R}^d$. *The functions of the form* $K_x(\cdot) = \exp\left(-\frac{\|x-\ \cdot\ \|^2}{2}\right)$ *are called* radial basis functions. *The radial basis functions each have unit norm and form a complete basis for* $\mathcal{H}_K$, *although they are not orthonormal. A complete orthonormal basis for* $\mathcal{H}_K$ *can be found by deriving a series representation for* K:

$$\begin{aligned} K(x, x') &= \exp\left(-\frac{\|x\|^2 + \|x'\|^2}{2}\right) \exp(\langle x, x' \rangle) \\ &= \exp\left(-\frac{\|x\|^2 + \|x'\|^2}{2}\right) \sum_{k=0}^{\infty} \frac{\langle x, x' \rangle^k}{k!} \end{aligned}$$

Combining analysis from the previous two examples, we find that $\mathcal{H}_K$ has a complete orthonormal basis consisting of functions of the form:

$$\exp\left(-\frac{\|x\|^2}{2}\right)\sqrt{\frac{1}{j_1!\cdots j_d!}}\prod_{i=1}^{d} x_i^{j_i}$$

for $d \geq 0$ and $(j_1,\ldots,j_d) \in \mathbb{Z}_+^d$. The functions in $\mathcal{H}_K$ thus include all the multivariate polynomials on $\mathbb{R}^d$.

The following proposition shows that Mercer kernels can be represented by series expansions in great generality.

PROPOSITION 4.5. *(Mercer's representation theorem, measure free version) Suppose $K(x,y)$ is a Mercer kernel on $\mathsf{X}\times\mathsf{X}$, where X is a closed subset of $\mathbb{R}^d$ (or more generally, X could be any complete separable metric space). Then there is a sequence of continuous functions (ψ_i) on X such that (4.15) and (4.16) hold, and $\psi_1,\psi_2,\ldots$ forms a complete orthonormal basis for the RKHS $\mathcal{H}_K$.*

PROOF. (Not self-contained. Reading is optional.) Mercer's theorem is usually stated and proved based on the integral operator associated with K and a reference measure ν on the Borel subsets of X with support X. The representations have the form $K(x,y) = \sum_{i=1}^{\infty}\lambda_i\psi_i(x)\psi_i(y)$, but the nonnegative constants λ_i can be incorporated into the functions ψ_i, because we are not concerned with orthonormality of the functions relative to the reference measure. Since X is assumed to be separable there exists a countable dense subset $\{q_1,q_2,\ldots\}$ of X. Given K, by putting a sufficiently small positive probability mass on each of these points we can construct a discrete probability measure ν which is strictly positive on X (i.e. every open nonempty set has strictly positive measure) such that $\int_{\mathsf{X}\times\mathsf{X}} K(x,y)^2\nu(dx)\nu(dy) < \infty$, from which it follows that $K \in \mathcal{A}(\mathsf{X},\nu)$, where $\mathcal{A}(\mathsf{X},\nu)$ is defined in [**FM09**]. Therefore we can apply [**FM09**, Theorem 2.4] to conclude the existence of a sequence of continuous functions (ψ_i) on X such that (4.15) is satisfied. The sequence satisfies (4.16) because the functions in the sequence are orthogonal in $L^2(\mathsf{X},\nu)$. The last statement of the proposition follows from Proposition 4.4. $\square$

EXAMPLE 4.5. *(Cauchy kernel function) The Cauchy kernel function is given by $K(x,x') = \frac{1}{1+\gamma\|x-x'\|^2}$ for $x,x' \in \mathbb{R}^d$, where γ is a scale parameter. The Cauchy kernel is similar to the Gaussian kernel, but it decreases much more slowly as $\|x-x'\| \to \infty$. The positive semidefinite property follows from the fact that the Fourier transform of $\frac{1}{1+\gamma\|x\|^2}$ is real and positive. Although Proposition 4.5 applies for this kernel, the proof of the proposition is nonconstructive, and there does not seem to be a natural way to find a complete orthonormal basis as there is for the case of Gaussian kernel function. Still, in applications, the Cauchy kernel can be used in much the same way as the Gaussian kernel, for example in (8.52) or (8.53).*

Kernels as measures of complexity of functions. Section 8.6 considers classifiers of the form $\mathrm{sgn}(f)$ and Chapter 9 considers regression functions of the form f, where f is assumed to be in some class of functions, and the ERM algorithm selects $\widehat{f}$ as the solution to an optimization problem. To avoid overfitting, the complexity of f is either constrained, or a penalty involving the complexity of f is added to the objective function. For many applications it is reasonable to optimize over functions f having the form $f(x) = \sum_j a_j \widetilde{\psi}_j(x)$, such that a finite or infinite set of functions $(\widetilde{\psi}_j)$ has been predetermined to fit the application well. In order to express prior knowledge about which coefficients are likely to be larger in the application, a sequence of positive weights $(w_i)_{i\geq 1}$ could be selected, such that the complexity of a function $f(x) = \sum_j a_j \widetilde{\psi}_j(x)$ could be defined to be $\sum_j a_j^2/w_j$. A larger weight w_j for a particular j, yields a smaller complexity cost for having a large value of a_j.

If the weights (w_j) are chosen such that K defined by $K(x,x') = \sum_j w_j \widetilde{\psi}_j(x)\widetilde{\psi}_j(x')$ is a Mercer kernel (roughly speaking, this requires that most of the weights are not too large) then we can also write $K(x,x') = \sum_i \psi_i(x)\psi_i(x')$, where $\psi_i = \sqrt{w_i}\widetilde{\psi}_i$. Therefore, since $f(x) = \sum_j \frac{a_j}{\sqrt{w_j}}\psi_j(x)$, it follows from Proposition 4.3 or 4.4 that $\|f\|_K^2 = \sum_j a_j^2/w_j$. In other words, the complexity measure is the squared RKHS norm of f for the kernel K. In this sense, the kernel K defines a measure of complexity of f.

Part 2

Basic Theory

CHAPTER 5

Formulation of the learning problem

Now that we have seen an informal statement of the learning problem, as well as acquired some technical tools in the form of concentration inequalities, we can proceed to define the learning problem formally. Recall that the basic goal is to be able to predict some random variable Y of interest from a correlated random observation X, where the predictor is to be constructed on the basis of n i.i.d. training samples $(X_1, Y_1), \dots, (X_n, Y_n)$ from the joint distribution of (X, Y). We will start by looking at an idealized scenario (often called the *realizable case* in the literature), in which Y is a *deterministic* function of X, and we happen to know the function class to which it belongs. This simple set-up will let us pose, in a clean form, the basic requirements a learning algorithm should satisfy. Once we are done with the realizable case, we can move on to the general setting, in which the relationship between X and Y is probabilistic and not known precisely. This is often referred to as the *model-free* or *agnostic* case.

This order of presentation is, essentially, historical. The first statement of the learning problem is hard to trace precisely, but the "modern" algorithmic formalization seems to originate with the 1984 work of Valiant [**Val84**] on learning Boolean formulae. Valiant has focused on *computationally efficient* learning algorithms. The agnostic (or model-free) formulation was first proposed and studied by Haussler [**Hau92**] in 1992.

The material in this chapter closely follows the exposition of Vidyasagar [**Vid03**, Ch. 3].

5.1. The realizable case

We start with an idealized scenario, now often referred to in the literature as the *realizable case*. The basic set-up is as follows. We have a set X (often called the *feature space* or *input space*) and a family $\mathcal{P}$ of probability distributions on X. We obtain an i.i.d. sample $X^n = (X_1, \dots, X_n)$ drawn according to some $P \in \mathcal{P}$, which we do not know (although it may very well be the case that $\mathcal{P}$ is a singleton, $|\mathcal{P}| = 1$, in which case we, of course, *do* know P). We will look at two basic problems:

(1) *Concept learning:* There is a class $\mathcal{C}$ of subsets of X, called the *concept class*, and an unknown *target concept* $C^* \in \mathcal{C}$ is picked by Nature. For each feature X_i in our sample X^n, we receive a binary *label* $Y_i = \mathbf{1}_{\{X_i \in C^*\}}$. The n feature-label pairs form the *training set*

$$(X_1, Y_1) = (X_1, \mathbf{1}_{\{X_1 \in C^*\}}), \dots, (X_n, Y_n) = (X_n, \mathbf{1}_{\{X_n \in C^*\}}). \tag{5.1}$$

The objective is to approximate the target concept C^* as accurately as possible.

(2) *Function learning:* There is a class $\mathcal{F}$ of functions $f : \mathsf{X} \to [0, 1]$, and an unknown *target function* $f^* \in \mathcal{F}$ is picked by nature. For each input point X_i in the sample

X^n, we receive a real-valued *output* $Y_i = f^*(X_i)$. The n input-output pairs

$$(X_1, Y_1) = (X_1, f^*(X_1)), \dots, (X_n, f^*(X_n)). \tag{5.2}$$

The objective is to approximate the target function f^* as accurately as possible. (**Note:** the requirement that f map X into $[0,1]$ is imposed primarily for technical convenience; using appropriate moment and/or tail behavior assumptions on P, it is possible to remove this requirement, but the resulting proofs will be somewhat laborious.)

We will now consider these two problems separately.

5.1.1. Concept learning. As we already stated, the goal of concept learning is to approximate the target concept C^* as accurately as possible on the basis of the training data (5.1). This is done by means of a *learning algorithm*. An algorithm of this sort should be capable of producing an approximation to C^* given the training set of the form (5.1) of any size n. More precisely:

Definition 5.1. *A* concept learning problem *is specified by a triple* $(\mathsf{X}, \mathcal{P}, \mathcal{C})$*, where* X *is the feature space,* $\mathcal{P}$ *is a family of probability distributions on* X*, and* $\mathcal{C}$ *is a concept class. A* learning algorithm *for* $(\mathsf{X}, \mathcal{P}, \mathcal{C})$ *is a sequence* $\mathcal{A} = \{A_n\}_{n=1}^\infty$ *of mappings*

$$A_n : (\mathsf{X} \times \{0,1\})^n \to \mathcal{C}.$$

If $\mathcal{P}$ consists of only one distribution P, then the mappings A_n may depend on P; otherwise, they may only depend on $\mathcal{P}$ as a whole. The idea behind the above definition is that for each training set size n we have a definite procedure for forming an approximation to the unknown target concept C^* on the basis of the training set of that size.

For brevity, let us denote by Z_i the ith training pair $(X_i, Y_i) = (X_i, \mathbf{1}_{\{X_i \in C^*\}})$, and let us denote by Z the set $\mathsf{X} \times \{0,1\}$. Given a training set $Z^n = (Z_1, \dots, Z_n) \in \mathsf{Z}^n$ and a learning algorithm $\mathcal{A}$, the approximation to C^* is

$$\widehat{C}_n = A_n(Z^n) = A_n(Z_1, \dots, Z_n) = A_n\left((X_1, \mathbf{1}_{\{X_1 \in C^*\}}), \dots, (X_n, \mathbf{1}_{\{X_n \in C^*\}})\right).$$

Note that $\widehat{C}_n$ is an element of the concept class $\mathcal{C}$ (by definition), and that it is a random variable since it depends on the random sample Z^n. It is often referred to as a *hypothesis* output by the learning algorithm $\mathcal{A}$.

How shall we measure the goodness of this approximation $\widehat{C}_n$? A natural thing to do is the following. Suppose now we draw a fresh feature X from the same distribution $P \in \mathcal{P}$ as the one that has generated the training feature set X^n and venture a *hypothesis* that X belongs to the target concept C^* if $X \in \widehat{C}_n$, i.e., if $\mathbf{1}_{\{X \in \widehat{C}_n\}} = 1$. When would we make a mistake, i.e., *misclassify* X? There are two mutually exclusive cases:

(1) X is in C^*, but not in $\widehat{C}_n$, i.e., $X \in C^* \cap \widehat{C}_n^c$, where $\widehat{C}_n^c = \mathsf{X} \backslash \widehat{C}_n$ is the complement of $\widehat{C}_n$ in X.
(2) X is not in C^*, but it is in $\widehat{C}_n$, i.e., $X \in (C^*)^c \cap \widehat{C}_n$.

Thus, we will misclassify X precisely when it happens to lie in the *symmetric difference*

$$C^* \triangle \widehat{C}_n := (C^* \cap \widehat{C}_n^c) \cup ((C^*)^c \cap \widehat{C}_n).$$

This will happen with probability $P(C^* \triangle \widehat{C}_n)$ — note, by the way, that this is a random number since $\widehat{C}_n$ depends on the training data Z^n. At any rate, we take the P-probability of the symmetric difference $C^* \triangle \widehat{C}_n$ as our measure of performance of $\mathcal{A}$. In order to streamline the notation, let us define the *risk* (or *loss*) of any $C \in \mathcal{C}$ w.r.t. C^* and P as

$$L_P(C, C^*) := P(C \triangle C^*) = P(X \in C \triangle C^*).$$

EXERCISE 5.1. *Prove that*

$$L_P(C, C^*) = \int_{\mathsf{X}} \left| \mathbf{1}_{\{x \in C\}} - \mathbf{1}_{\{x \in C^*\}} \right|^2 P(\mathrm{d}x).$$

In other words, $L_P(C, C^*)$ is the squared $L^2(P)$ norm of the difference of the indicator functions $I_C(\cdot) = \mathbf{1}_{\{\cdot \in C\}}$ and $I_{C^*}(\cdot) = \mathbf{1}_{\{\cdot \in C^*\}}$, $L_P(C, C^*) = \|I_C - I_{C^*}\|^2_{L^2(P)}$.

Roughly speaking, we will say that $\mathcal{A}$ is a good algorithm if

$$L_P(\widehat{C}_n, C^*) \to 0 \qquad \text{as } n \to \infty \tag{5.3}$$

for any $P \in \mathcal{P}$ and any $C^* \in \mathcal{C}$. Since $\widehat{C}_n$ is a random element of $\mathcal{C}$, the convergence in (5.3) can only be in some probabilistic sense. In order to make things precise, for any $C \in \mathcal{C}$ let us denote by P_C the joint distribution of a pair $Z = (X, Y)$, where $X \sim P$ and $Y = \mathbf{1}_{\{X \in C\}}$. Then we define the following two quantities:

$$r_{\mathcal{A}}(n, \varepsilon, P) := \sup_{C \in \mathcal{C}} P^n_C \left(Z^n \in \mathsf{Z}^n : L_P(A_n(Z^n), C) > \varepsilon \right)$$
$$\bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) := \sup_{P \in \mathcal{P}} r_{\mathcal{A}}(n, \varepsilon, P)$$

where P^n_C denotes the n-fold product of P. For a fixed P (which amounts to assuming that the features X^n were drawn i.i.d. from P), $r_{\mathcal{A}}(n, \varepsilon, P)$ quantifies the worst-case "size" of the set of "bad" samples, where we say that a sample X^n is bad if it causes the learning algorithm $\mathcal{A}$ to output a hypothesis $\widehat{C}_n = A_n(Z^n)$ whose risk is larger than ε. The worst case is over the entire concept class $\mathcal{C}$, since we do not know the target concept C^*. The quantity $\bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P})$ accounts for the fact that we do not know which $P \in \mathcal{P}$ has generated the training feature points.

With all these things defined, we can now state the following:

DEFINITION 5.2. *(PAC for concept learning in realizable learning setting) A learning algorithm $\mathcal{A} = \{A_n\}$ is* probably approximately correct *(or* PAC*) to accuracy ε if*

$$\lim_{n \to \infty} \bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) = 0. \tag{5.4}$$

We say that $\mathcal{A}$ is PAC if it is PAC to accurary ε for every $\varepsilon > 0$. The concept class $\mathcal{C}$ is called PAC learnable to accuracy ε w.r.t. $\mathcal{P}$ *if there exists an algorithm that is PAC to accuracy ε. Finally, we say that $\mathcal{C}$ is* PAC learnable *if there exists an algorithm that is PAC.*

The term "probably approximately correct," which seems to have first been introduced by Angluin [**Ang88**], is motivated by the following observations. First, the hypothesis $\widehat{C}_n$ output by $\mathcal{A}$ for some n is only an *approximation* to the target concept C^*. Thus, $L_P(\widehat{C}_n, C^*)$ will be, in general, nonzero. But if it is small, then we are justified in claiming that $\widehat{C}_n$ is *approximately correct*. Secondly, we may always encounter a bad sample, so $L_P(\widehat{C}_n, C^*)$ can be made small only *with high probability*. Thus, informally speaking, a PAC algorithm is one that "works reasonably well most of the time."

An equivalent way of phrasing the statement that a learning algorithm is PAC is as follows: For any $\varepsilon > 0$ and $\delta > 0$, there exists some $n(\varepsilon, \delta) \in \mathbb{N}$, such that

$$P_C^n\left(Z^n \in \mathsf{Z}^n : L_P(A_n(Z^n), C) > \varepsilon\right) \le \delta, \qquad \forall n \ge n(\varepsilon,\delta), \forall C \in \mathcal{C}, \forall P \in \mathcal{P}. \tag{5.5}$$

In this context, ε is called the *accuracy parameter*, while δ is called the *confidence parameter*. The meaning of this alternative characterization is as follows. If the sample size n is at least $n(\varepsilon, \delta)$, then we can state with confidence at least $1 - \delta$ that the hypothesis $\widehat{C}_n$ will correctly classify a fresh random point $X \in \mathsf{X}$ with probability at least $1 - \varepsilon$.

The two problems of interest to us are:

(1) Determine conditions under which a given concept class $\mathcal{C}$ is PAC learnable.
(2) Obtain upper and lower bounds on $n(\varepsilon, \delta)$ as a function of ε, δ. The following terminology is often used: the smallest number $n(\varepsilon, \delta)$ such that (5.5) holds is called the *sample complexity*.

5.1.2. Function learning. The goal of function learning is to construct an accurate approximation to an unknown target function $f^* \in \mathcal{F}$ on the basis of training data of the form (5.2). Analogously to the concept learning scenario, we have:

DEFINITION 5.3. *A* function learning problem *is specified by a triple* $(\mathsf{X}, \mathcal{P}, \mathcal{F})$*, where* X *is the input space,* $\mathcal{P}$ *is a family of probability distributions on* X*, and* $\mathcal{F}$ *is a class of functions* $f : \mathsf{X} \to [0, 1]$*. A* learning algorithm *for* $(\mathsf{X}, \mathcal{P}, \mathcal{F})$ *is a sequence* $\mathcal{A} = \{A_n\}_{n=1}^{\infty}$ *of mappings*

$$A_n : (\mathsf{X} \times [0, 1])^n \to \mathcal{F}.$$

As before, let us denote by Z_i the input-output pair $(X_i, Y_i) = (X_i, f^*(X_i))$ and by Z the product set $\mathsf{X} \times [0, 1]$. Given a training set $Z^n = (Z_1, \dots, Z_n) \in \mathsf{Z}^n$ and a learning algorithm $\mathcal{A}$, the approximation to f^* is

$$\widehat{f}_n = A_n(Z^n) = A_n\left((X_1, f^*(X_1)), \dots, (X_n, f^*(X_n))\right).$$

As in the concept learning setting, $\widehat{f}_n$ is a *random element* of the function class $\mathcal{F}$.

In order to measure the performance of $\mathcal{A}$, we again imagine drawing a fresh input point $X \in \mathsf{X}$ from the same distribution $P \in \mathcal{P}$ that has generated the training inputs X^n. A natural error metric is the squared loss $|\widehat{f}_n(X) - f^*(X)|^2$. As before, we can define the *risk* (or *loss*) of any $f \in \mathcal{F}$ w.r.t. f^* and P as

$$L_P(f, f^*) := \mathbf{E}_P|f(X) - f^*(X)|^2 = \|f - f^*\|^2_{L^2(P)} = \int_{\mathsf{X}} |f(x) - f^*(x)|^2 P(\mathrm{d}x). \tag{5.6}$$

Thus, the quantity of interest is the risk of $\widehat{f}_n$:

$$L_P(\widehat{f}_n, f^*) = \int_{\mathsf{X}} |\widehat{f}_n(x) - f^*(x)|^2 P(\mathrm{d}x).$$

Keep in mind that $L_P(\widehat{f}_n, f^*)$ is a random variable, as it depends on $\widehat{f}_n$, which in turn depends on the random sample $Z^n \in \mathsf{Z}^n$.

REMARK 5.1. The concept learning problem is, in fact, a special case of the function learning problem. Indeed, fix a concept class $\mathcal{C}$ and consider the function class $\mathcal{F}$ consisting of the indicator functions of the sets in $\mathcal{C}$:

$$\mathcal{F} = \{I_C : C \in \mathcal{C}\}.$$

Then for any $f = I_C$ and $f^* = I_{C^*}$ we will have

$$L_P(f, f^*) = \|I_C - I_{C^*}\|^2_{L^2(P)} = P(C \triangle C^*),$$

which is the error metric we have defined for concept learning.

If for each $f \in \mathcal{F}$ we denote by P_f the joint distribution of a pair $Z = (X, Y)$, where $X \sim P$ and $Y = f(X)$, then for a given learning problem $(\mathsf{X}, \mathcal{P}, \mathcal{F})$ and a given algorithm $\mathcal{A}$ we can define

$$\begin{aligned} r_{\mathcal{A}}(n, \varepsilon, P) &:= \sup_{f \in \mathcal{F}} P^n_f \left(Z^n \in \mathsf{Z}^n : L_P(A_n(Z^n), f) > \varepsilon \right) \\ \bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) &:= \sup_{P \in \mathcal{P}} r_{\mathcal{A}}(n, \varepsilon, P) \end{aligned}$$

for every $n \in \mathbb{N}$ and $\varepsilon > 0$.The meaning of these quantities is exactly parallel to the corresponding quantities in concept learning, and leads to the following definition:

Definition 5.4. *(PAC for function learning in realizable learning setting) A learning algorithm $\mathcal{A} = \{A_n\}$ is* PAC to accuracy ε *if*

$$\lim_{n \to \infty} \bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) = 0,$$

and PAC if it is PAC to accuracy ε for all $\varepsilon > 0$. A function class $\mathcal{F} = \{f : \mathsf{X} \to [0, 1]\}$ is PAC-learnable (to accuracy ε) *w.r.t. $\mathcal{P}$ if there exists an algorithm $\mathcal{A}$ that is PAC for $(\mathsf{X}, \mathcal{P}, \mathcal{F})$ (to accuracy ε).*

An equivalent way of stating that $\mathcal{A}$ is PAC is that, for any $\varepsilon, \delta > 0$ there exists some $n(\varepsilon, \delta) \in \mathbb{N}$ such that

$$P^n \left(Z^n \in \mathsf{Z}^n : L_P(A_n(Z^n), f) > \varepsilon \right) \le \delta, \qquad \forall n \ge n(\varepsilon, \delta), \forall f \in \mathcal{F}, \forall P \in \mathcal{P}.$$

The smallest $n(\varepsilon, \delta) \in \mathbb{N}$ for which the above inequality holds is termed the *sample complexity*.

5.2. Examples of PAC-learnable concept classes

To make these ideas concrete, let us consider two examples of PAC-learnable concept classes. Given what we know at this point, the only way to show that a given concept class is PAC-learnable is to exhibit an algorithm which is PAC. Later on, we will develop generic tools that will allow us to determine PAC-learnability without having to construct an algorithm for each separate case.

5.2.1. Finite concept classes. First, we show that any finite concept class is PAC-learnable. Thus, consider a triple $(\mathsf{X}, \mathcal{P}, \mathcal{C})$, where $\mathcal{X}$ and $\mathcal{P}$ are arbitrary, but the concept class is finite: $|\mathcal{C}| < \infty$.

Let $Z^n = (Z_1, \ldots, Z_n)$ be a training set, where, as usual, $Z_i = (X_i, Y_i)$ with $Y_i = \mathbf{1}_{\{X_i \in C^*\}}$. We say that the ith training pair Z_i is a *positive example* if $Y_i = 1$ (i.e., if $X_i \in C^*$), and is a *negative example* otherwise. By hypothesis, $\mathcal{C} = \{C_m\}_{m=1}^M$. Consider the (random) set

$$\mathcal{F}(Z^n) := \left\{ m \in [M] : Y_i = \mathbf{1}_{\{X_i \in C_m\}} \text{ for all } i \in [n] \right\}.$$

In other words, $\mathcal{F}(Z^n)$ consists of all concepts in $\mathcal{C}$ that are consistent with the training data. If $C^* = C_{m^*}$, then evidently $m^* \in \mathcal{F}(Z^n)$, so $\mathcal{F}(Z^n)$ is nonempty. We consider an algorithm that returns an arbitrary element of $\mathcal{F}(Z^n)$, say, the smallest one:

$$\widehat{C}_n = A_n(Z^n) = C_{\widehat{m}_n}, \qquad \text{where } \widehat{m}_n = \min\{m : m \in \mathcal{F}(Z^n)\}. \tag{5.7}$$

THEOREM 5.1. *The algorithm $\mathcal{A}$ defined in (5.7) satisfies*

$$\overline{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) \le M(1-\varepsilon)^n. \tag{5.8}$$

Therefore, this algorithm is PAC, so the class $\mathcal{C}$ is PAC-learnable.

PROOF. Fix some $P \in \mathcal{P}$ and $\varepsilon > 0$ and consider the set

$$\mathcal{B}_{m^*,P}(\varepsilon) := \{m \in [M] : P(C_m \triangle C_{m^*}) > \varepsilon\}.$$

This set is deterministic, and evidently

$$\begin{aligned} P^n\left[P(\widehat{C}_n \triangle C^*) > \varepsilon\right] &= P^n\left[\widehat{m}_n \in \mathcal{B}_{m^*,P}\right] \\ &= \sum_{m \in \mathcal{B}_{m^*,P}} P^n[\widehat{m}_n = m]. \end{aligned}$$

For any $m \in \mathcal{B}_{m^*,P}$, the event $\{\widehat{m}_n = m\}$ happens only if $m \in \mathcal{F}(Z^n) \cap \mathcal{B}_{m^*,P}$. Since $C^* = C_{m^*}$ fits the training data perfectly [and, in particular, $m^* \in \mathcal{F}(Z^n)$], C_m will fit the training data perfectly if and only if none of the X_i's fall into $C_{m^*} \triangle C_m$. If $m \in \mathcal{F}(Z^n) \cap \mathcal{B}_{m^*,P}$, then it must also be the case that $P(C_m \triangle C_{m^*}) > \varepsilon$. Thus,

$$\begin{aligned} P^n[\widehat{m}_n = m] &\le P^n\left[X_i \notin C_m \triangle C_{m^*} \text{ for all } i \in [n]\right] \\ &\le (1-\varepsilon)^n, \end{aligned}$$

where the second inequality uses the fact that the X_i's are independent and the fact that, for all $m \in \mathcal{B}_{m^*,P}$, $P(C_m \triangle C_{m^*}) > \varepsilon$. Since $|\mathcal{B}_{m^*,P}| \le M$, we obtain the bound

$$P^n\left[P(\widehat{C}_n \triangle C^*) > \varepsilon\right] \le M(1-\varepsilon)^n.$$

Since this bound holds for all $P \in \mathcal{P}$ and for all choices of C^*, we get (5.8). □

COROLLARY 5.1. *The sample complexity of learning a finite concept class of cardinality M satisfies*

$$n(\varepsilon, \delta) \ge \frac{1}{\varepsilon} \log\left(\frac{M}{\delta}\right).$$

PROOF. From (5.8), we see that $\overline{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) \le \delta$ for all n satisfying $M(1-\varepsilon)^n \le \delta$. Since $(1-\varepsilon)^n \le e^{-n\varepsilon}$, a sufficient condition is $\delta \ge \varepsilon^{-1}\log(M/\delta)$. □

5.2.2. Axis-parallel rectangles. We take $\mathsf{X} = [0,1]^2$, the unit square in the plane, let $\mathcal{P}$ be the class of all probability distributions on X (w.r.t. the usual Borel σ-algebra), and let $\mathcal{C}$ be the collection of all *axis-parallel rectangles*: that is, a set C is in $\mathcal{C}$ if and only if it is of the form

$$\begin{aligned} C &= [a_1, b_1] \times [a_2, b_2] \\ &= \left\{(x_1, x_2) \in [0,1]^2 : a_1 \le x_1 \le b_1,\ a_2 \le x_2 \le b_2\right\} \end{aligned}$$

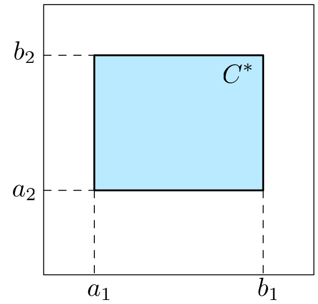

FIGURE 1. An axis-parallel rectangle.

for some $0 \le a_1 \le b_1 \le 1$ and $0 \le a_2 \le b_2 \le 1$ (see Figure 1).

We now describe our learning algorithm. Given a training set $Z^n = \big(Z_1, \ldots, Z_n\big) = \big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$, we partition the examples into positive and negative ones, as before. Our algorithm $\mathcal{A} = \{A_n\}_{n=1}^{\infty}$ is the following intuitive rule: for each n, we take

$$\widehat{C}_n = A_n(Z^n) = \text{smallest rectangle } C \in \mathcal{C} \text{ that contains all positive examples in } Z^n. \tag{5.9}$$

Figure 1 shows a particular instance of this algorithm. We will now prove the following result, originally due to Blumer et al. [**BEHW89**]:

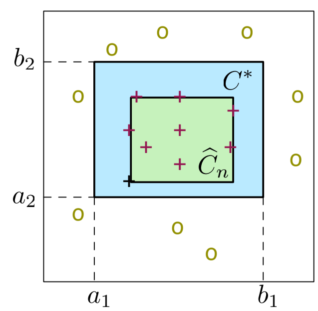

FIGURE 2. The hypothesis $\widehat{C}_n$ produced by algorithm (5.9). Positive (resp., negative) examples are shown as crosses (resp., zeros).

THEOREM 5.2. *The algorithm $\mathcal{A}$ defined in (5.9), i.e., the one that returns the smallest axis-parallel rectangle that encloses all positive examples in Z^n, satisfies*

$$\bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) \le 4(1 - \varepsilon/4)^n. \tag{5.10}$$

Therefore, this algorithm is PAC, and the class $\mathcal{C}$ is PAC-learnable.

PROOF. Since no positive example can lie outside C^*, the hypothesis $\widehat{C}_n$ produced by the algorithm (5.9) must lie inside C^*: $\widehat{C}_n \subseteq C^*$. Therefore,

$$\widehat{C}_n \triangle C^* = C^* \cap (\widehat{C}_n)^c \equiv C^* \backslash \widehat{C}_n. \tag{5.11}$$

If $P(C^*) < \varepsilon$, then from (5.11) it follows that $L_P(\widehat{C}_n, C^*) = P\big(C^* \backslash \widehat{C}_n\big) \le P(C^*) < \varepsilon$. Thus, we will assume that $P(C^*) \ge \varepsilon$. Suppose that $C^* = [a_1, b_1] \times [a_2, b_2]$ and $\widehat{C}_n = [\widehat{a}_1, \widehat{b}_1] \times [\widehat{a}_2, \widehat{b}_2]$, and consider the following four rectangles:

$$
\begin{aligned}
V_1 &= [a_1, \widehat{a}_1) \times [a_2, b_2], \\
V_2 &= (\widehat{b}_1, b_1] \times [a_2, b_2], \\
H_1 &= [a_1, b_1] \times [a_2, \widehat{a}_2), \\
H_2 &= [a_1, b_1] \times (\widehat{b}_2, b_2]
\end{aligned}
$$

(see Figure 3). From (5.11), we see that the symmetric difference $\widehat{C}_n \triangle C^*$ is exactly equal

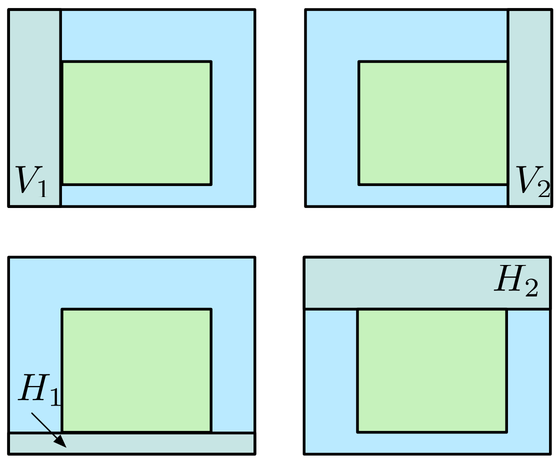

FIGURE 3. The four rectangles V_1, V_2, H_1, H_2 used in the proof of Theorem 5.2. The target concept C^* is in blue, the hypothesis $\widehat{C}_n$ returned by our algorithm is green.

to the union of these four rectangles, which we will denote by E:

$$\widehat{C}_n \triangle C^* = E := V_1 \cup V_2 \cup H_1 \cup H_2.$$

(Note, by the way, that V_1, V_2, H_1, H_2, E are all *random* rectangles, since they depend on the hypothesis $\widehat{C}_n$ and thus on the random training set Z^n.) Our goal is show that $P(E) \le \varepsilon$ with high probability.

We claim that, with probability at least $1 - 4(1-\varepsilon/4)^n$, each of the rectangles V_1, V_2, H_1, H_2 has P-probability of no more than $\varepsilon/4$. Assuming this is the case, then $P(E) = P(V_1 \cup V_2 \cup H_1 \cup H_2) \le \varepsilon$ with probability at least $1 - 4(1-\varepsilon/4)^n$. From this, we conclude that, for any C^*,

$$P^n_{C^*}\left(Z^n \in \mathsf{Z}^n : L_P(A_n(Z^n), C^*) > \varepsilon\right) \le P^n_{C^*}\Big(Z^n \in \mathsf{Z}^n : P(E) > \varepsilon\Big) \le 4(1-\varepsilon/4)^n. \tag{5.12}$$

Since this bound holds for all C^* and for all $P \in \mathcal{P}$, we get (5.10).

It now remains to prove the claim. We first focus on $P(V_1)$. To that end, let A_1 be the smallest rectangle of the form $[a_1, a] \times [a_2, b_2]$, such that $P(A_1) \geq \varepsilon/4$. Such a rectangle exists by the continuity of probability, and, moreover, $P([a_1, a) \times [a_2, b_2]) \leq \varepsilon/4$. Consider the event

$$\bigcup_{i=1}^{n} \{X_i \in A_1\}, \tag{5.13}$$

i.e., that A_1 is hit by at least one training example. If this event occurs, then $V_1 \subset [a_1, a) \times [a_2, b_2]$, so if this event occurs, $P(V_1) \leq \varepsilon/4$. In other words,

$$\bigcup_{i=1}^{n} \{X_i \in A_1\} \subseteq \{P(V_1) \leq \varepsilon/4\}.$$

or, taking the contrapositive,

$$\{P(V_1) > \varepsilon/4\} \subseteq \bigcap_{i=1}^{n} \{X_i \notin A_1\}.$$

The probability of the event that there are no training examples in A_1 can be computed as

$$\begin{aligned} \mathbf{P}\left(\bigcap_{i=1}^{n} \{X_i \notin A_1\}\right) &= \prod_{i=1}^{n} \mathbf{P}\left(X_i \notin A_1\right) && (5.14) \\ &= \left[\mathbf{P}\left(X \notin A_1\right)\right]^n && (5.15) \\ &\leq \left(1 - \varepsilon/4\right)^n, && (5.16) \end{aligned}$$

where (5.14) is by independence of the X_i's, (5.15) follows from the fact that the X_i's are identically distributed, and (5.16) follows from the fact that $P(A_1) \geq \varepsilon/4$ by construction. Therefore,

$$\mathbf{P}\Big[P(V_1) > \varepsilon/4\Big] \leq \mathbf{P}\Big[X_i \notin A_1, \forall i\Big] \leq (1 - \varepsilon/4)^n.$$

Similar reasoning applies to V_2, H_1, H_2. Thus, with probability at least $1 - 4(1 - \varepsilon/4)^n$,

$$P(V_1) \leq \varepsilon/4, \quad P(V_2) \leq \varepsilon/4, \quad P(H_1) \leq \varepsilon/4, \quad P(H_2) \leq \varepsilon/4,$$

and the claim is proved. □

Corollary 5.2. *The sample complexity of learning axis-parallel rectangles satisfies*

$$n(\varepsilon, \delta) \geq \frac{4 \log(4/\delta)}{\varepsilon}. \tag{5.17}$$

Proof. From (5.10), $\bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) \leq \delta$ for all n such that $4(1 - \varepsilon/4)^n \leq \delta$. Following the same reasoning as in the proof of Corollary 5.1, we get (5.17). □

5.3. Agnostic (or model-free) learning

The realizable setting we have focused on in Section 5.1 rests on certain assumptions, which are not always warranted:

- The assumption that the target concept C^* belongs to $\mathcal{C}$ (or that the target function f^* belongs to $\mathcal{F}$) means that we are trying to fit a hypothesis to data, which are *a priori* known to have been generated by some member of the model class defined by $\mathcal{C}$ (or by $\mathcal{F}$). However, in general we may not want to (or be able to) assume much about the data generation process, and instead would like to find the best fit to the data at hand using an element of some model class of our choice.
- The assumption that the training features (or inputs) are labelled noiselessly by $\mathbf{1}_{\{x \in C^*\}}$ (or by $f(x)$) rules out the possibility of noisy measurements or observations.
- Finally, even if the above assumption were true, we would not necessarily have *a priori* knowledge of the concept class $\mathcal{C}$ (or the function class $\mathcal{F}$) containing the target concept (or function). In that case, the best we could hope for is to pick our own model class and seek the best *approximation* to the unknown target concept (or function) among the elements of that class.

The *model-free learning problem* (also referred to as the *agnostic case*), introduced by Haussler [**Hau92**], takes a more general decision-theoretic approach and removes the above restrictions. It has the following ingredients:

- Sets X, Y, and U
- A class $\mathcal{P}$ of probability distributions on $\mathsf{Z} := \mathsf{X} \times \mathsf{Y}$
- A class $\mathcal{F}$ of functions $f : \mathsf{X} \to \mathsf{U}$ (the *hypothesis space*)
- A *loss function* $\ell : \mathsf{Y} \times \mathsf{U} \to [0, 1]$

The learning process takes place as follows. We obtain an i.i.d. sample $Z^n = (Z_1, \ldots, Z_n)$, where each $Z_i = (X_i, Y_i)$ is drawn from the same fixed but unknown $P \in \mathcal{P}$. A *learning algorithm* is a sequence $\mathcal{A} = \{A_n\}_{n=1}^{\infty}$ of mappings

$$A_n : \mathsf{Z}^n \to \mathcal{F}.$$

As before, let

$$\widehat{f}_n = A_n(Z^n) = A_n(Z_1, \ldots, Z_n) = A_n((X_1, Y_1), \ldots, (X_n, Y_n)).$$

This is the hypothesis emitted by the learning algorithm based on the *training data* Z^n. Note that, by definition, $\widehat{f}_n$ is a *random element* of the hypothesis space $\mathcal{F}$, and that it maps each point $x \in \mathsf{X}$ to a point $u = \widehat{f}_n(x) \in \mathsf{U}$. Following the same steps as in the realizable case, we evaluate the goodness of $\widehat{f}_n$ by its expected loss

$$L_P(\widehat{f}_n) := \mathbf{E}_P\big[\ell(Y, \widehat{f}_n(X))\big|Z^n\big] = \int_{\mathsf{X}\times\mathsf{Y}} \ell(y, \widehat{f}_n(x)) P(\mathrm{d}x, \mathrm{d}y),$$

where the expectation is w.r.t. a random couple $(X, Y) \in \mathsf{Z}$ drawn according to the same P but independently of Z^n. Note that $L_P(\widehat{f}_n)$ is a random variable since so is $\widehat{f}_n$. In general, we can define the expected risk w.r.t. P for every f in our hypothesis space by

$$L_P(f) := \mathbf{E}_P[\ell(Y, f(X))] = \int_{\mathsf{X}\times\mathsf{Y}} \ell(y, f(x)) P(\mathrm{d}x, \mathrm{d}y)$$

as well as the *minimum risk*

$$L_P^*(\mathcal{F}) := \inf_{f \in \mathcal{F}} L_P(f).$$

Conceptually, $L_P^*(\mathcal{F})$ is the best *possible* performance of any hypothesis in $\mathcal{F}$ when the samples are drawn from P; similarly, $L_P(\widehat{f}_n)$ is the *actual* performance of the algorithm with access to a training sample of size n. It is clear from definitions that

$$0 \le L_P^*(\mathcal{F}) \le L_P(\widehat{f}_n) \le 1.$$

The goal of learning is to guarantee that $L_P(\widehat{f}_n)$ is as close as possible to $L_P^*(\mathcal{F})$, whatever the true $P \in \mathcal{P}$ happens to be. In order to speak about this quantitatively, we need to assess the probability of getting a "bad" sample. To that end, we define, similarly to what we have done earlier, the quantity

$$r_{\mathcal{A}}(n,\varepsilon) := \sup_{P\in\mathcal{P}} P^n\left(Z^n \in \mathsf{Z}^n : L_P(\widehat{f}_n) > L_P^*(\mathcal{F}) + \varepsilon\right) \tag{5.18}$$

for every $\varepsilon > 0$. Thus, a sample $Z^n \sim P^n$ is declared to be "bad" if it leads to a hypothesis whose expected risk on an independent test point $(X,Y) \sim P$ is greater than the smallest possible loss $L_P^*(\mathcal{F})$ by more than ε. We have the following:

Definition 5.5. *(PAC for model-free setting) We say that a learning algorithm for a problem* $(\mathsf{X},\mathsf{Y},\mathsf{U},\mathcal{P},\mathcal{F},\ell)$ *is* PAC to accuracy ε *if*

$$\lim_{n\to\infty} r_{\mathcal{A}}(n,\varepsilon) = 0.$$

An algorithm that is PAC to accuracy ε for every $\varepsilon > 0$ is said to be PAC. A learning problem specified by a tuple $(\mathsf{X},\mathsf{Y},\mathsf{U},\mathcal{P},\mathcal{F},\ell)$ *is* model-free *(or* agnostically*)* learnable *(to accuracy ε) if there exists an algorithm for it which is PAC (to accuracy ε).*

Let us look at two examples.

5.3.1. Function learning in the realizable case. First we show that the model-free framework contains the realizable set-up as a special case. To see this, let X be an arbitrary space and let $\mathsf{Y} = \mathsf{U} = [0,1]$. Let $\mathcal{F}$ be a class of functions $f : \mathsf{X} \to [0,1]$. Let $\mathcal{P}_{\mathsf{X}}$ be a family of probability distributions P_X on X. To each P_X and each $f \in \mathcal{F}$ associate a probability distribution $P_{X,f}$ on $\mathsf{X}\times\mathsf{Y}$ as follows: let $X \sim P_X$, and let the conditional distribution of Y given $X = x$ be given by

$$P_{Y|X,f}(B|X=x) = \mathbf{1}_{\{f(x)\in B\}}$$

for all (measurable) sets $B \subseteq \mathsf{Y}$. The resulting joint distribution $P_{X,f}$ is then uniquely defined by its action on the rectangles $A \times B$, $A \subseteq \mathsf{X}$ and $B \subseteq \mathsf{Y}$:

$$P_{X,f}(A\times B) := \int_A P_{Y|X,f}(B|x)P_X(\mathrm{d}x) = \int_A \mathbf{1}_{\{f(x)\in B\}}P_X(\mathrm{d}x)$$

Finally, let $\mathcal{P} = \{P_{X,f} : f \in \mathcal{F}, P_X \in \mathcal{P}_{\mathsf{X}}\}$. Finally, let $\ell(y,u) := |y-u|^2$.

Now, fixing a probability distribution $P \in \mathcal{P}$ is equivalent to fixing some $P_X \in \mathcal{P}_{\mathsf{X}}$ and some $f \in \mathcal{F}$. A random element of $\mathsf{Z} = \mathsf{X}\times\mathsf{Y}$ drawn according to such a P has the form $(X, f(X))$, where $X \sim P_X$. An i.i.d. sequence $(X_1,Y_1),\dots,(X_n,Y_n)$ drawn according to P therefore has the form

$$(X_1, f(X_1)),\dots,(X_n, f(X_n)),$$

which is precisely what we had in our discussion of function learning in the realizable case. Next, for any $P = P_{X,f} \in \mathcal{P}$ and any other $g \in \mathcal{F}$, we have

$$\begin{aligned} L_{P_{X,f}}(g) &= \int_{\mathsf{X}\times\mathsf{Y}} |y - g(x)|^2 P_{X,f}(\mathrm{d}x, \mathrm{d}y) \\ &= \int_{\mathsf{X}\times\mathsf{Y}} \mathbf{1}_{\{y=f(x)\}} |y - g(x)|^2 P_X(\mathrm{d}x) \\ &= \int_{\mathsf{X}} |f(x) - g(x)|^2 P_X(\mathrm{d}x) \\ &= \|f - g\|^2_{L^2(P_X)}, \end{aligned}$$

which is precisely the risk $L_{P_X}(g, f)$ that we have considered in our function learning formulation earlier. Moreover,

$$L^*_{P_{X,f}} = \inf_{g\in\mathcal{F}} L_{P_{X,f}}(g) = \inf_{g\in\mathcal{F}} \|f - g\|^2_{L^2(P_X)} \equiv 0.$$

Therefore,

$$\begin{aligned} r_{\mathcal{A}}(n, \varepsilon) &= \sup_{P_{X,f}\in\mathcal{P}} P^n_{X,f}\left(Z^n \in \mathsf{Z}^n : L_{P_{X,f}}(\widehat{f}_n) > L^*_{P_{X,f}} + \varepsilon\right) \\ &= \sup_{P_X\in\mathcal{P}_{\mathsf{X}}} \sup_{f\in\mathcal{F}} P^n_X\left(X^n \in \mathsf{X}^n : L_P(\widehat{f}_n, f) > \varepsilon\right) \\ &\equiv \bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}_{\mathsf{X}}). \end{aligned}$$

Thus, the function learning problem in the realizable case can be covered under the model-free framework as well.

5.3.2. Learning to classify with noisy labels. An agnostic classification problem can be built up from a realizable classification problem, as explained in this section. Start with a realizable classification problem with binary labels $(\mathsf{X}, \mathcal{P}_{\mathsf{X}}, \mathcal{C})$, under the usual 0-1 loss. Here $\mathcal{P}_{\mathsf{X}}$ is a family of probability measures on (the Borel subsets of) X. Let C^* denote the target (i.e. true) concept and let C be another concept. Then for a given $P_{\mathsf{X}} \in \mathcal{P}_{\mathsf{X}}$ and target concept C^*, the expected loss for using concept C to classify is $L_{P_{\mathsf{X}}}(C^*, C) = P_{\mathsf{X}}(C^* \triangle C)$. A learning algorithm $\mathcal{A} = (A_n)_{\{n\geq n\}}$ is sought to produce $\widehat{C}_n = A_n(Z^n)$ that makes $P_X(C^* \triangle \widehat{C}_n)$ small with high probability.

Let $0 \leq \eta < 1/2$ denote a crossover probability. Given the above realizable classification problem, we can define a corresponding agnostic one by modeling the labels as noisy labels. Specifically, the agnostic model is denoted by $(\mathsf{X}, \mathsf{Y} = \{0, 1\}, \mathcal{P}, \mathcal{C})$, again under 0-1 loss. An element P_{X,C^*} of $\mathcal{P}$ corresponds to a (P_{X}, C^*) pair, such that P_{X,C^*} is the joint probability distribution of (X, Y) where X has probability distribution P_{X} and $Y = \mathbf{1}_{\{X\in C^*\}} \oplus W$, where W is independent of X with the Bernoulli(η) probability distribution, and "$\oplus$" denotes XOR (i.e. modulo 2) addition. In words, the original label of an input X, $\mathbf{1}_{\{X\in C^*\}}$, is toggled with probability η to produce the label Y.

This model fits into the agnostic framework if it is interpreted appropriately, as follows. Given the true concept is C^* and an instance of X and W, we interpret $\mathbf{1}_{\{X\in C^*\}}$ as a nominal or preliminary label, and $Y = \mathbf{1}_{\{X\in C^*\}} \oplus W$ as the true label. Thus, as usual in this section, the training data and test data are generated independently with the same joint distribution of (X, Y). Assuming the true concept is C^* and C is a given classifier, the loss for using C to predict Y is one if and only if $\mathbf{1}_{\{X \in C\}} \neq Y$. Therefore, the expected loss for truth P_{X,C^*} and classifier C satisfies the following:

$$\begin{aligned}
L_{P_{\mathsf{X},C^*}}(C) &= P_{\mathsf{X},C^*}(\mathbf{1}_{\{X \in C\}} \neq Y) \\
&= P_X(\mathbf{1}_{\{X \in C\}} \neq \mathbf{1}_{\{X \in C^*\}} \oplus W) \\
&= P_X(W \neq \mathbf{1}_{\{X \in C^* \triangle C\}}) \\
&= (1-\eta) P_X(C^* \triangle C) + \eta(1 - P_X(C^* \triangle C)) \\
&= \eta + (1-2\eta) P_X(C^* \triangle C)
\end{aligned}$$

Thus, $L_{P_{\mathsf{X},C^*}}(C)$ is a (linear) increasing function of $P_X(C^* \triangle C)$. Hence, to have $L_{P_{\mathsf{X},C^*}}(\widehat{C}_n)$ close to its minimum possible value, we want $\widehat{C}_n$ to make $P_X(C^* \triangle \widehat{C}_n)$ as small as possible, just as in the realizable case. In other words, we want to maximize the probability that $\widehat{C}_n$ correctly predicts the nominal label determined by C^*. So, in the end, we could view the nominal label $\mathbf{1}_{\{X \in C^*\}}$ as the true label, and Y as a noisy version of the true label.

5.4. Empirical risk minimization

Having formulated the model-free learning problem, we must now turn to the question of how to construct PAC learning algorithms (and the related question of when a hypothesis class is PAC-learnable in the model-free setting).

We will first start with a heuristic argument and then make it rigorous. Suppose we are faced with the learning problem specified by $(\mathsf{X}, \mathsf{Y}, \mathsf{U}, \mathcal{P}, \mathcal{F}, \ell)$. Given a training set $Z^n = (Z_1, \ldots, Z_n)$, where each $Z_i = (X_i, Y_i)$ is independently drawn according to some unknown $P \in \mathcal{P}$, what should we do? The first thing to note is that, for any hypothesis $f \in \mathcal{F}$, we can approximate its risk $L_P(f)$ by the *empirical risk*

$$\frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i)), \tag{5.19}$$

whose expectation w.r.t. the distribution of Z^n is clearly equal to $L_P(f)$. In fact, since ℓ is bounded between 0 and 1, Hoeffding's inequality tells us that

$$\left| \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i)) - L_P(f) \right| < \varepsilon \qquad \text{with probability at least } 1 - 2e^{-2n\varepsilon^2}.$$

We can express these statements more succinctly if we define, for each $f \in \mathcal{F}$, the function $\ell_f : \mathsf{Z} \to [0,1]$ by

$$\ell_f(z) \equiv \ell_f(x,y) := \ell(y, f(x)). \tag{5.20}$$

Then the empirical risk (5.19) is just the expectation of ℓ_f w.r.t. the empirical distribution P_n:

$$P_n(\ell_f) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i)),$$

and, since $L_P(f) = \mathbf{E}_P[\ell(Y, f(X))] = P(\ell_f)$, we will have

$$|P_n(\ell_f) - P(\ell_f)| < \varepsilon \qquad \text{with probability at least } 1 - 2e^{-2n\varepsilon^2}. \tag{5.21}$$

Now, given the data Z^n we can compute the empirical risks $P_n(\ell_f)$ for every f in our hypothesis class $\mathcal{F}$. Since (5.21) holds for each $f \in \mathcal{F}$ individually, we may intuitively claim that the empirical risk for each f is a sufficiently accurate estimator of the corresponding true risk $L_P(f) \equiv P(\ell_f)$. Thus, a reasonable learning strategy would be to find any $\widehat{f}_n \in \mathcal{F}$ that would *minimize* the empirical risk, i.e., take

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} P_n(\ell_f) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)). \tag{5.22}$$

The reason why we would expect something like (5.22) to work is as follows: if a given f^* is a minimizer of $L_P(f) = P(\ell_f)$ over $\mathcal{F}$,

$$f^* = \arg\min_{f \in \mathcal{F}} P(\ell_f),$$

then its empirical risk, $P_n(\ell_{f^*})$, will be close to $L_P(f^*) = P(\ell_{f^*}) = L_P^*(\mathcal{F})$ with high probability. Moreover, it makes sense to expect that, in some sense, $\widehat{f}_n$ defined in (5.22) would be close to f^*, resulting in something like

$$P(\ell_{\widehat{f}_n}) \approx P_n(\ell_{\widehat{f}_n}) \approx P_n(\ell_{f^*}) \approx P(\ell_{f^*})$$

with high probability.

Unfortunately, this is not true in general. However, as we will now see, it is true under certain regularity conditions on the objects $\mathcal{P}$, $\mathcal{F}$, and ℓ. In order to state these regularity conditions precisely, let us define the *induced loss function class*

$$\mathcal{L}_{\mathcal{F}} := \{\ell_f : f \in \mathcal{F}\}.$$

Each $\ell_f \in \mathcal{L}_{\mathcal{F}}$ corresponds to the hypothesis $f \in \mathcal{F}$ via (5.20). Now, for any $n \in \mathbb{N}$ and any $\varepsilon > 0$ let us define

$$q(n, \varepsilon) := \sup_{P \in \mathcal{P}} P^n \left(Z^n \in \mathsf{Z}^n : \sup_{f \in \mathcal{F}} |P_n(\ell_f) - P(\ell_f)| \geq \varepsilon \right). \tag{5.23}$$

For a fixed $P \in \mathcal{P}$, quantity $\sup_{f \in \mathcal{F}} |P_n(\ell_f) - P(\ell_f)|$ is the *worst-case deviation* between the empirical means $P_n(\ell_f)$ and their expectations $P(\ell_f)$ over the entire hypothesis class $\mathcal{F}$. Given P, we say that an i.i.d. sample $Z^n \in \mathsf{Z}^n$ is "bad" if there exists at least one $f \in \mathcal{F}$, for which

$$|P_n(\ell_f) - P(\ell_f)| \geq \varepsilon.$$

Equivalently, a sample is bad if

$$\sup_{f \in \mathcal{F}} |P_n(\ell_f) - P(\ell_f)| \geq \varepsilon.$$

The quantity $q(n, \varepsilon)$ then compensates for the fact that P is unknown by considering the *worst case* over the entire class $\mathcal{P}$. With this in mind, we make the following definition:

Definition 5.6. *We say that the induced class $\mathcal{L}_{\mathcal{F}}$ has the* uniform convergence of empirical means (UCEM) *property w.r.t. $\mathcal{P}$ if*

$$\lim_{n \to \infty} q(n, \varepsilon) = 0$$

for every $\varepsilon > 0$.

THEOREM 5.3. *If the induced class $\mathcal{L}_{\mathcal{F}}$ has the UCEM property, then the* empirical risk minimization (ERM) *algorithm of* (5.22) *is PAC.*

PROOF. Fix $\varepsilon, \delta > 0$. We will now show that we can find a sufficiently large $n(\varepsilon, \delta)$, such that $r_{\mathcal{A}}(n, \varepsilon) \le \delta$ for all $n \ge n(\varepsilon, \delta)$, where $r_{\mathcal{A}}(n, \varepsilon)$ is defined in (5.18).

Let $f^* \in \mathcal{F}$ minimize the true risk w.r.t. P, i.e., $P(f^*) = L_P^*(\mathcal{F})$. For any n, we have

$$\begin{aligned} L_P(\widehat{f}_n) - L_P^* &= P\big(\ell_{\widehat{f}_n}\big) - P\big(f^*\big) \\ &= \underbrace{P\big(\ell_{\widehat{f}_n}\big) - P_n\big(\ell_{\widehat{f}_n}\big)}_{T_1} + \underbrace{P_n\big(\ell_{\widehat{f}_n}\big) - P_n\big(\ell_{f^*}\big)}_{T_2} + \underbrace{P_n\big(\ell_{f^*}\big) - P\big(\ell_{f^*}\big)}_{T_3}, \end{aligned}$$

where in the second line we have added and subtracted $P_n\big(\ell_{\widehat{f}_n}\big)$ and $P_n\big(\ell_{f^*}\big)$. We will now analyze the behavior of the three terms, T_1, T_2, and T_3. Since $\widehat{f}_n$ minimizes the empirical risk $P_n(\ell_f)$ over all $f \in \mathcal{F}$, we will have

$$T_2 = P_n\big(\ell_{\widehat{f}_n}\big) - P_n\big(\ell_{f^*}\big) \le 0.$$

Next,

$$T_1 = P\big(\ell_{\widehat{f}_n}\big) - P_n\big(\ell_{\widehat{f}_n}\big) \le \sup_{f \in \mathcal{F}} \left[P_n(\ell_f) - P(\ell_f)\right] \le \sup_{f \in \mathcal{F}} \left|P_n(\ell_f) - P(\ell_f)\right|,$$

and the same upper bound holds for T_3. Hence,

$$L_P(\widehat{f}_n) - L_P^*(\mathcal{F}) \le 2 \sup_{f \in \mathcal{F}} \left|P_n(\ell_f) - P(\ell_f)\right|. \tag{5.24}$$

Now, since $\mathcal{L}_{\mathcal{F}}$ has the UCEM property, we can find some sufficiently large $n_0(\varepsilon, \delta)$, such that

$$q(n, \varepsilon/2) = \sup_{P \in \mathcal{P}} P^n \left(Z^n \in \mathsf{Z}^n : \sup_{f \in \mathcal{F}} \left|P_n(\ell_f) - P(\ell_f)\right| \ge \varepsilon/2 \right) \le \delta, \qquad \forall n \ge n_0(\varepsilon, \delta).$$

From this it follows that, for all $n \ge n_0(\varepsilon, \delta)$, we will have

$$P^n \left(Z^n : \sup_{f \in \mathcal{F}} \left|P_n(\ell_f) - P(\ell_f)\right| \ge \varepsilon/2 \right) \le \delta, \qquad \forall P \in \mathcal{P}.$$

From (5.24), we see that

$$L_P(\widehat{f}_n) \ge L_P^*(\mathcal{F}) + \varepsilon \qquad \Longrightarrow \qquad \sup_{f \in \mathcal{F}} \left|P_n(\ell_f) - P(\ell_f)\right| \ge \varepsilon/2$$

for all n. However, for all $n \ge n_0(\varepsilon, \delta)$ the latter event will occur with probability at most δ, no matter which P is in effect. Therefore, for all $n \ge n_0(\varepsilon, \delta)$ we will have

$$\begin{aligned} r_{\mathcal{A}}(n, \varepsilon) &= \sup_{P \in \mathcal{P}} P^n \left(Z^n : L_P(\widehat{f}_n) > L_P^*(\mathcal{F}) + \varepsilon \right) \\ &\le \sup_{P \in \mathcal{P}} P^n \left(Z^n : \sup_{f \in \mathcal{F}} \left|P_n(\ell_f) - P(\ell_f)\right| \ge \varepsilon/2 \right) \\ &\equiv q(n, \varepsilon/2) \\ &\le \delta, \end{aligned}$$

which is precisely what we wanted to show. Thus, $r_{\mathcal{A}}(n, \varepsilon) \to 0$ as $n \to \infty$ for every $\varepsilon > 0$, which means that the ERM algorithm is PAC. $\square$

This theorem shows that the UCEM property of the induced class $\mathcal{L}_\mathcal{F}$ is a sufficient condition for the ERM algorithm to be PAC. Now the whole affair rests on us being able to establish the UCEM property for various "interesting" and "useful" problem specifications. This will be our concern in the chapters ahead. However, let us give you a hint of what to expect. Let us start with the following result:

PROPOSITION 5.1 (Finite hypothesis classes). *Consider a learning problem* $(\mathsf{X}, \mathsf{Y}, \mathsf{U}, \mathcal{P}, \mathcal{F}, \ell)$, *where the hypothesis class is finite:* $|\mathcal{F}| < \infty$. *Then the induced function class* $\mathcal{L}_\mathcal{F}$ *satisfies*

$$q(n, \varepsilon) \le 2|\mathcal{F}|e^{-2n\varepsilon^2}, \tag{5.25}$$

and therefore has the UCEM property.

PROOF. By assumption, $\mathcal{F} = \{f_1, \ldots, f_M\}$, where $M := |\mathcal{F}| < \infty$. Therefore, for any $P \in \mathcal{P}$,

$$\begin{aligned} P^n & \left(Z^n \in \mathsf{Z}^n : \sup_{f \in \mathcal{F}} |P_n(\ell_f) - P(\ell_f)| \ge \varepsilon \right) \\ &= P^n \left(\bigcup_{m=1}^M \{ Z^n \in \mathsf{Z}^n : |P_n(\ell_{f_m}) - P(\ell_{f_m})| \ge \varepsilon \} \right) \\ &\le \sum_{m=1}^M P^n \big(Z^n \in \mathsf{Z}^n : |P_n(\ell_{f_m}) - P(\ell_{f_m})| \ge \varepsilon \big), \end{aligned}$$

by the union bound. Since the loss function ℓ takes values in $[0, 1]$, each of the terms in the sum above can be estimated using the Hoeffding bound:

$$P^n \big(Z^n \in \mathsf{Z}^n : |P_n(\ell_{f_m}) - P(\ell_{f_m})| \ge \varepsilon \big) \le 2e^{-2n\varepsilon^2}.$$

Combining these estimates and taking the supremum over $P \in \mathcal{P}$, we get (5.25). ☐

As we will see shortly, in many cases, even if the hypothesis class $\mathcal{F}$ is infinite, we will be able to show that the induced class $\mathcal{L}_\mathcal{F}$ is so well-behaved that the bound

$$\mathbf{E}_{P^n} \left[\sup_{f \in \mathcal{F}} |P_n(\ell_f) - P(\ell_f)| \right] \le \frac{C_{\mathcal{F},\ell}}{\sqrt{n}} \tag{5.26}$$

holds for *every* P, where $C_{\mathcal{F},\ell} > 0$ is some constant that depends only on the characteristics of the hypothesis class $\mathcal{F}$ and the loss function ℓ. Since ℓ_f is bounded between 0 and 1, the function

$$g(Z^n) := \sup_{f \in \mathcal{F}} |P_n(\ell_f) - P(\ell_f)|$$

has bounded differences with constants $c_1 = \ldots = c_n = 1/n$. McDiarmid's inequality then tells us that, for any $t > 0$,

$$P^n \Big(g(Z^n) - \mathbf{E} g(Z^n) \ge t \Big) \le e^{-2nt^2}. \tag{5.27}$$

Let

$$n_0(\varepsilon, \delta) := \max \left\{ \frac{4C_{\mathcal{F},\ell}^2}{\varepsilon^2}, \frac{2}{\varepsilon^2} \log \left(\frac{1}{\delta} \right) \right\} + 1. \tag{5.28}$$

Then for any $n \geq n_0(\varepsilon, \delta)$

$$
\begin{aligned}
P^n\Big(g(Z^n) \geq \varepsilon\Big) &= P^n\Big(g(Z^n) - \mathbf{E}g(Z^n) \geq \varepsilon - \mathbf{E}g(Z^n)\Big) \\
&\leq P^n\left(g(Z^n) - \mathbf{E}g(Z^n) \geq \varepsilon - \frac{C_{\mathcal{F},\ell}}{\sqrt{n}}\right) && \text{because of (5.26)} \\
&\leq P^n\left(g(Z^n) - \mathbf{E}g(Z^n) \geq \frac{\varepsilon}{2}\right) && \text{because } n > \frac{4C_{\mathcal{F},\ell}^2}{\varepsilon^2} \\
&\leq e^{-n\varepsilon^2/2} && \text{using (5.27) with } t = \varepsilon/2 \\
&< \delta && \text{because } n > \frac{2}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)
\end{aligned}
$$

for *any* probability distribution P over $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$. Thus, we have derived a very important fact: If the induced loss class $\mathcal{L}_{\mathcal{F}}$ satisfies (5.26), then (a) it has the UCEM property, and consequently is model-free learnable using the ERM algorithm, and (b) the sample complexity is quadratic in $1/\varepsilon$ and logarithmic in $1/\delta$. Our next order of business, taken up in the next two chapters, will be to derive sufficient conditions on $\mathcal{F}$ and ℓ for something like (5.26) to hold.

5.5. The mismatched minimization lemma

Since it arises repeatedly, we isolate an idea used in the proof of Theorem 5.3 so we can easily refer to it later. Suppose we'd like to find a minimizer of a function G defined on some domain U, but only an approximation, $\widehat{G}$, of G is available. Then the following lemma implies a minimizer of $\widehat{G}$ nearly minimizes G as well, if $\widehat{G}$ is uniformly close to G.

Lemma 5.1. *(Mismatched minimization lemma) Suppose that $\widehat{G}$ is an ε uniform approximation of G for some $\varepsilon > 0$, meaning that $|G(u) - \widehat{G}(u)| \leq \varepsilon$ for all $u \in \mathsf{U}$.*
(a) (Single version) For any $\widehat{u} \in \mathsf{U}$, $G(\widehat{u}) \leq \widehat{G}(\widehat{u}) + \varepsilon$.
(b) (Double version) Suppose that u^* is a minimizer of $\widehat{G}$, meaning that $u^* \in \mathsf{U}$ and $\widehat{G}(u^*) \leq \widehat{G}(u)$ for all $u \in \mathsf{U}$. Then $G(u^*) \leq \inf_{u \in \mathsf{U}} G(u) + 2\varepsilon$.

Proof. Part (a) is immediate from the assumption $|G(u) - \widehat{G}(u)| \leq \varepsilon$ for all $u \in \mathsf{U}$. For any $u \in \mathsf{U}$, $G(u) \geq \widehat{G}(u) - \varepsilon \geq \widehat{G}(u^*) - \varepsilon \geq G(u^*) - 2\varepsilon$. Therefore, $\inf_{u \in \mathsf{U}} G(u) \geq G(u^*) - 2\varepsilon$, which is equivalent to part (b). The proof of part (b) is illustrated in Figure 4. $\square$

Lemma 5.1 implies (5.24) as follows. Let P be fixed. The expected loss as a function of the predictor, $f \mapsto P(\ell_f)$, is approximated by the empirical expected loss as a function of the predictor, $f \mapsto P_n(\ell_f)$. The bound $|P(\ell_f) - P_n(\ell_f)| \leq \varepsilon$ holds uniformly over $f \in \mathcal{F}$ for $\varepsilon = \sup_{f \in \mathcal{F}} |P(\ell_f) - P_n(\ell_f)|$. So (5.24) is an instance of the double version of Lemma 5.1.

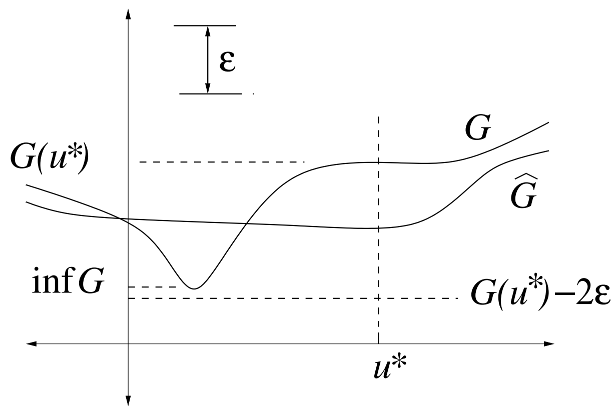

FIGURE 4. Illustration of the double version of the mismatched minimization lemma

CHAPTER 6

Empirical Risk Minimization: Abstract risk bounds and Rademacher averages

In the last chapter, we have left off with a theorem that gave a sufficient condition for the *Empirical Risk Minimization* (ERM) algorithm

$$\widehat{f}_n = \operatorname*{arg\,min}_{f \in \mathcal{F}} P_n(\ell_f) \tag{6.1}$$

$$= \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) \tag{6.2}$$

to be PAC for a given learning problem with hypothesis space $\mathcal{F}$ and loss function ℓ. This condition pertained to the behavior of the uniform deviation of empirical means from true means over the *induced class* $\mathcal{L}_{\mathcal{F}} = \{\ell_f : f \in \mathcal{F}\}$. Specifically, we proved that ERM is a PAC algorithm if

$$\lim_{n\to\infty} \sup_{P \in \mathcal{P}} P^n \left(\sup_{f \in \mathcal{F}} |P_n(\ell_f) - P(\ell_f)| \ge \varepsilon \right) = 0, \qquad \forall \varepsilon > 0, \tag{6.3}$$

where $\mathcal{P}$ is the class of probability distributions generating the training data.

6.1. An abstract framework for ERM

To study ERM in a general framework, we will adopt a simplified notation often used in the literature. We have a space Z and a class $\mathcal{F}$ of functions $f : \mathsf{Z} \to [0,1]$. Let $\mathcal{P}(\mathsf{Z})$ denote the space of all probability distributions on Z. For each sample size n, the training data are in the form of an n-tuple $Z^n = (Z_1, \dots, Z_n)$ of Z-valued random variables drawn according to some unknown $P \in \mathcal{P}$. For each P, we can compute the *expected risk* of any $f \in \mathcal{F}$ by

$$P(f) = \mathbf{E}_P f(Z) = \int_{\mathsf{Z}} f(z) P(\mathrm{d}z). \tag{6.4}$$

The *minimum risk* over $\mathcal{F}$ is

$$L^*_P(\mathcal{F}) := \inf_{f \in \mathcal{F}} P(f). \tag{6.5}$$

A learning algorithm is a sequence $\mathcal{A} = \{A_n\}_{n \ge 1}$ of mappings $A_n : \mathsf{Z}^n \to \mathcal{F}$, and the objective is to ensure that

$$P(\widehat{f}_n) \approx L^*_P(\mathcal{F}) \qquad \text{eventually with high probability.} \tag{6.6}$$

The ERM algorithm works by taking

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} P_n(f) \tag{6.7}$$

$$= \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i). \tag{6.8}$$

This way of writing down our problem hides most of the ingredients that were specified in Haussler's framework of model-free learning, so it is important to keep in mind that Z is an input/output pair (X, Y) and the functions $f \in \mathcal{F}$ are really the induced losses for some loss function ℓ and hypothesis class $\mathcal{G}$. That is, they are functions of the form $\ell_g(x, y) = \ell(g(x), y)$, or, in the case if classification for a set of concepts $\mathcal{C}$ and $0-1$ loss, functions of the form $\ell_C(x, y) = \mathbf{1}_{\{\mathbf{1}_{\{x \in C\}} \neq y\}}$. However, recalling our discussion of unsupervised learning problems in Chapter 1, we do not insist on splitting Z into input X and output Y, nor do we need to imagine any particular structure for f.

We have seen (Theorem 5.3) that a uniform bound on the deviation empirical means in $\mathcal{F}$ is a suffiicent condition for the consistency of ERM. In order to have a clean way of keeping track of all the relevant quantities, let us introduce some additional notation. First of all, we need a way of comparing the behavior of any two probability distributions P and P' with respect to the class $\mathcal{F}$. A convenient way of doing this is through the $\mathcal{F}$-*seminorm*

$$\|P - P'\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |P(f) - P'(f)| \tag{6.9}$$

$$= \sup_{f \in \mathcal{F}} |\mathbf{E}_P f - \mathbf{E}_{P'} f| \tag{6.10}$$

$$= \sup_{f \in \mathcal{F}} \left| \int_{\mathsf{Z}} f(z) P(\mathrm{d}z) - \int_{\mathsf{Z}} f(z) P'(\mathrm{d}z) \right|. \tag{6.11}$$

We say that $\|\cdot\|_{\mathcal{F}}$ is a *seminorm* because it has all the properties of a norm (in particular, it satisfies the triangle inequality), but it may happen that $\|P - P'\|_{\mathcal{F}} = 0$ for $P \neq P'$. Next, given a random sample Z^n we define the *uniform deviation*

$$\Delta_n(Z^n) := \|P_n - P\|_{\mathcal{F}} \equiv \sup_{f \in \mathcal{F}} |P_n(f) - P(f)|. \tag{6.12}$$

To keep things simple, we do not indicate the underlying distribution P or the function class $\mathcal{F}$ explicitly. We will do this from now on, unless some confusion is possible, in which case we will use appropriate indices. Thus, we will write $L(f)$, $L^*(\mathcal{F})$, etc., and you should always keep in mind that all expectations are computed w.r.t. the (unknown) data-generating distribution $P \in \mathcal{P}(\mathsf{Z})$. In the same spirit, we will denote by $P_n(f)$ the *empirical risk* of f on the sample Z^n:

$$P_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(Z_i). \tag{6.13}$$

The following result is key to understanding the role of the uniform deviations $\Delta_n(Z^n)$ in controlling the performance of the ERM algorithm.

Proposition 6.1. *The generalization loss for a learning algorithm satisfies:*

$$P(\widehat{f}_n) \leq L^*(\mathcal{F}) + 2\Delta_n(Z^n) \qquad \textit{(if algorithm is ERM)} \tag{6.14}$$

$$P(\widehat{f}_n) \leq P_n(\widehat{f}_n) + \Delta_n(Z^n) \qquad \textit{(for any algorithm).} \tag{6.15}$$

Proof. The proposition follows immediately from the mismatched minimization lemma, Lemma 5.1, with (6.14) corresponding to the double version of Lemma 5.1 and (6.15) corresponding to the single version. Let us give another proof in order to get comfortable with the abstract notation of this section. Let f^* be any minimizer of $P(f)$ over $\mathcal{F}$. Then

$$\begin{aligned} P(\widehat{f}_n) - L^*(\mathcal{F}) &= P(\widehat{f}_n) - P(f^*) \\ &= P(\widehat{f}_n) - P_n(\widehat{f}_n) + P_n(\widehat{f}_n) - P_n(f^*) + P_n(f^*) - P(f^*), \end{aligned}$$

where $P_n(\widehat{f}_n) - P_n(f^*) \leq 0$ by definition of ERM,

$$P(\widehat{f}_n) - P_n(\widehat{f}_n) \leq \sup_{f \in \mathcal{F}} [P_n(f) - P(f)] \leq \|P_n - P\|_{\mathcal{F}} = \Delta_n(Z^n),$$

and the same holds for $P_n(f^*) - P(f^*)$. This proves (6.14), while (6.15) is immediate from (6.12). □

Remark 6.1. The bounds (6.14) and (6.15) are both useful in practice, and they have different meanings. The bound (6.14) says that, if the uniform deviation $\Delta_n(Z^n)$ is small, then the expected risk of the ERM hypothesis will be close to the minimum risk $L^*(\mathcal{F})$. That is nice to know even though $L^*(\mathcal{F})$ is typically not computable from the data. The bound (6.15) says that the empirical estimate $P_n(\widehat{f}_n)$ is an accurate estimate of the generalization performance of $\widehat{f}_n$, and, of course, $P_n(\widehat{f}_n)$ is computable from the data. Both bounds suggest that the success of ERM depends on how small the uniform deviation $\Delta_n(Z^n)$ can be. Thus, we need to develop tools for analyzing the behavior of $\Delta_n(Z^n)$.

6.2. Bounding the uniform deviation: Rademacher averages

Motivated by Proposition 6.1, we would like to have conditions implying $\Delta_n(Z^n)$ is small with high probability. That can be achieved by first bounding $\mathbf{E}[\Delta_n(Z^n)]$ and then showing that the distribution of $\Delta_n(Z^n)$ is concentrated about its mean using McDiarmid's inequality. The following definition will play an important role for bounding $\mathbf{E}[\Delta_n(Z^n)]$.

Definition 6.1. *Let $\mathcal{A} \subset \mathbb{R}^n$ with $\mathcal{A}$ bounded. The* Rademacher average *of $\mathcal{A}$, denoted by $R_n(\mathcal{A})$, is defined by*

$$R_n(\mathcal{A}) = \mathbf{E}\left[\sup_{a \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i a_i \right| \right],$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher (i.e., ± 1 with equal probability) random variables.

To motivate the use of Rademacher averages, suppose that P and $\mathcal{F}$ are such that if $Z^n = (Z_1, \ldots, Z_n)$ is distributed according to P^n, then

$$\frac{1}{n} \sum_{i=1}^n f(Z_i) \approx P(f) \text{ for all } f \in \mathcal{F}, \text{ with high probability.} \tag{6.16}$$

Condition (6.16) has to do with $P, \mathcal{F}$, and n only; it doesn't matter which random vector Z^n is used, as long as it has distribution P^n. So if $Z_{n+1}, \ldots, Z_{2n}$ are n more random variables that are independent, each with distribution P, then

$$\frac{1}{n}\sum_{i=n+1}^{2n} f(Z_i) \approx P(f) \text{ for all } f \in \mathcal{F}, \text{ with high probability.} \tag{6.17}$$

If (6.16), and hence (6.17), are true, then the left-hand side of (6.16) minus the left-hand side of (6.17) is approximately zero, for all $f \in \mathcal{F}$, with high probability. That can be written as:

$$\frac{1}{n}\sum_{i=1}^{2n} \varepsilon_i f(Z_i) \approx 0 \text{ for all } f \in \mathcal{F}, \text{ with high probability,} \tag{6.18}$$

where $\varepsilon_1 = \cdots = \varepsilon_n = 1$ and $\varepsilon_{n+1} = \cdots = \varepsilon_{2n} = -1$. Furthermore, if $Z_1, \ldots, Z_{2n}$ are mutually independent with distribution P, and if $\pi : [2n] \to [2n]$ is a random permutation, uniformly distributed over all $(2n)!$ possibilities, and independent of Z^{2n}, then

$$\sum_{i=1}^{2n} \varepsilon_i f(Z_i) \stackrel{\mathrm{d}}{=} \sum_{i=1}^{2n} \varepsilon_i f(Z_{\pi(i)}) = \sum_{i=1}^{2n} \varepsilon_{\pi^{-1}(i)} f(Z_i) = \sum_{i=1}^{2n} \widetilde{\varepsilon}_i f(Z_i), \tag{6.19}$$

where $\widetilde{\varepsilon}_i = \varepsilon_{\pi^{-1}(i)}$, so that $\widetilde{\varepsilon}$ is uniformly distributed over all ± 1 vectors of length $2n$ with zero sum. The distribution of $\widetilde{\varepsilon}$ is close to the distribution of a vector of n iid Rademacher variables, if n is at least moderately large. To summarize, (6.16) implies that

$$\frac{1}{2n}\sum_{i=1}^{2n} \widetilde{\varepsilon}_i f(Z_i) \approx 0 \text{ for all } f \in \mathcal{F}, \text{ with high probability.} \tag{6.20}$$

It is reasonable to think the converse is true as well. If (6.20) is true, it means that the lefthand sides of (6.16) and (6.17), which are independent of each other, are close to each other with high probability. It seems that forces (6.16) and (6.17) to be true. That intuition is behind the theorem given next.

Consider a class $\mathcal{F}$ of functions $f : \mathsf{Z} \to [0,1]$ from our formulation of the ERM problem. The key result is that $\mathbf{E}[\Delta_n(Z^n)]$ is controlled by the mean of the Rademacher averages of the *random sets*

$$\mathcal{F}(Z^n) := \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\}. \tag{6.21}$$

A useful way to think about $\mathcal{F}(Z^n)$ is as a projection of $\mathcal{F}$ onto the random sample Z^n.

THEOREM 6.1. *Fix a space Z and let $\mathcal{F}$ be a class of functions $f : \mathsf{Z} \to [0,1]$. Then for any $P \in \mathcal{P}(\mathsf{Z})$*

$$\mathbf{E}\Delta_n(Z^n) \leq 2\mathbf{E}R_n(\mathcal{F}(Z^n)). \tag{6.22}$$

PROOF. The idea of the proof is to use a symmetrization argument due to Vapnik-Chervonenkis, and focused on by Giné and Zinn [**GZ84**]. Fix $P \in \mathcal{P}$ throughout the proof. By definition,

$$\mathbf{E}\left[\Delta_n(Z^n)\right] = \mathbf{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(Z_i) \; - \; P(f) \right| \right].$$

Note that the mapping

$$y \mapsto \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) \;-\; y(f)\right|\right]$$

is a convex mapping of $y = (y(f) : f \in \mathcal{F})$ to $\mathbb{R}$, because the absolute value function is convex, the supremum of a set of convex functions is convex, and the expectation of a random convex function is convex. Let $\overline{Z}^n$ be an independent copy of Z^n. For example, we could take $\overline{Z}_i = Z_{n+i}$ where $Z_1, \ldots, Z_{2n}$ are distributed as above. Note that $\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(\overline{Z}_i)\right] = P(f)$. So by Jensen's inequality,

$$\mathbf{E}\left[\Delta_n(Z^n)\right] \le \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - f(\overline{Z}_i)\right|\right].$$

For each i, $(f(Z_i), f(\overline{Z}_i)) \overset{\mathrm{d}}{=} (f(\overline{Z}_i), f(Z_i))$, so that $f(Z_i) - f(\overline{Z}_i) \overset{\mathrm{d}}{=} f(\overline{Z}_i) - f(Z_i)$. That is, the distribution of $f(Z_i) - f(\overline{Z}_i)$ is symmetric. Thus, if $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher random variables, $(f(Z_i) - f(\overline{Z}_i))_{1\le i\le n} \overset{\mathrm{d}}{=} (\varepsilon_i(f(Z_i) - f(\overline{Z}_i)))_{1\le i\le n}$. Thus,

$$\begin{aligned}
\mathbf{E}\left[\Delta_n(Z^n)\right] &\le \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i(f(Z_i) - f(\overline{Z}_i))\right|\right] \\
&\le \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i)\right|\right] + \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(\overline{Z}_i))\right|\right] = 2\mathbf{E}\left[R_n(\mathcal{F}(Z^n))\right].
\end{aligned}$$

$\square$

The above theorem and Proposition 6.1 imply the following key result on ERM.

Corollary 6.1. *For any $P \in \mathcal{P}(\mathsf{Z})$ and any n, a learning algorithms $\widehat{f}_n$ satisfies the bound*

$$P(\widehat{f}_n) \le L^*(\mathcal{F}) + 4\,\mathbf{E}R_n(\mathcal{F}(Z^n)) + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}} \text{ (if ERM is used)} \tag{6.23}$$

$$P(\widehat{f}_n) \le P_n(\widehat{f}_n) + 2\,\mathbf{E}R_n(\mathcal{F}(Z^n)) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}} \text{ (for any algorithm)} \tag{6.24}$$

with probability at least $1-\delta$.

Proof. By the first bound in Proposition 6.1,

$$P(\widehat{f}_n) \le L^*(\mathcal{F}) + 2\Delta_n(Z^n) = L^*(\mathcal{F}) + 2\mathbf{E}[\Delta_n(Z^n)] + (2\Delta_n(Z^n) - 2\mathbf{E}[\Delta_n(Z^n)])$$

with probability one, and by Theorem 6.1, $\mathbf{E}[2\Delta_n(Z^n)] \le 4\mathbf{E}[R_n(\mathcal{F}(Z^n))]$. It thus suffices to prove the probability that $2\Delta_n(Z^n)$ exceeds its mean by more than $\varepsilon = \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}$ is less than or equal to δ. Examining the definition (6.12) of $\triangle(Z^n)$ shows that, as a function of Z^n, it has the bounded difference property for constants $c_i = \frac{1}{n}$ for all i. Thus, by McDiarmid's inequality,

$$P\{2\Delta_n \geq 2\mathbf{E}[\Delta_n] + \varepsilon\} = P\{\Delta_n \geq \mathbf{E}[\Delta_n] + \varepsilon/2\}$$
$$\leq \exp\left(-\frac{2(\varepsilon/2)^2}{n(1/n)^2}\right) = \exp\left(-n\varepsilon^2/2\right) = \delta,$$

as desired. $\square$

6.3. Structural results for Rademacher averages

The results developed above highlight the fundamental role played by Rademacher averages in bounding the generalization error of the ERM algorithm. In order to use these bounds, we need to get a better handle on the behavior of Rademacher averages. Recall that the Rademacher average of a bounded set $\mathcal{A} \subset \mathbb{R}^n$ is defined as

$$R_n(\mathcal{A}) := \frac{1}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{a\in\mathcal{A}}\left|\sum_{i=1}^n \varepsilon_i a_i\right|\right]$$

(see Def. 6.1). It will also be convenient to introduce an alternative version, without the absolute value:

$$R_n^\circ(\mathcal{A}) := \frac{1}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^n \varepsilon_i a_i\right].$$

Both variants are used in the literature, and each has its pros and cons.

There is a natural geometrical interpretation of Rademacher averages. Note that if we let ε denote the vector of Rademacher random variables, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^*$, then $\frac{1}{\sqrt{n}}\varepsilon$ is a random unit length vector, and $R_n(\mathcal{A}) = \frac{1}{\sqrt{n}}\mathbf{E}\left[\sup_{a\in\mathcal{A}\cup-\mathcal{A}}\langle a, \frac{1}{\sqrt{n}}\varepsilon\rangle\right]$. Note that $\sup_{a\in\mathcal{A}\cup-\mathcal{A}}\langle a, \frac{1}{\sqrt{n}}\varepsilon\rangle$ is the half-width of the smallest slab normal to ε that contains $\mathcal{A}\cup-\mathcal{A}$. Thus, $R_n(\mathcal{A})\sqrt{n}$ is the average, over ε, of such half-widths. For a similar geometrical interpretation for R°, use the fact that ε and $-\varepsilon$ have the same distribution to obtain $R_n^\circ(\mathcal{A}) = \frac{1}{\sqrt{n}}\mathbf{E}\left[\sup_{a\in\mathcal{A}}\langle a, \frac{1}{\sqrt{n}}\varepsilon\rangle\right] = -\frac{1}{\sqrt{n}}\mathbf{E}\left[\inf_{a\in\mathcal{A}}\langle a, \frac{1}{\sqrt{n}}\varepsilon\rangle\right]$. Therefore,

$$R_n^\circ(\mathcal{A}) = \frac{1}{2\sqrt{n}}\mathbf{E}\left[\sup_{a\in\mathcal{A}}\langle a, \frac{1}{\sqrt{n}}\varepsilon\rangle - \inf_{a\in\mathcal{A}}\langle a, \frac{1}{\sqrt{n}}\varepsilon\rangle\right]$$

Thus, $R_n^\circ(\mathcal{A})\sqrt{n}$ is the average half-width of the smallest slab containing $\mathcal{A}$, such that the slab is normal to ε, averaged over the possible values of ε.[1] See the illustration in Figure 6.3.

First, we collect some results that describe how Rademacher averages behave with respect to several operations on sets. Let $\mathcal{A}, \mathcal{B}$ be two bounded subsets of $\mathbb{R}^n$. In addition to the union $\mathcal{A}\cup\mathcal{B}$ and the intersection $\mathcal{A}\cap\mathcal{B}$, we can also define:

- the translate of $\mathcal{A}$ by $v \in \mathbb{R}^n$,

$$\mathcal{A} + v := \{a + v : a \in \mathcal{A}\};$$

[1]Similar measures for the size of a set $\mathcal{A}$ have been studied, corresponding to taking ε to be uniformly distributed over the unit sphere in $\mathbb{R}^n$ or taking ε to be a vector of independent standard normal random variables.

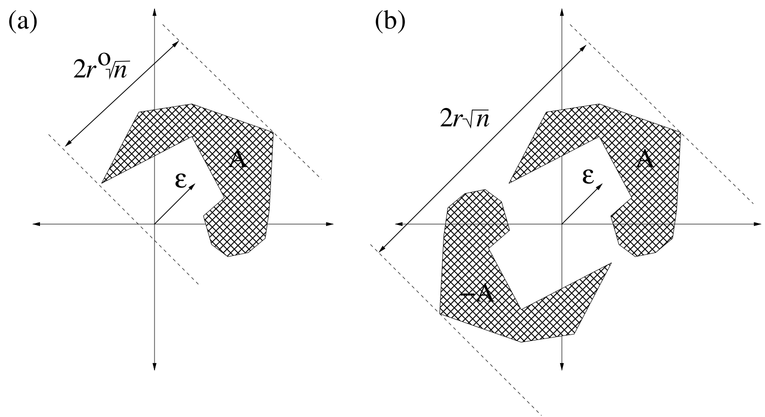

FIGURE 1. Illustration of the geometrical interpretation of Rademacher averages. $R_n^\circ(\mathcal{A})$ and $R_n(\mathcal{A})$, respectively, are the averages of r° and r shown, for ε uniformly distributed over the corners of the hypercube $[-1,1]^n$.

- the scaling of $\mathcal{A}$ by $c \in \mathbb{R}$,

$$c\mathcal{A} := \{ca : a \in \mathcal{A}\};$$

- the *Minkowski sum* of $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{A} + \mathcal{B} := \{a + b : a \in \mathcal{A}, b \in \mathcal{B}\};$$

- the *convex hull* and the *absolute convex hull* of $\mathcal{A}$:

$$\mathrm{conv}(\mathcal{A}) = \left\{ \sum_{m=1}^{N} c_m a^{(m)} : N \geq 1, c_m \geq 0 \text{ for } m \in [N], c_1 + \ldots + c_N = 1, \right.$$
$$\left. a^{(1)}, \ldots, a^{(N)} \in \mathcal{A} \right\}$$
$$\mathrm{absconv}(\mathcal{A}) = \left\{ \sum_{m=1}^{N} c_m a^{(m)} : N \geq 1, |c_1| + \ldots + |c_N| = 1, \right.$$
$$\left. a^{(1)}, \ldots, a^{(N)} \in \mathcal{A} \right\} = \mathrm{conv}(\mathcal{A} \cup (-\mathcal{A}))$$

Some basic properties of Rademacher averages are as follows:

(1) $R_n^\circ(\mathcal{A}) \leq R_n(\mathcal{A}) = R_n^\circ(\mathcal{A} \cup -\mathcal{A})$
(2) $R_n^\circ(\mathcal{A}) = R_n(\mathcal{A})$ if $\mathcal{A} = -\mathcal{A}$
(3) $R_n^\circ(\mathcal{A} + v) = R_n^\circ(\mathcal{A})$ for any $v \in \mathbb{R}^n$,
(4) $R_n(\mathcal{A} \cup \mathcal{B}) \leq R_n(\mathcal{A}) + R_n(\mathcal{B})$
(5) $R_n^\circ(\mathcal{A} + \mathcal{B}) = R_n^\circ(\mathcal{A}) + R_n^\circ(\mathcal{B})$
(6) $R_n(c\mathcal{A}) = |c| R_n(\mathcal{A})$
(7) $R_n(\mathcal{A}) = R_n(\mathrm{conv}(\mathcal{A}))$
(8) $R_n(\mathcal{A}) = R_n(\mathrm{absconv}(\mathcal{A}))$

These are all straightforward to prove. In particular, $R_n(\mathcal{A}) = R_n(\mathrm{conv}(\mathcal{A}))$ follows from the geometrical interpretation above, because a slab in $\mathbb{R}^d$ contains $\mathcal{A}$ if and only if it contains $\mathrm{conv}(\mathcal{A})$. Also, since $\mathrm{absconv}(\mathcal{A}) = \mathrm{conv}(\mathcal{A} \cup (-\mathcal{A}))$, it follows that

$$R_n(\mathrm{absconv}(\mathcal{A})) = R_n(\mathrm{conv}(\mathcal{A} \cup (-\mathcal{A}))) = R_n(\mathcal{A} \cup (-\mathcal{A})) = R_n(\mathcal{A}).$$

The properties listed above show what happens to Rademacher averages when we form combinations of sets. This will be useful to us later, when we talk about hypothesis classes made up of simpler classes by means of operations like set-theoretic unions, intersections, complements or differences, logical connectives, or convex and linear combinations.

The next result, often referred to as the finite class lemma, is based on the use of subgaussian random variables – see Section 2.5.

Lemma 6.1 (Finite class lemma). *If $\mathcal{A} = \{a^{(1)}, \ldots, a^{(N)}\} \subset \mathbb{R}^n$ is a finite set with $\|a^{(j)}\| \le L$ for all $j = 1, \ldots, N$ and $N \ge 2$, then*

$$R_n(\mathcal{A}) \le \frac{2L\sqrt{\log N}}{n}. \tag{6.25}$$

Proof. Let ε^n be a vector of n i.i.d. Rademacher variables, and for every $k \in [N]$ let

$$Y_k := \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i a_i^{(k)}.$$

A Rademacher random variable ε_i is mean zero and $|\varepsilon_i| \le 1$, so it is subgaussian with scale parameter one by Hoeffding's lemma, Lemma 2.1. Therefore, $\varepsilon_i a_i^{(k)}$ is subgaussian with scale parameter $|a_i^{(k)}|$. A sum of independent subgaussian random variables is also subgaussian, with the scale parameter of the sum given by the same formula as the standard deviation of the sum. Specifically, Y_k is subgaussian with scale paremeter ν given by $\nu^2 = \frac{1}{n^2}\sum_i (a_i^{(k)})^2 = \|a^{(k)}\|^2/n^2 \le (L/n)^2$. Since $|x| = \max\{x, -x\}$, we can write

$$R_n(\mathcal{A}) = \mathbf{E}\left[\max\left\{Y_1, -Y_1, \ldots, Y_N, -Y_N\right\}\right]. \tag{6.26}$$

The $2N$ random variables $\pm Y_1, \ldots, \pm Y_N$ are ν-subgaussian with $\nu = L/n$. Therefore, Lemma 2.3 implies

$$R_n(\mathcal{A}) \le \frac{L\sqrt{2\log(2N)}}{n} \le \frac{2L\sqrt{\log N}}{n},$$

where the last step uses the fact that $\log(2N) \le 2\log N$ for $N \ge 2$. □

Next we discuss the so-called *contraction principle* for Rademacher averages that will be used in later chapters. Suppose $n \ge 1$, and for $1 \le i \le n$ let $\varphi_i : \mathbb{R} \to \mathbb{R}$. Let $\varphi \circ v = (\varphi_1(v_1), \ldots, \varphi_n(v_n))$. Given a subset $\mathcal{A}$ of $\mathbb{R}^n$, let $\varphi \circ \mathcal{A} = \{\varphi \circ v : v \in A\}$. Recall that $R_n^\circ(\mathcal{A})$ is the variation of $R_n(\mathcal{A})$ obtained by omitting the absolute value symbols in the definition of $R_n(\mathcal{A})$.

Proposition 6.2 (Contraction principles for Rademacher averages). *If $\mathcal{A}$ is a bounded subset of $\mathbb{R}^n$ and for $i \in [n]$, $\varphi_i : \mathbb{R} \to \mathbb{R}$ is an M-Lipschitz continuous function, then $R_n^\circ(\varphi \circ \mathcal{A}) \le M R_n^\circ(\mathcal{A})$. Furthermore, if $\varphi_i(0) = 0$ for all i (i.e., $\varphi(\mathbf{0}) = \mathbf{0}$) then $R_n(\varphi \circ \mathcal{A}) \le 2M R_n(\mathcal{A})$.*

PROOF. We first prove the contraction property for R_n^o. By scaling, without loss of generality we can assume that $M = 1$. Since the functions $\varphi_1, \ldots, \varphi_n$ can be introduced one at a time, it suffices to consider the case that all the functions φ_i are equal to the identity function, except φ_1. That is, it suffices to show that $R_n^\circ(\mathcal{A}) = R_n^\circ(\mathcal{A}_1)$, where $\mathcal{A}_1 = \{(\varphi(a_1), a_2, \ldots, a_n) : a \in \mathcal{A}\}$ and $\varphi : \mathbb{R} \to \mathbb{R}$ is Lipschitz-continuous with constant one. Averaging over the values of ε_1, we get

$$\begin{aligned}
R_n^\circ(\mathcal{A}) &= \frac{1}{2n}\mathbf{E}\left\{\sup_{a\in\mathcal{A}}\left(a_1 + \sum_{i=2}^n \varepsilon_i a_i\right) + \sup_{a'\in\mathcal{A}}\left(-a'_1 + \sum_{i=2}^n \varepsilon_i a'_i\right)\right\} \\
&= \frac{1}{2n}\mathbf{E}\left\{\sup_{a,a'\in\mathcal{A}}\left(a_1 - a'_1 + \sum_{i=2}^n \varepsilon_i a_i + \sum_{i=2}^n \varepsilon_i a'_i\right)\right\} \\
&= \frac{1}{2n}\mathbf{E}\sup_{a,a'\in\mathcal{A}}\left(|a_1 - a'_1| + \sum_{i=2}^n \varepsilon_i a_i + \sum_{i=2}^n \varepsilon_i a'_i\right), \qquad (6.27)
\end{aligned}$$

where the last equality comes from the fact that a and a' can be swapped. Applying the same equations with $\mathcal{A}$ replaced by $\mathcal{A}_1$ yields:

$$R_n^\circ(\mathcal{A}_1) = \frac{1}{2n}\mathbf{E}\sup_{a,a'\in\mathcal{A}}\left(|\phi(a_1) - \phi(a'_1)| + \sum_{i=2}^n \varepsilon_i a_i + \sum_{i=2}^n \varepsilon_i a'_i\right). \qquad (6.28)$$

Comparison of (6.27) and (6.28) and using the assumption $|\varphi(a_1) - \varphi(a'_1)| \le |a_1 - a'_1|$ yields that $R_n^\circ(\mathcal{A}_1) \le R_n^\circ(\mathcal{A})$. This completes the proof of $R_n^\circ(\varphi \circ \mathcal{A}) \le MR_n^\circ(\mathcal{A})$.

We shall show how the contraction principle for R_n follows from the contraction principle for R_n^o. Let $\mathbf{0}$ denote the zero vector in $\mathbb{R}^n$. We shall also use the fact that $R_n^\circ(\mathcal{A} \cup \mathcal{B}) \le R_n^\circ(\mathcal{A}) + R_n^\circ(\mathcal{B})$ if $\mathbf{0} \in \mathcal{A}$ and $\mathbf{0} \in \mathcal{B}$. We find

$$\begin{aligned}
R_n(\varphi \circ \mathcal{A}) &= R_n^\circ((\varphi \circ \mathcal{A}) \cup (-\varphi \circ \mathcal{A}) \cup \{\mathbf{0}\}) \\
&\le R_n^\circ((\varphi \circ \mathcal{A}) \cup \{\mathbf{0}\}) + R_n^\circ((-\varphi \circ \mathcal{A}) \cup \{\mathbf{0}\}) \\
&= R_n^\circ(\varphi \circ (\mathcal{A} \cup \{\mathbf{0}\})) + R_n^\circ(-\varphi \circ (\mathcal{A} \cup \{\mathbf{0}\})) \\
&= 2R_n^\circ(\varphi \circ (\mathcal{A} \cup \{\mathbf{0}\})) \\
&\le 2MR_n^\circ(\mathcal{A} \cup \{\mathbf{0}\}) \le 2MR_n(\mathcal{A}).
\end{aligned}$$

□

6.4. Spoiler alert: A peek into the next two chapters

We will start exploring the implications of the finite class lemma, Lemma 6.1, more fully in the next two chapters, but we can give a brief preview here. Consider a learning problem of the type described in Section 6.1 in the special case when $\mathcal{F}$ consists of *binary-valued* functions on Z, i.e., $\mathcal{F}(\mathsf{Z}) = \{f(z) : z \in \mathsf{Z}, f \in \mathcal{F}\} \subseteq \{0, 1\}$. From Theorem 6.1, we know that

$$\mathbf{E}\Delta_n(Z^n) \le 2\mathbf{E}R_n(\mathcal{F}(Z^n)), \qquad (6.29)$$

where

$$\mathcal{F}(Z^n) := \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\}. \qquad (6.30)$$

Note that because each f can take values 0 or 1, $\mathcal{F}(Z^n) \subseteq \{0,1\}^n$. Moreover, since for any $Z^n \in \mathsf{Z}^n$ and any $f \in \mathcal{F}$ we have

$$\sqrt{\sum_{i=1}^{n} |f(Z_i)|^2} \le \sqrt{n}, \tag{6.31}$$

the set $\mathcal{F}(Z^n)$ for a *fixed* Z^n satisfies the conditions of the finite class lemma with $N = |\mathcal{F}(Z^n)| \le 2^n$ and $L = \sqrt{n}$. Hence,

$$R_n(\mathcal{F}(Z^n)) \le 2\sqrt{\frac{\log|\mathcal{F}(Z^n)|}{n}}. \tag{6.32}$$

In general, since $\log|\mathcal{F}(Z^n)| \le n$, the bound just says that $R_n(\mathcal{F}(Z^n)) \le 2$, which is not that useful. However, as we will see in the next two chapters, for a broad range of binary function classes $\mathcal{F}$ it will not be possible to pick out every single element in $\{0,1\}^n$ by taking the random "slices" $\mathcal{F}(Z^n)$, provided n is sufficiently large. To make these notions precise, let us define the quantity

$$\mathbb{S}_n(\mathcal{F}) := \sup_{z^n \in \mathsf{Z}^n} |\mathcal{F}(z^n)|, \tag{6.33}$$

which is called the *nth shatter coefficient* of $\mathcal{F}$. Then we have the bound

$$R_n(\mathcal{F}(Z^n)) \le 2\sqrt{\frac{\log \mathbb{S}_n(\mathcal{F})}{n}}. \tag{6.34}$$

Next, let

$$V(\mathcal{F}) := \max\left\{n \in \mathbb{N} : \mathbb{S}_n(\mathcal{F}) = 2^n\right\}. \tag{6.35}$$

This number is the famous *Vapnik–Chervonenkis* (or *VC*) *dimension* of $\mathcal{F}$, which has originated in their work [**VC71**]. It is clear that if $\mathbb{S}_n(\mathcal{F}) < 2^n$ for some n, then $\mathbb{S}_m(\mathcal{F}) < 2^m$ for all $m > n$. Hence, $V(\mathcal{F})$ is always well-defined (though it may be infinite). When it is finite, we say that $\mathcal{F}$ is a *VC class*. What this means is that, for n large enough, a certain structure emerges in the sets $\mathcal{F}(z^n)$, which prevents us from being able to form any combination of binary labels by sweeping through the entire $\mathcal{F}$. A fundamental result, which was independently derived by Sauer [**Sau72**] and Shelah [**She72**] in different contexts (combinatorics and mathematical logic respectively) and also appeared in a weaker form in the original work of Vapnik and Chervonenkis [**VC71**], says the following:

LEMMA 6.2 (Sauer–Shelah). *If $\mathcal{F}$ is a VC class, i.e., $V(\mathcal{F}) < \infty$, then*

$$\mathbb{S}_n(\mathcal{F}) \le \sum_{i=0}^{V(\mathcal{F})} \binom{n}{i} \le (n+1)^{V(\mathcal{F})}. \tag{6.36}$$

Thus, the finite class lemma and the Sauer-Shelah lemma combine to give the following important result, which we will revisit in the next two chapters:

THEOREM 6.2. *If $\mathcal{F}$ is a VC class of binary functions, then, with probability one,*

$$R_n(\mathcal{F}(Z^n)) \le 2\sqrt{\frac{V(\mathcal{F})\log(n+1)}{n}}. \tag{6.37}$$

Of course, the right hand side of (6.37) is therefore also an upper bound on $\mathbf{E}R_n(\mathcal{F}(Z^n))$, and such bound can be combined with Corollary 6.1. Consequently, for a VC class $\mathcal{F}$, the risk of ERM computed on an i.i.d. sample of size n from an arbitrary distribution $P \in \mathcal{P}(\mathsf{Z})$ is bounded by

$$P(\widehat{f}_n) \leq L^*(\mathcal{F}) + 8\sqrt{\frac{V(\mathcal{F})\log(n+1)}{n}} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}} \tag{6.38}$$

with probability at least $1-\delta$. In fact, using a much more refined technique called *chaining* originating in the work of Dudley [**Dud78**], it is possible to remove the logarithm in (6.37) to obtain the bound

$$R_n(\mathcal{F}(Z^n)) \leq C\sqrt{\frac{V(\mathcal{F})}{n}}, \tag{6.39}$$

where $C > 0$ is some universal constant independent of n and $\mathcal{F}$. We will not cover chaining in this class, but we will use the above formula.

To summarize, Figure 2 shows a diagram of the proof that if the set of classifiers has low VC dimension, then the ERM algorithm is PAC.

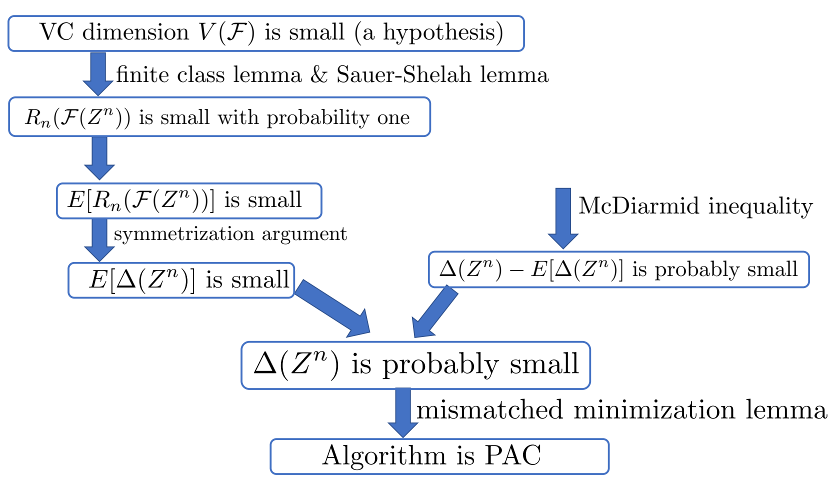

FIGURE 2. Path for proving Theorem 8.1, which, roughly speaking, states that the ERM algorithm is PAC for a model free learning problem $(\mathsf{X}, \mathcal{F}, \mathcal{P})$ such that the functions in $\mathcal{F}$ are binary valued, if the VC dimension of $\mathcal{F}$ is finite.

CHAPTER 7

Vapnik–Chervonenkis classes

A key result on the ERM algorithm, proved in the previous lecture, was that

$$P(\widehat{f}_n) \le L^*(\mathcal{F}) + 4\mathbf{E}R_n(\mathcal{F}(Z^n)) + \sqrt{\frac{2\log(1/\delta)}{n}}$$

with probability at least $1-\delta$. The quantity $R_n(\mathcal{F}(Z^n))$ appearing on the right-hand side of the above bound is the *Rademacher average* of the random set

$$\mathcal{F}(Z^n) = \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\},$$

often referred to as the *projection* of $\mathcal{F}$ onto the sample Z^n. From this we see that a sufficient condition for the ERM algorithm to produce near-optimal hypotheses with high probability is that the expected Rademacher average $\mathbf{E}R_n(\mathcal{F}(Z^n)) = \widetilde{O}(1/\sqrt{n})$, where the $\widetilde{O}(\cdot)$ notation indicates that the bound holds up to polylogarithmic factors in n, i.e., there exists some positive polynomial function $p(\cdot)$ such that

$$\mathbf{E}R_n(\mathcal{F}(Z^n)) \le O\left(\sqrt{\frac{p(\log n)}{n}}\right).$$

Hence, a lot of effort in statistical learning theory is devoted to obtaining tight bounds on $\mathbf{E}R_n(\mathcal{F}(Z^n))$.

One way to guarantee an $\widetilde{O}(1/\sqrt{n})$ bound on $\mathbf{E}R_n$ is if the effective size of the random set $\mathcal{F}(Z^n)$ is finite and grows polynomially with n. Then the finite class lemma will tell us that

$$R_n(\mathcal{F}(Z^n)) = O\left(\sqrt{\frac{\log n}{n}}\right).$$

In general, a reasonable notion of "effective size" is captured by various *covering numbers* (see, e.g., the lecture notes by Mendelson [**Men03**] or the recent monograph by Talagrand [**Tal14**] for detailed expositions of the relevant theory). In this chapter, we will look at a simple combinatorial notion of effective size for classes of *binary-valued* functions. This particular notion has originated with the work of Vapnik and Chervonenkis [**VC71**], and was historically the first such notion to be introduced into statistical learning theory. It is now known as the *Vapnik–Chervonenkis* (or *VC*) *dimension*.

7.1. Vapnik–Chervonenkis dimension: definition

DEFINITION 7.1. *Let $\mathcal{C}$ be a class of (measurable) subsets of some space Z. We say that a finite set $S = \{z_1, \ldots, z_n\} \subset \mathsf{Z}$ is* shattered *by $\mathcal{C}$ if for every subset $S' \subseteq S$ there exists some $C \in \mathcal{C}$ such that $S' = S \cap C$.*

In other words, $S = \{z_1, \ldots, z_n\}$ is shattered by $\mathcal{C}$ if, for any binary n-tuple $b = (b_1, \ldots, b_n) \in \{0,1\}^n$, there exists some $C \in \mathcal{C}$ such that

$$\big(\mathbf{1}_{\{z_1 \in C\}}, \ldots, \mathbf{1}_{\{z_n \in C\}}\big) = b$$

or, equivalently, if

$$\Big\{\big(\mathbf{1}_{\{z_1 \in C\}}, \ldots, \mathbf{1}_{\{z_n \in C\}}\big) : C \in \mathcal{C}\Big\} = \{0,1\}^n,$$

where we consider any two $C_1, C_2 \in \mathcal{C}$ as equivalent if $\mathbf{1}_{\{z_i \in C_1\}} = \mathbf{1}_{\{z_i \in C_2\}}$ for all $1 \le i \le n$.

Definition 7.2. *The* Vapnik–Chervonenkis dimension *(or the* VC dimension*) of $\mathcal{C}$ is*

$$V(\mathcal{C}) := \sup\Big\{|S| : S \textit{ is shattered by } \mathcal{C}\Big\}.$$

If $V(\mathcal{C}) < \infty$, we say that $\mathcal{C}$ is a VC class *(of sets).*

We can express the VC dimension in terms of *shatter coefficients* of $\mathcal{C}$: Let

$$\mathbb{S}_n(\mathcal{C}) := \sup_{S \subset \mathsf{Z}, |S|=n} |\{S \cap C : C \in \mathcal{C}\}|$$

denote the nth *shatter coefficient* of $\mathcal{C}$, where for each fixed S we consider any two $C_1, C_2 \in \mathcal{C}$ as equivalent if $S \cap C_1 = S \cap C_2$. Then $\mathcal{C}$ shatters a set $S \subset \mathsf{Z}$ with $|S| = n$ if and only if $|\{S \cap C : C \in \mathcal{C}\}| = 2^n$. Therefore, $V(\mathcal{C})$ is also given by:

$$V(\mathcal{C}) = \sup\Big\{n \in \mathbb{N} : \mathbb{S}_n(\mathcal{C}) = 2^n\Big\}.$$

The VC dimension $V(\mathcal{C})$ may be infinite, but it is always well-defined. The following lemma shows that $V(\mathcal{C}) = \inf\Big\{n \in \mathbb{N} : \mathbb{S}_{n+1}(\mathcal{C}) < 2^{n+1}\Big\}$.

Lemma 7.1. *If $\mathbb{S}_n(\mathcal{C}) < 2^n$, then $\mathbb{S}_m(\mathcal{C}) < 2^m$ for all $m > n$.*

Proof. It suffices to prove the contrapositive, namely, that if $m > n$ and $\mathbb{S}_m(\mathcal{C}) = 2^m$, then $\mathbb{S}_n(\mathcal{C}) = 2^n$. So suppose $m > n$ and $\mathbb{S}_m(\mathcal{C}) = 2^m$. By the assumption $\mathbb{S}_m(\mathcal{C}) = 2^m$, there exists $S = \{z_1, \ldots, z_m\} \subset \mathsf{Z}$, such that for every binary n-tuple $b = (b_1, \ldots, b_n)$ we can find some $C \in \mathcal{C}$ satisfying

$$\big(\mathbf{1}_{\{z_1 \in C\}}, \ldots, \mathbf{1}_{\{z_n \in C\}}, \mathbf{1}_{\{z_{n+1} \in C\}}, \ldots, \mathbf{1}_{\{z_m \in C\}}\big) = (b_1, \ldots, b_n, 0, \ldots, 0). \tag{7.1}$$

From (7.1) it immediately follows that

$$\big(\mathbf{1}_{\{z_1 \in C\}}, \ldots, \mathbf{1}_{\{z_n \in C\}}\big) = (b_1, \ldots, b_n). \tag{7.2}$$

Since $b = (b_1, \ldots, b_n)$ was arbitrary, we see from (7.2) that $\mathbb{S}_n(\mathcal{C}) = 2^n$. $\square$

There is a one-to-one correspondence between binary-valued functions $f : \mathsf{Z} \to \{0,1\}$ and subsets of Z:

$$\forall f : \mathsf{Z} \to \{0,1\} \text{ let } C_f := \{z : f(z) = 1\}$$
$$\forall C \subseteq \mathsf{Z} \text{ let } f_C := \mathbf{1}_{\{C\}}.$$

Thus, we can extend the concept of shattering, as well as the definition of the VC dimension, to any class $\mathcal{F}$ of functions $f : \mathsf{Z} \to \{0,1\}$:

Definition 7.3. *Let $\mathcal{F}$ be a class of functions $f : \mathsf{Z} \to \{0,1\}$, or let $\mathcal{F}$ be a class of functions $f : \mathsf{Z} \to \{-1,1\}$. We say that a finite set $S = \{z_1, \dots, z_n\} \subset \mathsf{Z}$ is shattered by $\mathcal{F}$ if it is shattered by the class*

$$\mathcal{C}_{\mathcal{F}} := \{C_f : f \in \mathcal{F}\},$$

where $C_f := \{z \in \mathsf{Z} : f(z) = 1\}$. The nth shatter coefficient of $\mathcal{F}$ is $\mathbb{S}_n(\mathcal{F}) = \mathbb{S}_n(\mathcal{C}_{\mathcal{F}})$, and the VC dimension of $\mathcal{F}$ is defined as $V(\mathcal{F}) = V(\mathcal{C}_{\mathcal{F}})$.

In light of these definitions, we can equivalently speak of the VC dimension of a class of sets or a class of binary-valued functions.

7.2. Examples of Vapnik–Chervonenkis classes

7.2.1. Semi-infinite intervals. Let $\mathsf{Z} = \mathbb{R}$ and take $\mathcal{C}$ to be the class of all intervals of the form $(-\infty, t]$ as t varies over $\mathbb{R}$. We will prove that $V(\mathcal{C}) = 1$. In view of Lemma 7.1, it suffices to show that (1) any one-point set $S = \{a\}$ is shattered by $\mathcal{C}$, and (2) no two-point set $S = \{a, b\}$ is shattered by $\mathcal{C}$.

Given $S = \{a\}$, choose any $t_1 < a$ and $t_2 > a$. Then $(-\infty, t_1] \cap S = \varnothing$ and $(-\infty, t_2] \cap S = S$. Thus, S is shattered by $\mathcal{C}$. This holds for every one-point set S, and therefore we have proved (1). To prove (2), let $S = \{a, b\}$ and suppose, without loss of generality, that $a < b$. Then there exists no $t \in \mathbb{R}$ such that $(-\infty, t] \cap S = \{b\}$. This follows from the fact that if $b \in (-\infty, t] \cap S$, then $t \geq b$. Since $b > a$, we must have $t > a$, so that $a \in (-\infty, t] \cap S$ as well. Since a and b are arbitrary, we see that no two-point subset of $\mathbb{R}$ can be shattered by $\mathcal{C}$.

7.2.2. Closed intervals. Again, let $\mathsf{Z} = \mathbb{R}$ and take $\mathcal{C}$ to be the class of all intervals of the form $[s, t]$ for all $s, t \in \mathbb{R}$. Then $V(\mathcal{C}) = 2$. To see this, we will show that (1) any two point set $S = \{a, b\}$ can be shattered by $\mathcal{C}$ and that (2) no three-point set $S = \{a, b, c\}$ can be shattered by $\mathcal{C}$.

For (1), let $S = \{a, b\}$ and suppose, without loss of generality, that $a < b$. Choose four points $t_1, t_2, t_3, t_4 \in \mathbb{R}$ such that $t_1 < t_2 < a < t_3 < b < t_4$. There are four subsets of S: $\varnothing$, $\{a\}$, $\{b\}$, and $\{a, b\} = S$. Then

$$[t_1, t_2] \cap S = \varnothing, \quad [t_2, t_3] \cap S = \{a\}, \quad [t_3, t_4] \cap S = \{b\}, \quad [t_1, t_4] \cap S = S.$$

Hence, S is shattered by $\mathcal{C}$. This holds for every two-point set in $\mathbb{R}$, which proves (1). To prove (2), let $S = \{a, b, c\}$ be an arbitrary three-point set with $a < b < c$. Then the intersection of any $[t_1, t_2] \in \mathcal{C}$ with S containing a and c must necessarily contain b as well. This shows that no three-point set can be shattered by $\mathcal{C}$, so by Lemma 7.1 we conclude that $V(\mathcal{C}) = 2$.

7.2.3. Closed halfspaces. Let $\mathsf{Z} = \mathbb{R}^2$, and let $\mathcal{C}$ consist of all closed halfspaces, i.e., sets of the form

$$\{z = (z_1, z_2) \in \mathbb{R}^2 : w_1 z_1 + w_2 z_2 \geq b\}$$

for all choices of $w_1, w_2, b \in \mathbb{R}$ such that $(w_1, w_2) \neq (0, 0)$. Then $V(\mathcal{C}) = 3$.

To see that $\mathbb{S}_3(\mathcal{C}) = 2^3 = 8$, it suffices to consider any set $S = \{z_1, z_2, z_3\}$ of three *non-collinear* points. Then it is not hard to see that for any $S' \subseteq S$ it is possible to choose a closed halfspace $C \in \mathcal{C}$ that would contain S', but not S. To see that $\mathbb{S}_4(\mathcal{C}) < 2^4$, we must look at all four-point sets $S = \{z_1, z_2, z_3, z_4\}$. There are two cases to consider:

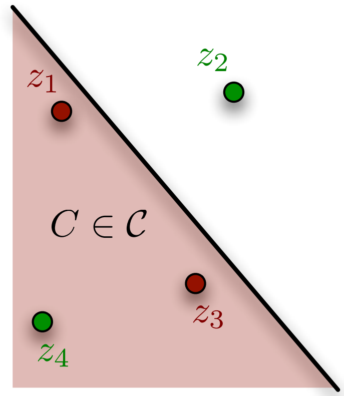

FIGURE 1. Impossibility of shattering an affinely independent four-point set in $\mathbb{R}^2$ by closed halfspaces.

(1) One point in S lies in the convex hull of the other three. Without loss of generality, let's suppose that $z_1 \in \mathrm{conv}(S')$ with $S' = \{z_2, z_3, z_4\}$. Then there is no closed halfspace C, such that $C \cap S = S'$. The reason for this is that every such halfspace C is a convex set. Hence, if $S' \subset C$, then any point in $\mathrm{conv}(S')$ is contained in C as well.

(2) No point in S is in the convex hull of the remaining points. This case, when S is an *affinely independent set*, is shown in Figure 1. Let us partition S into two disjoint subsets, S_1 and S_2, each consisting of "opposite" points. In the figure, $S_1 = \{z_1, z_3\}$ and $S_2 = \{z_2, z_4\}$. Then it is easy to see that there is no halfspace $\mathcal{C}$ whose boundary could separate S_1 from its complement S_2. This is, in fact, the (in)famous "XOR counterexample" of Minsky and Papert [**MP69**], which has demonstrated the impossibility of universal concept learning by one-layer perceptrons.

Since any four-point set in $\mathbb{R}^2$ falls under one of these two cases, we have shown that no such set can be shattered by $\mathcal{C}$. Hence, $V(\mathcal{C}) = 3$.

More generally, if $\mathsf{Z} = \mathbb{R}^d$ and $\mathcal{C}$ is the class of all closed halfspaces

$$\left\{ z \in \mathbb{R}^d : \sum_{j=1}^{d} w_j z_j \geq b \right\}$$

for all $w = (w_1, \dots, w_d) \in \mathbb{R}^d$ such that at least one of the w_j's is nonzero and all $b \in \mathbb{R}$, then $V(\mathcal{C}) = d + 1$ [**WD81**]; we will see a proof of this fact shortly.

7.2.4. Axis-parallel rectangles. Let $\mathsf{Z} = \mathbb{R}^2$, and let $\mathcal{C}$ consist of all "axis-parallel" rectangles, i.e., sets of the form $C = [a_1, b_1] \times [a_2, b_2]$ for all $a_1, b_1, a_2, b_2 \in \mathbb{R}$. Then $V(\mathcal{C}) = 4$.

First we exhibit a four-point set $S = \{z_1, z_2, z_3, z_4\}$ that is shattered by $\mathcal{C}$. It suffices to take $z_1 = (-1, 0)$, $z_2 = (1, 0)$, $z_3 = (0, -1)$, $z_4 = (0, 1)$. To show that no five-point set is shattered by $\mathcal{C}$, consider an arbitrary $S = \{z_1, z_2, z_3, z_4, z_5\}$. Of these, pick any one point with the smallest first coordinate and any one point with the largest first coordinate, and likewise for the second coordinate (refer to Figure 2), for a total of at most four. Let S' denote the set consisting of these points; in Figure 2, $S' = \{z_1, z_2, z_3, z_4\}$. Then it is easy

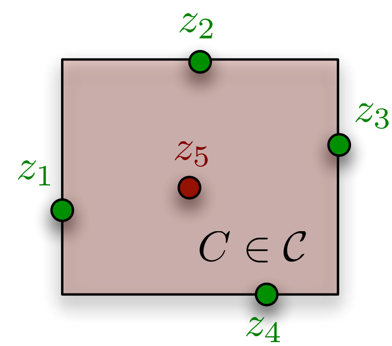

FIGURE 2. Impossibility of shattering a five-point set by axis-parallel rectangles.

to see that any $C \in \mathcal{C}$ that contains the points in S' must contain all the points in $S \backslash S'$ as well. Hence, no five-point set in $\mathbb{R}^2$ can be shattered by $\mathcal{C}$, so $V(\mathcal{C}) = 4$.

The same argument also works for axis-parallel rectangles in $\mathbb{R}^d$, i.e., all sets of the form $C = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_d, b_d]$, leading to the conclusion that the VC dimension of the set of all axis-parallel rectangles in $\mathbb{R}^d$ is equal to $2d$.

7.2.5. Sets determined by finite-dimensional function spaces – Dudley classes. The following result is due to Dudley [**Dud78**]. Let Z be an arbitrary set, and let $\mathcal{G}$ be an m-dimensional linear space of functions $g : \mathsf{Z} \to \mathbb{R}$, which means there exist m linearly independent functions $\psi_1, \ldots, \psi_m \in \mathcal{G}$ such that each $g \in \mathcal{G}$ has a (unique) representation of the form

$$g = \sum_{j=1}^{m} c_j \psi_j,$$

for some coefficients $c_1, \ldots, c_m \in \mathbb{R}$. Let h be an arbitrary function on Z, not necessarily in $\mathcal{G}$, and let $\mathcal{G} + h = \{g + h : g \in \mathcal{G}\}$. Consider the class of classifiers

$$\mathsf{pos}(\mathcal{G} + h) := \Big\{ \mathsf{pos}(g + h) : g \in \mathcal{G} \Big\},$$

where $\mathsf{pos}(g + h) := \{z \in \mathsf{Z} : g(z) + h(z) \geq 0\}$.

PROPOSITION 7.1 (VC Dimension of Dudley classes). $V(\mathsf{pos}(\mathcal{G} + h)) = m$.

PROOF. (Proof that $V(\mathsf{pos}(\mathcal{G} + h)) \geq m$.) Since the m functions $\psi_1, \ldots, \psi_m$ are linearly independent, there exist $z_1, \ldots, z_m \in \mathsf{Z}$ such that the $m \times m$ matrix $(\psi_j(z_i))_{1 \leq i,j \leq m}$ is full rank. That is because the points z_i can be selected greedily such that the rank of the first i rows of the matrix is i for $1 \leq i \leq m$. We shall show that $\mathsf{pos}(\mathcal{G} + h)$ shatters $\{z_1, \ldots, z_m\}$. To that end, fix an arbitrary $b \in \{-1, 1\}^m$.[1] Since the square matrix $(\psi_j(z_i))_{1 \leq i,j \leq m}$ has full rank, there exists an m vector $c^b = (c_1^b, \ldots, c_m^b)$ such that $b_i = \sum_{j=1}^m \psi_j(z_i) c_j^b + h(z_i)$ for $i \in [m]$. Therefore, letting $\psi^b = \sum \psi_j c_j^b$, we note that $\psi^b \in \mathcal{G}$ and $\psi^b(z_i) + h(z_i) = b_i$ for $1 \leq i \leq m$, so $z_i \in \mathsf{pos}(\psi^b + h)$ if and only if $b_i = 1$, for $1 \leq i \leq m$. Therefore, $\mathsf{pos}(\mathcal{G} + h)$ shatters $\{z_1, \ldots, z_m\}$, which completes the proof that $V(\mathsf{pos}(\mathcal{G} + h)) \geq m$.

(Proof that $V(\mathsf{pos}(\mathcal{G} + h)) \leq m$.) To prove this, we need to show that no set of $m + 1$ points in Z can be shattered by $\mathsf{pos}(\mathcal{G} + h)$. For the sake of argument by contradiction, fix

[1] Note that here we are using the two-point set $\{-1, 1\}$ rather than $\{0, 1\}$. Why?

$m+1$ arbitrary points $z_1, \ldots, z_{m+1} \in \mathsf{Z}$ and suppose that $\mathsf{pos}(\mathcal{G}+h)$ shatters $\{z_1, \ldots, z_{m+1}\}$. Let

$$\mathcal{G}|_{z^{m+1}} = \{(g(z_1), \ldots, g(z_{m+1})) : g \in \mathcal{G}\}.$$

Note that $\mathcal{G}|_{z^{m+1}}$ is a linear subspace of $\mathbb{R}^{m+1}$ due to the fact that $\mathcal{G}$ is a linear space, and the dimension of $\mathcal{G}|_{z^{m+1}}$ is less than or equal to m, because the rank of $\mathcal{G}$ is m. Therefore, there exists some nonzero vector $v = (v_1, \ldots, v_{m+1}) \in \mathbb{R}^{m+1}$ orthogonal to $\mathcal{G}|_{z^{m+1}}$, i.e., for every $g \in \mathcal{G}$

$$v_1 g(z_1) + \ldots + v_{m+1} g(z_{m+1}) = 0. \tag{7.3}$$

Consider two cases:
(a) If $v_1 h(z_1) + \ldots + v_{m+1} h(z_{m+1}) = 0$ (for example if $h \equiv 0$), it can be assumed that $v_i < 0$ for some i, because if not, the vector v could be replaced by $-v$. Let b be defined by $b_i = \mathbf{1}_{\{v_i \geq 0\}}$. Then for some $g^b \in \mathcal{G}$,

$$\left(\mathbf{1}_{\{g^b(z_1)+h(z_1)\geq 0\}}, \ldots, \mathbf{1}_{\{g^b(z_{m+1})+h(z_{m+1})\geq 0\}}\right) = b$$

and $\sum_i v_i(g^b(z_i) + h(z_i)) = 0$. But each term of this sum is nonnegative and the i^{th} term with $v_i < 0$ is strictly positive, contradicting the sum equal to zero.
(b) If case (a) does not hold, then by replacing v by $-v$ if necessary, it can be assumed that $v_1 h(z_1) + \ldots + v_{m+1} h(z_{m+1}) < 0$. Let b be defined by $b_i = \mathbf{1}_{\{v_i \geq 0\}}$. Then for some $g^b \in \mathcal{G}$,

$$\left(\mathbf{1}_{\{g^b(z_1)+h(z_1)\geq 0\}}, \ldots, \mathbf{1}_{\{g^b(z_{m+1})+h(z_{m+1})\geq 0\}}\right) = b$$

and $\sum_i v_i(g^b(z_i) + h(z_i)) = \sum_i v_i h(z_i) < 0$, but each term in the first sum is greater than or equal to zero, a contradiction. So in either case we reach a contradiction, so $V(\mathsf{pos}(\mathcal{G}+h)) \leq m$. $\square$

This result can be used to bound the VC dimension of many classes of sets:

- Let $\mathcal{C}$ be the class of all closed halfspaces in $\mathbb{R}^d$. Then any $C \in \mathcal{C}$ can be represented in the form $C = \{z : g(z) \geq 0\}$ for $g(z) = \langle w, z\rangle - b$ with some nonzero $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. The set $\mathcal{G}$ of all such affine functions on $\mathbb{R}^d$ is a linear space of dimension $d+1$, so by the above result we have $V(\mathcal{C}) = d+1$.
- Let $\mathcal{C}$ be the class of all closed balls in $\mathbb{R}^d$, i.e., sets of the form

  $$C = \left\{z \in \mathbb{R}^d : \|z - x\|^2 \leq r^2\right\}$$

  where $x \in \mathbb{R}^d$ is the *center* of C and $r \in \mathbb{R}^+$ is its *radius*. The concept class corresponds to the class of binary valued functions $\mathsf{pos}(\mathcal{G}_o)$, where $\mathcal{G}_o$ is the set of functions $g : \mathbb{R}^d \to \mathbb{R}$ of the form:

$$g(z) = r^2 - \|z - x\|^2 = r^2 - \sum_{j=1}^{d} |z_j - x_j|^2. \tag{7.4}$$

Expanding the second expression for g in (7.4), we get

$$\begin{aligned} g(z) &= 2\sum_{j=1}^{d} x_j z_j + r^2 - \sum_{j=1}^{d} x_j^2 - \sum_{j=1}^{d} z_j^2 \\ &= \sum_{j=1}^{d+1} c_j \psi_j(z) + h(z) \end{aligned}$$

where $c_j = x_j$ and $\psi_j(z) = 2z_j$ for $j \in [d]$, $c_{d+1} = r^2 - \sum_{j=1}^d x_j^2$, $\psi_{d+1}(z) = 1$, and $h(z) = -\sum_{j=1}^d z_j^2$. Note that $c_{d+1} = r^2 - \sum_{j=1}^d c_j^2 \geq -\sum_{j=1}^d c_j^2$. Thus, the constants $c_1, \dots, c_{d+1}$ are not quite arbitrary because they satisfy the constraint $c_{d+1} \geq -\sum_{j=1}^d c_j^2$. Let $\mathcal{G}$ denote the linear span of $\psi_1, \dots, \psi_{d+1}$. Then $\mathcal{G}_o \subset \mathcal{G} + h$ and $\mathcal{G}$ has dimension $d+1$. Therefore, $V(\mathcal{C}) = V(\mathsf{pos}(\mathcal{G}_o)) \leq V(\mathsf{pos}(\mathcal{G}+h)) = d+1$.

Next we show that that $V(\mathcal{C}) \geq d+1$. By the first part of the proof of Proposition 7.1, there exists a set of $d+1$ points in $\mathbb{R}^d$ that is shattered by the set of closed half spaces in the strict sense that none of the points is on the boundary of any of the 2^{d+1} half spaces that shatter the points. Therefore, there exists a set of 2^{d+1} balls, each having a very large radius, that shatter the points as well. So $V(\mathcal{C}) \geq d+1$. In conclusion, we find $V(\mathcal{C}) = d+1$.

7.2.6. VC dimension vs. number of parameters. Looking back at all these examples, one may get the impression that the VC dimension of a set of binary-valued functions is just the number of parameters. This is not the case. Consider the following one-parameter family of functions:

$$g_\theta(z) := \sin(\theta z), \qquad \theta \in \mathbb{R}.$$

However, the class of sets

$$\mathcal{C} = \Big\{ \{z \in \mathbb{R} : g_\theta(z) \geq 0\} : \theta \in \mathbb{R} \Big\}$$

has infinite VC dimension. Indeed, for any n, any collection of numbers $z_1, \dots, z_n \in \mathbb{R}$, and any binary string $b \in \{0,1\}^n$, one can always find some $\theta \in \mathbb{R}$ such that

$$\operatorname{sgn}(\sin(\theta z_i)) = \begin{cases} +1, & \text{if } b_i = 1 \\ -1, & \text{if } b_i = 0 \end{cases}.$$

7.3. Growth of shatter coefficients: the Sauer–Shelah lemma

The importance of VC classes in learning theory arises from the fact that, as n tends to infinity, the fraction of subsets of any $\{z_1, \dots, z_n\} \subset \mathsf{Z}$ that are shattered by a given VC class $\mathcal{C}$ tends to zero. We will prove this fact in this section by deriving a sharp bound on the shatter coefficients $\mathbb{S}_n(\mathcal{C})$ of a VC class $\mathcal{C}$. This bound has been (re)discovered at least three times, first in a weak form by Vapnik and Chervonenkis [**VC71**] in 1971, then independently and in different contexts by Sauer [**Sau72**] and Shelah [**She72**] in 1972. In strict accordance with Stigler's law of eponymy[2], it is known in the statistical learning literature as the *Sauer–Shelah lemma*.

[2] "No scientific discovery is named after its original discoverer" (http://en.wikipedia.org/wiki/Stigler's_law_of_eponymy)

Let us first set up some notation. Given integers $n, d \geq 1$, let $\binom{n}{\leq d}$ denote the number of subsets of a set of cardinality n with cardinality less than or equal to d. It follows that $\binom{n}{\leq d} = 2^n$ for $d \geq n$, and, in general,

$$\binom{n}{\leq d} = \sum_{i=0}^{d} \binom{n}{i},$$

with the convention that $\binom{n}{i} = 0$ for $i > n$.

LEMMA 7.2 (Sauer–Shelah lemma). *Let $\mathcal{C}$ be a class of subsets of some space Z with $V(\mathcal{C}) = d < \infty$. Then for all n,*

$$\mathbb{S}_n(\mathcal{C}) \leq \binom{n}{\leq d}. \tag{7.5}$$

Also, $\binom{n}{\leq d} \leq (n+1)^d$ and, for $n \geq d$, $\binom{n}{\leq d} \leq \left(\frac{ne}{d}\right)^d$.

Let $\mathcal{C}$ be a VC class of subsets of some space Z. The Sauer-Shelah lemma implies

$$\limsup_{n\to\infty} \frac{\mathbb{S}_n(\mathcal{C})}{2^n} \leq \lim_{n\to\infty} \frac{(n+1)^{V(\mathcal{C})}}{2^n} = 0.$$

In other words, as n becomes large, the fraction of subsets of an arbitrary n-element set $\{z_1, \ldots, z_n\} \subset \mathsf{Z}$ that are shattered by $\mathcal{C}$ becomes negligible. Moreover, combining the bounds of the Sauer–Shelah lemma with the finite class lemma for Rademacher averages, Lemma 6.1 (with $L = \sqrt{n}$ and $N = (n+1)^{V(\mathcal{F})}$), we get the following:

THEOREM 7.1. *Let Z be an arbitrary set and let $\mathcal{F}$ be a class of binary-valued functions $f : \mathsf{Z} \to \{0, 1\}$, or a class of functions $f : \mathsf{Z} \to \{-1, 1\}$. Let Z^n be an i.i.d. sample of size n drawn according to an arbitrary probability distribution $P \in \mathcal{P}(\mathsf{Z})$. Then, with probability one,*

$$R_n(\mathcal{F}(Z^n)) \leq 2\sqrt{\frac{V(\mathcal{F})\log(n+1)}{n}}.$$

A more refined *chaining technique* [**Dud78**] can be used to remove the logarithm in the above bound:

THEOREM 7.2. *There exists an absolute constant $C > 0$, such that under the conditions of the preceding theorem, with probability one,*

$$R_n(\mathcal{F}(Z^n)) \leq C\sqrt{\frac{V(\mathcal{F})}{n}}.$$

7.3.1. Proof of the Sauer-Shelah lemma. Lemma 7.2 is proved in this section, based on the technique of shifting [**Fra87, Fra91**]. We first write the definition of VC dimension in case the base set consists of $[n] = \{1, \ldots, n\}$, and subsets of $[n]$ are represented as binary vectors. Let $\mathcal{U} \subset \{0, 1\}^n$. Given $A \subset [n]$, if $b \in \{0, 1\}^n$ let $\pi_A(b) \triangleq (b_i : i \in A)$, called the *restriction* of b to A, and let $\pi_A(S) = \{\pi_A(b) : b \in S\}$. The *support* of a binary vector b is $\{i : b_i = 1\}$. A nonempty set $A \subset [n]$ is said to be shattered by $\mathcal{U}$ if $\pi_A(\mathcal{U})$ contains all $2^{|A|}$ possible sequences. The VC dimension of $\mathcal{U}$, denoted by $V(\mathcal{U})$, is the maximum d such that there exists a subset of $[n]$ of cardinality d that is shattered by $\mathcal{U}$.

Lemma 7.3 (Sauer-Shelah lemma, simple setting). *Suppose* $\mathcal{U} \subset \{0,1\}^n$ *with* $V(\mathcal{U}) = d$. *Then* $|\mathcal{U}| \leq \binom{n}{\leq d}$. *Also,* $\binom{n}{\leq d} \leq (n+1)^d$ *and, for* $n \geq d$, $\binom{n}{\leq d} \leq \left(\frac{ne}{d}\right)^d$.

Lemma 7.2 follows by applying Lemma 7.3 to $\mathcal{U} = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}$ for arbitrary $z_1, \ldots, z_n \in \mathsf{Z}$, because for any such $\mathcal{U}$, $V(\mathcal{U}) \leq V(\mathcal{F})$.

Proof of Lemma 7.3. It is shown below that there exists $\mathcal{V} \subset \{0,1\}^n$ satisfying the following three properties:

(1) $|\mathcal{V}| = |\mathcal{U}|$
(2) For any $A \subset [n]$, if A is shattered by $\mathcal{V}$ then A is shattered by $\mathcal{U}$
(3) $\mathcal{V}$ is downward closed. By definition, this means, for any $b, b' \in \{0,1\}^n$ such that $b' \leq b$ (bitwise) and $b \in \mathcal{V}$, it holds that $b' \in \mathcal{V}$.

This will complete the proof, because the third property implies that $\mathcal{V}$ shatters the support of any vector in $\mathcal{V}$. Therefore by property 2, $\mathcal{U}$ shatters the support of any vector in $\mathcal{V}$. [3] Thus, any vector in $\mathcal{V}$ can have at most d nonzero bits. So $|\mathcal{U}| = |\mathcal{V}| \leq \binom{n}{\leq d}$. It remains to show there exists $\mathcal{V} \subset \{0,1\}^n$ satisfying the three properties above. For that we refer to the shifting algorithm described in pseudocode as Algorithm 1, describing how $\mathcal{U}$ is transformed into $\mathcal{V}$. For $i \in [n]$, τ_i denotes the toggle function. It operates on binary vectors such that $\tau_i(b)$ is obtained from b by flipping the i^{th} bit of b from 0 to 1 or vice versa. The algorithm is illustrated in Figure 3.

Algorithm 1 Shifting algorithm

```
Input: U ⊂ {0,1}^n for some n ≥ 1
  1.  for i in [n] :
  2.      for b in U :
  3.          if b_i = 1 and τ_i(b) ∉ U :
  4.              replace b by τ_i(b)
  5.  repeat steps 1-4 until no further changes occur
  6.  return V = U
```

The algorithm terminates because the total number of 1's strictly decreases each time a change occurs in steps 1-4. Property 1 is true because $|\mathcal{U}|$ is never changed by the algorithm. Property 3 is true because otherwise more changes would have been possible in steps 1-4. It remains to check property 2. Consider the block of steps 2-4, executed for some $i \in [n]$. Let $\mathcal{U}$ denote the state just before execution of steps 2-4 and let $\mathcal{U}'$ denote the state just after execution of those steps. Suppose $A \subset [n]$ such that $\mathcal{U}'$ shatters A. It suffices to prove $\mathcal{U}$ also shatters $\mathcal{A}$.

The only changes to $\mathcal{U}$ made during the block of steps 2-4 is that the i^{th} bit of some vectors in $\mathcal{U}$ might be changed from 1 to 0. So if $i \notin A$ then $\pi_A(\mathcal{U}) = \pi_A(\mathcal{U}')$, so that $\mathcal{U}$ also shatters A.

[3] So the number of sets shattered by $\mathcal{U}$ is greater than or equal to the cardinality of $\mathcal{U}$, a result known as Pajor's lemma [**Paj85**].

$$\text{Initial } \mathcal{U}: \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad \text{After 2-4 for } i \in \{1,2\}: \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ \mathbf{0} & 1 & 0 & 0 & 0 \\ \mathbf{0} & 1 & 0 & 1 & 0 \\ 1 & \mathbf{0} & 1 & 1 & 0 \\ 1 & \mathbf{0} & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

FIGURE 3. Illustration of the shifting algorithm. The input set $\mathcal{U}$ is the set of rows of the matrix on the left. The matrix on the right represents $\mathcal{U}$ after steps 2-4 of Algorithm 1 have been applied for $i \in \{1,2\}$. Bits that were toggled are shown in bold font. The input $\mathcal{U}$ shatters all singleton sets, all pairs except for $\{1,2\}$, and the set $\{2,3,4\}$. After steps 2-4 have been applied for $i \in \{1,2\}$, $\mathcal{U}$ as shown on the right, shatters all singleton sets of columns, all pairs except $\{1,2\},\{1,3\}$ and $\{1,4\}$, and no triplets.

So suppose $i \in A$. Let b denote an arbitrary binary vector indexed by A. Since $\mathcal{U}'$ shatters A there is a vector $b' \in \mathcal{U}'$ such that $b = \pi_A(b')$. We need to show that $b = \pi_A(b'')$ for some $b'' \in \mathcal{U}$. If $b_i = 1$ then, since the algorithm only turns 1's into 0's, $b' \in \mathcal{U}$, so it suffices to take $b'' = b'$. If $b_i = 0$, then since $\mathcal{U}'$ shatters A there must be another vector $b''' \in \mathcal{U}'$ such that $\pi_A(b''') = \tau_i(b)$. Since $b_i''' = 1$ and $b''' \in \mathcal{U}'$, it must be that $b''' \in \mathcal{U}$. Since b''' was not modified by the block of steps 2-4, it must be that $\tau_i(b''') \in \mathcal{U}$. Therefore, it suffices to take $b'' = \tau_i(b''')$. Thus, $\mathcal{U}$ shatters A as claimed.

To prove the first upper bound on $\binom{n}{\le d}$, recall that $\binom{n}{\le d}$ is the number of subsets of n distinct objects with cardinality less than or equal to d. The bound is true because there are $(n+1)^d$ ways to make d draws with replacement from the set $\{0,1,\cdots,n\}$, and the set of nonzero numbers so drawn can be made to equal any subset of $\{1,\ldots,n\}$ with cardinality less than or equal to d. For example, if $d=5$ and $n=10$, the draws $(4,6,3,0,0)$ correspond to the set $\{3,4,6\}$ and the draws $(4,6,0,0,0)$ correspond to the set $\{4,6\}$.

The second upper bound on $\binom{n}{\le d}$ follows from:

$$\left(\frac{d}{n}\right)^d \binom{n}{\le d} = \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^d \le \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \overset{(a)}{=} \left(1+\frac{d}{n}\right)^n \overset{(b)}{\le} \mathrm{e}^d,$$

where (a) is due to the binomial theorem, and (b) follows from $1+x \le \mathrm{e}^x$. $\square$

CHAPTER 8

Binary classification

In the first section of this chapter, we apply the results of the previous two chapters about ERM, Rademacher averages and VC dimension, directly to the concept learning problem. This brings to fruition the sneak preview given in Section 6.4. Then in Section 8.2 we turn to the use of surrogate loss functions, which offers a way to regularize the empirical risk, yielding in some cases convex optimization problems, and a parameter to trade between approximation accuracy and generalization ability. Motivated by Section 8.2 the remaining sections focus on various classes of classifiers and bounds on their Rademacher averages, which can be used to give probabilistic performance guarantees.

8.1. The fundamental theorem of concept learning

A binary classification learning problem can be modeled as a triple $(\mathsf{X}, \mathcal{P}, \mathcal{C})$, where X is a general set, $\mathcal{P}$ (also denoted by $\mathcal{P}(\mathsf{Z})$) is a set of probability distributions over $\mathsf{Z} = \mathsf{X} \times \{0,1\}$, and $\mathcal{C}$ is a class of subsets of X, with the following interpretation. A fresh observation is modeled as a random couple $Z = (X, Y)$, where $X \in \mathsf{X}$ is called the *feature vector* and $Y \in \{0,1\}$ is called the *label* [1]. In the spirit of the model-free framework, we assume that the relationship between the features and the labels is stochastic and described by an unknown probability distribution $P \in \mathcal{P}(\mathsf{Z})$. As usual, we consider the case when we are given an i.i.d. sample of length n from P. The goal is to learn a concept, i.e., a set $\widehat{C} \subset \mathsf{X}$ such that the probability of classification error, $\mathbf{P}(\mathbf{1}_{\{X \in \widehat{C}\}} \neq Y)$, is small. As we have seen before, the optimal choice is the *Bayes classifier*, $C^*_{Bayes} = \{x : \eta(x) \geq 1/2\}$, where $\eta(x) := \mathbf{P}[Y = 1 | X = x]$ is the *regression function*. However, since we make no assumptions on P, in general we cannot hope to learn the Bayes classifier. Instead, we focus on a more realistic goal: We fix a collection $\mathcal{C}$ of concepts and then use the training data to come up with a hypothesis $\widehat{C}_n \in \mathcal{C}$, such that

$$\mathbf{P}(\mathbf{1}_{\{X \in \widehat{C}_n\}} \neq Y) \approx \inf_{C \in \mathcal{C}} \mathbf{P}(\mathbf{1}_{\{X \in C\}} \neq Y)$$

with high probability.

By way of notation, let us write $L_P(C)$ for the classification error of C, i.e., $L_P(C) := \mathbf{P}(\mathbf{1}_{\{X \in C\}} \neq Y)$, and let $L^*_P(\mathcal{C})$ denote the smallest classification error attainable over $\mathcal{C}$:

$$L^*_P(\mathcal{C}) := \inf_{C \in \mathcal{C}} L_P(C).$$

We will assume that a minimizing $C^* \in \mathcal{C}$ exists.

[1] In the next section we will switch from using $\{0,1\}$, which connects more directly to concept classification, to $\{-1,1\}$, which is more convenient for discussion of surrogate loss.

The empirical risk minimization (ERM) learning algorithm for this problem was defined in Chapter 5. Given the training data $Z^n = (Z_1, \ldots, Z_n)$, the ERM algorithm returns a concept $\widehat{C}_n$ such that

$$\widehat{C}_n \in \arg\max_{C \in \mathcal{C}} L_{P_n}(C),$$

where P_n is the empirical distribution of the training data. Chapter 5 showed that probabilistic performance guarantees for ERM can be given if uniform convergence of empirical means (UCEM) holds. Chapter 6 abstracts away the loss functions and considers a class of functions $\mathcal{F}$ defined on a space Z. A key result of the chapter, based on a symmetrization technique, is that the UCEM property can be established if there are suitable upper bounds on the average Rademacher complexity of $\mathcal{F}$. In turn, Chapter 7 gives upper bounds on Rademacher averages for classes of binary functions in terms of the VC dimension of those classes. To combine these results we trace backwards from the abstract formulation of Chapter 6 to the original problem.

For the concept learning problem with the usual 0-1 loss, each concept C corresponds to a function $\ell_C : \mathsf{Z} \to \{0,1\}$ defined by $\ell_C(z) = \mathbf{1}_{\{y \neq \mathbf{1}_{\{x \in C\}}\}}$. Let $\mathcal{F}_{\mathcal{C}} = \{\ell_C : C \in \mathcal{C}\}$. Thus, $\mathcal{F}_{\mathcal{C}}$ is a class of binary valued functions induced by $\mathcal{C}$ and the 0-1 loss function. The following lemma shows that the VC dimension of $\mathcal{C}$ is the same as the VC dimension of the induced class of functions.

Lemma 8.1. *Given a concept class $\mathcal{C}$ consisting of subsets of a base space X, let $\mathcal{F}_{\mathcal{C}} = \{\ell_C : C \in \mathcal{C}\}$ denote the set of binary valued functions on $\mathsf{X} \times \{0,1\}$ induced by the 0-1 loss. In other words, $\ell_C((x,y)) = \mathbf{1}_{\{y \neq \mathbf{1}_{\{x \in C\}}\}}$. Then $V(\mathcal{C}) = V(\mathcal{F}_{\mathcal{C}})$.*

Proof. To begin, we claim that for any $\{x_1, \ldots, x_n\} \subset \mathsf{X}$, $\mathcal{C}$ shatters $\{x_1, \ldots, x_n\}$ if and only if $\mathcal{F}_{\mathcal{C}}$ shatters $\{(x_1, 0), \ldots, (x_n, 0)\}$. For any $x \in \mathsf{X}$, if $z = (x, 0)$, $\ell_C(z) = \ell_C((x,0)) = \mathbf{1}_{\{0 \neq \mathbf{1}_{\{x \in C\}}\}} = \mathbf{1}_{\{x \in C\}}$. Thus, for any $b \in \{0,1\}^n$ and any $C \in \mathcal{C}$,

$$(\mathbf{1}_{\{x_1 \in C\}}, \ldots, \mathbf{1}_{\{x_n \in C\}}) = b \text{ if and only if } (\ell_C((x_1,0)), \ldots, \ell_C((x_n,0))) = b.$$

The claim follows.

The next step is to show $V(\mathcal{C}) \leq V(\mathcal{F}_{\mathcal{C}})$. Let $n = V(\mathcal{C})$ if $V(\mathcal{C}) < \infty$ and n be an arbitrary positive integer if $V(\mathcal{C}) = \infty$. Then there exists $\{x_1, \ldots, x_n\} \subset \mathsf{X}$ that is shattered by $\mathcal{C}$. Thus, by the claim just shown, $\{(x_1, 0), \ldots, (x_n, 0)\}$ is shattered by $\mathcal{F}_{\mathcal{C}}$. Therefore $n \leq V(\mathcal{F}_{\mathcal{C}})$. Thus, $V(\mathcal{C}) \leq V(\mathcal{F}_{\mathcal{C}})$.

The next claim is that $\mathcal{F}_{\mathcal{C}}$ does not shatter any two point set of the form $\{(x,0),(x,1)\}$. For any $x \in \mathsf{X}$ and any $C \in \mathcal{C}$, $\ell_C((x,0)) \neq \ell_C((x,1))$. Thus, there does not exist $C \in \mathcal{C}$ such that $(\ell_C((x,0)), \ell_C((x,1))) = (1,1)$. Therefore, $\mathcal{F}_{\mathcal{C}}$ does not shatter $\{(x,0),(x,1)\}$, as claimed.

It remains to show $V(\mathcal{C}) \geq V(\mathcal{F}_{\mathcal{C}})$. Let $n = V(\mathcal{F}_{\mathcal{C}})$ if $V(\mathcal{F}_{\mathcal{C}}) < \infty$ and n be an arbitrary positive integer if $V(\mathcal{F}_{\mathcal{C}}) = \infty$. Then there exists $\{z_1, \ldots, z_n\} \subset \mathsf{X} \times \{0,1\}$ that is shattered by $\mathcal{F}_{\mathcal{C}}$. For each i, $z_i = (x_i, b'_i)$ for some $b'_i \in \{0,1\}$. Since $\mathcal{F}_{\mathcal{C}}$ does not shatter any two point set of the form $\{(x,0),(x,1)\}$, the x_i's are distinct. Note that $\ell_C(z_i) = b'_i \oplus \ell_C((x_i, 0))$, where "$\oplus$" denotes modulo two addition. Therefore, $\{(x_1, 0), \ldots, (x_n, 0)\}$ is also shattered by $\mathcal{F}_{\mathcal{C}}$. Therefore, by the claim shown at the beginning of the proof, $\{x_1, \ldots, x_n\}$ is shattered by $\mathcal{C}$. Thus, $n \leq V(\mathcal{C})$, so that $V(\mathcal{C}) \geq V(\mathcal{F}_{\mathcal{C}})$. $\square$

Warning: In what follows, we will use C or c to denote various absolute constants; their values may change from line to line.

The bounds in Theorems 7.1 and 7.2 on $R_n(\mathcal{F}(Z^n))$ that hold with probability one are also bounds on $\mathbf{E}R_n(\mathcal{F}(Z^n))$, so either of them can be combined with Corollary 6.1. In particular, using the first part of Corollary 6.1 together with Lemma 8.1 implies the following bounds for ERM for concept learning. The thread of the proof we have given is pictured in Fig. 2.

THEOREM 8.1 (Performance bounds for concept learning by ERM). *Consider an agnostic concept learning problem* $(\mathsf{X}, \mathcal{P}, \mathcal{C})$, *and let* $\delta > 0$. *For any* $P \in \mathcal{P}$, *the ERM algorithm satisfies*

$$L_P(\widehat{C}_n) \le L_P^*(\mathcal{C}) + 8\sqrt{\frac{V(\mathcal{C})\log(n+1)}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{8.1}$$

with probability at least $1-\delta$. *There is a universal constant* C *so that for any probability distribution* P *on* Z *and* $\delta \in (0,1)$, *the ERM algorithm satisfies*

$$L_P(\widehat{C}_n) \le L_P^*(\mathcal{C}) + C\sqrt{\frac{V(\mathcal{C})}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{8.2}$$

with probability at least $1-\delta$.

Hence, a concept learning problem $(\mathsf{X}, \mathcal{P}, \mathcal{C})$ is PAC learnable if $V(\mathcal{C}) < +\infty$. The converse is also true; such results are sometimes called "no free lunch" theorems. We thus have the following corollary.

COROLLARY 8.1. *(Fundamental theorem of concept learning) A concept learning problem* $(\mathsf{X}, \mathcal{P}, \mathcal{C})$ *is PAC learnable if and only if* $V(\mathcal{C}) < +\infty$.

8.1.1. Linear and generalized linear discriminant rules. Generalized linear discriminant rules correspond to using a Dudley class of concepts, as described in Section 7.2.5, and repeated here. Let Z be an arbitrary set, and let $\mathcal{G}$ be an m-dimensional linear space of functions $g : \mathsf{Z} \to \mathbb{R}$, which means there exist m linearly independent functions $\psi_1, \dots, \psi_m \in \mathcal{G}$ such that each $g \in \mathcal{G}$ has a (unique) representation of the form

$$g = \sum_{j=1}^{m} c_j \psi_j,$$

for some coefficients $c_1, \dots, c_m \in \mathbb{R}$. Let h be an arbitrary function on Z, not necessarily in $\mathcal{G}$, and let $\mathcal{G} + h = \{g + h : g \in \mathcal{G}\}$. Consider the class of classifiers:

$$\mathsf{pos}(\mathcal{G} + h) := \Big\{\{z \in \mathsf{Z} : g(z) + h(z) \ge 0\} : g \in \mathcal{G}\Big\}.$$

By Proposition 7.1, the VC dimension of the class is m. Thus, Theorem 8.1 holds for the family of Dudley classifiers with $V(\mathcal{C}) = m$.

One of the simplest classification rules (and one of the first to be studied) is a *linear discriminant rule*. Linear discriminant rules are the half-space classifiers in $\mathsf{X} = \mathbb{R}^d$, obtained by thresholding linear functions. They correspond to concepts $C \in \mathcal{C}$ that can be represented in the form

$$C = \{x : \langle w, z\rangle \ge b\} = \{z : g(z) \ge 0\}$$

for $g(z) = \langle w, z\rangle - b$ with some nonzero $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. The set $\mathcal{G}$ of all such affine functions on $\mathbb{R}^d$ is a linear space of dimension $d+1$, so by the above result we have $V(\mathcal{C}) = d+1$. Thus, Theorem 8.1 holds for the family of half-space classifiers with $V(\mathcal{C}) = d+1$.

8.1.2. Two fundamental issues. As Theorem 8.1 shows, the ERM algorithm applied to the collection of all (generalized) linear discriminant rules is guaranteed to work well in the sense that the classification error of the output hypothesis will, with high probability, be close to the optimum achievable by any discriminant rule with the given structure. The same argument extends to any collection of concepts $\mathcal{C}$ with VC dimension much smaller than the sample size n. In other words, with high probability the difference $L_P(\widehat{C}) - L_P^*(\mathcal{C})$ will be small. However, precisely because the VC dimension of $\mathcal{G}$ cannot be too large, the approximation properties of $\mathcal{C}$ will be limited. Another problem is computational. For instance, the problem of finding an empirically optimal linear discriminant rule is NP-hard. In other words, unless P is equal to NP, there is no hope of coming up with an efficient ERM algorithm for linear discriminant rules that would work for all feature space dimensions d. If d is fixed, then it is possible to enumerate all projections of a given sample Z^n onto the class of indicators of all halfspaces in $O(n^{d-1}\log n)$ time, which allows for an exhaustive search for an ERM solution, but the usefulness of this naive approach is limited to $d < 5$.

8.2. Risk bounds for combined classifiers via surrogate loss functions

One way to sidestep the above approximation-theoretic and computational issues is to replace the 0–1 Hamming loss function that gives rise to the probability of error criterion with some other loss function. What we gain is the ability to bound the performance of various complicated classifiers built up by combining simpler *base classifiers* in terms of the complexity (e.g, the VC dimension) of the collection of the base classifiers, as well as considerable computational advantages, especially if the problem of minimizing the empirical surrogate loss turns out to be a convex programming problem. What we lose, though, is that, in general, we will not be able to compare the generalization error of the learned classifier to the minimum classification risk. Instead, we will have to be content with the fact that the generalization error will be close to the smallest *surrogate loss*.

For the remainder of this chapter, we shall assume that the labels take values ± 1*, and we will consider* ± 1 *valued classifiers* $g : \mathsf{X} \to \{-1, 1\}$ *rather than concepts* $C \subset \mathsf{X}$. Consider classifiers of the form

$$g_f(x) = \operatorname{sgn} f(x) \equiv \begin{cases} 1, & \text{if } f(x) \geq 0 \\ -1, & \text{otherwise} \end{cases} \tag{8.3}$$

where $f : \mathsf{X} \to \mathbb{R}$ is some function. The generalized linear discriminant rules considered in the previous section have such form, with the family of f's having finite linear dimension. As we shall see, other families of functions f can be used. For a joint distribution P of (X, Y), the risk, or probability of classification error, of using g_f satisfies

$$L(g_f) = \mathbf{P}(g_f(X) \neq Y) = \mathbf{P}(Y g_f(X) \leq 0) \overset{(a)}{\leq} \mathbf{P}(Y f(X) \leq 0) = \mathbf{E}[\mathbf{1}_{\{-Yf(X) \geq 0\}}]. \tag{8.4}$$

(The inequality (a) in (8.4) is sometimes strict because if $f(X) = 0$ and $Y = 1$, then $Yf(X) \leq 0$, even though there is no classification error, i.e. $Y g_f(X) = 1 > 0$.) From now on, when dealing with classifiers of the form (8.3), we write $L(f)$ instead of $L(g_f)$ to keep the notation simple.

The idea of surrogate loss function is to replace $\mathbf{1}_{\{x\geq 0\}}$ at the right end of (8.4) by a continuous, often convex, function that dominates it. That is, suppose $\varphi : \mathbb{R} \to \mathbb{R}_+$ is such that

(1) φ is continuous
(2) φ is nondecreasing
(3) $\varphi(x) \geq \mathbf{1}_{\{x\geq 0\}}$ for all $x \in \mathbb{R}$.

We call φ a *penalty function*, similar to use of the term for constrained optimization problems. The *surrogate loss function*, or φ-loss function, corresponding to penalty function φ is defined by $\ell_\varphi(y,u) \triangleq \varphi(-yu)$. Note that the surrogate loss function is greater than or equal to the original loss function derived from 0-1 loss: $\ell(y,u) \leq \ell_\varphi(y,u)$ for all $(y,u) \in \{-1,1\} \times \mathbb{R}$.

Table 1 displays some popular examples of penalty functions, along with their Lipschitz constants, and surrogate loss functions.

TABLE 1. Some popular penalty functions

Name	Penalty function $\varphi(x)$	M_φ	surrogate loss function $\ell_\varphi(y,u)$
exponential	e^x	–	e^{-yu}
logit	$\log_2(1+\mathrm{e}^x)$	$\frac{1}{\ln 2}$	$\log_2(1+\mathrm{e}^{-yu})$
hinge	$(1+x)_+$	1	$(1-yu)_+$
ramp	$\min\left\{1, \left(1+\frac{x}{\gamma}\right)_+\right\}$	$\frac{1}{\gamma}$	$\min\left\{1, \left(1-\frac{yu}{\gamma}\right)_+\right\}$

In order to avoid overuse of the letter "L," denote the φ*-risk* of f by

$$A_\varphi(f) := \mathbf{E}[\varphi(-Yf(X))]$$

and its empirical version

$$A_{\varphi,n}(f) := \frac{1}{n}\sum_{i=1}^{n} \varphi(-Y_i f(X_i)).$$

Since the surrogate loss is greater than or equal to the 0-1 loss, $L(f) \leq A_\varphi(f)$ and $L_n(f) \leq A_{\varphi,n}(f)$.

With these preliminaries out of the way, we can state and prove the basic surrogate loss bound, due to Koltchinskii and Panchenko [**KP02**]. We shall again appeal to the uniform approximation method to bound $A_\varphi(\widehat{f})$ (and hence $L(\widehat{f})$.) A key role is played by the maximum deviation of empirical (surrogate) risk from general (surrogate) risk, defined by

$$\Delta_n(Z^n) := \sup_{f\in\mathcal{F}} |A_\varphi(f) - A_{\varphi,n}(f)|\,.$$

LEMMA 8.2. *Consider a class $\mathcal{F}$ of functions from X into $\mathbb{R}$, and let φ be a penalty function such that:*

(i) For some $B > 0$, $\varphi(-yf(x)) \in [0, B]$ for all $(x, y) \in \mathsf{X} \times \{0, 1\}$ and all $f \in \mathcal{F}$.
(ii) φ is Lipschitz-continuous with constant M_φ: $|\varphi(u) - \varphi(v)| \leq M_\varphi |u - v|$ for $u, v \in \mathbb{R}$.

Then for any n and $t \geq 0$, with probability at least $1 - \mathrm{e}^{-2t^2}$,

$$\Delta_n(Z^n) \leq 4M_\varphi \mathbf{E} R_n(\mathcal{F}(X^n)) + \frac{Bt}{\sqrt{n}} \tag{8.5}$$

Proof. Note that $A_\varphi(f)$ and $A_{\varphi,n}(f)$ are the general and empirical averages, respectively, of the surrogate loss function $\ell_{\varphi,f}(x, y) = \varphi(-yf(x))$. For a reason that will become apparent, we'd prefer to have functions with value 0 when $f(x) = 0$, so we will work with functions of the form $\varphi(-yf(x)) - \varphi(0)$. This works out because subtraction of $\varphi(0)$ has the same effect on the generalization risk as on the empirical risk:

$$\begin{aligned} A_\varphi(f) - A_{\varphi,n}(f) &= P(\ell_{\varphi,f}) - P_n(\ell_{\varphi,f}) \\ &= P(\ell_{\varphi,f} - \varphi(0)) - P_n(\ell_{\varphi,f} - \varphi(0)) \end{aligned}$$

Thus, with $\mathcal{H}_\varphi$ being the class of functions on $\mathsf{Z} = \mathsf{X} \times \{0, 1\}$ of the form $\varphi(-yf(x)) - \varphi(0)$, for $f \in \mathcal{F}$, the familiar symmetrization argument from Section 6.2 yields

$$\mathbf{E}\Delta_n(Z^n) \leq 2\mathbf{E} R_n\left(\mathcal{H}_\varphi(Z^n)\right). \tag{8.6}$$

Next, consider the class of functions $\mathcal{H}$ of the form $h(x, y) = -yf(x)$ for $f \in \mathcal{F}$. We shall now show that the multiplicative structure of the loss function with $y \in \{-1, 1\}$ implies that for any sample $Z^n = (X^n, Y^n)$, $R_n(\mathcal{H}(Z^n)) = R_n(\mathcal{F}(X^n))$. That is, given X^n, no matter how the n points $X_1, \ldots, X_n$ are labeled with ± 1's to get Y^n, the Rademacher average of $R_n(\mathcal{H}(Z^n))$ is the same. It is because if ε_i is a Rademacher random variable and Y_i is a fixed value in $\{1, -1\}$, then $\varepsilon_i Y_i$ has the same distribution as ε_i. In detail:

$$\begin{aligned} R_n(\mathcal{H}(Z^n)) &= \frac{1}{n} \mathbf{E}_\varepsilon \left[\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n \varepsilon_i h(Z_i) \right| \right] \\ &= \frac{1}{n} \mathbf{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i Y_i f(X_i) \right| \right] \\ &= \frac{1}{n} \mathbf{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right] \\ &= R_n(\mathcal{F}(X^n)) \end{aligned} \tag{8.7}$$

The next step is to apply the contraction principle for Rademacher averages, Proposition 6.2, using the mapping $F : \mathbb{R} \to \mathbb{R}$ defined by $F(v) = \varphi(v) - \varphi(0)$. Note that $\mathcal{H}_\varphi = F \circ \mathcal{H}$, $F(0) = 0$ (due to the subtraction of $\varphi(0)$) as required for the contraction principle, and F is M_φ Lipschitz continuous. The contraction principle and the fact (8.7) imply

$$R_n(\mathcal{H}_\varphi(Z^n)) = R_n(F \circ \mathcal{H}(Z^n)) \leq 2M_\varphi R_n(\mathcal{H}(Z^n)) = 2M_\varphi R_n(\mathcal{F}(X^n)).$$

Taking the expectation over Z^n and combining with (8.6) yields

$$\mathbf{E}\Delta_n(Z^n) \leq 4M_\varphi \mathbf{E} R_n\left(\mathcal{F}(X^n)\right). \tag{8.8}$$

Now, since the surrogate loss functions $\ell_{\varphi,f}(x,y)$ take values in $[0,B]$, the function $Z^n \mapsto \Delta_n(Z^n)$ has bounded differences with $c_1 = \ldots = c_n = B/n$. Therefore, from (8.8) and from McDiarmid's inequality, we have for every $t > 0$ that

$$\mathbf{P}\left(\Delta_n(Z^n) \geq 4M_\varphi \mathbf{E} R_n(\mathcal{F}(X^n)) + \frac{Bt}{\sqrt{n}}\right) \leq \mathbf{P}\left(\Delta_n(Z^n) \geq \mathbf{E}\Delta_n(Z^n) + \frac{Bt}{\sqrt{n}}\right) \leq e^{-2t^2}.$$

$\square$

Lemma 8.2, together with Proposition 6.1, imply the following theorem.

THEOREM 8.2. *Consider any learning algorithm $\mathcal{A} = \{\mathcal{A}_n\}_{n=1}^\infty$, where, for each n, the mapping $\mathcal{A}_n$ receives the training sample $Z^n = (Z_1, \ldots, Z_n)$ as input and produces a function $\widehat{f}_n : \mathsf{X} \to \mathbb{R}$ from some class $\mathcal{F}$. Let φ be a penalty function satisfying the conditions of Lemma 8.2. The following hold:*

- *(single version) For any n and $t \geq 0$,*

$$L(\widehat{f}_n) \leq A_\varphi(\widehat{f}_n) \leq A_{\varphi,n}(\widehat{f}_n) + 4M_\varphi \mathbf{E} R_n(\mathcal{F}(X^n)) + \frac{Bt}{\sqrt{n}}$$

  with probability at least $1 - \mathrm{e}^{-2t^2}$. (If $B = 1$ and $\mathrm{e}^{-2t^2} = \delta$ then $\frac{Bt}{\sqrt{n}} = \sqrt{\frac{\ln(1/\delta)}{2n}}$ as usual, but we leave in t for generality used in the proof of the next theorem.)
- *(double version) If A is an ERM algorithm for the surrogate loss, for any n and $\delta \in (0,1)$, the following bound holds with probability at least $1 - \mathrm{e}^{-2t^2}$:*

$$L(\widehat{f}_n) \leq A_\varphi(\widehat{f}_n) \leq A^*_\varphi(\mathcal{F}) + 8M_\varphi \mathbf{E} R_n(\mathcal{F}(X^n)) + \frac{2Bt}{\sqrt{n}}$$

Through a simple application of the union bound, we can extend Theorem 8.2 to the case of a possibly countably infinite family $\{\varphi_k\}_{k\geq 1}$ of penalty functions. Both a single and double version holds; for brevity we state the single version.

THEOREM 8.3. *Let $\{\varphi_k\}_{k\geq 1}$ be a family of penalty functions, where each φ_k takes values in $[0,1]$ and is Lipschitz-continuous with constant M_{φ_k}. Then, for any n and any $t > 0$,*

$$L(f) \leq \inf_{k\geq 1}\left\{A_{\varphi_k,n}(f) + 4M_{\varphi_k}\mathbf{E}R_n(\mathcal{F}(X^n)) + \sqrt{\frac{\log k}{n}}\right\} + \frac{t}{\sqrt{n}}, \qquad \forall f \in \mathcal{F} \tag{8.9}$$

with probability at least $1 - 2e^{-2t^2}$.

PROOF. For each $k \geq 1$, by Theorem 8.2,

$$L(f) \leq A_{\varphi_k,n}(f) + 4M_{\varphi_k}\mathbf{E}R_n(\mathcal{F}(X^n)) + \sqrt{\frac{\log k}{n}} + \frac{t}{\sqrt{n}}, \qquad \forall f \in \mathcal{F} \tag{8.10}$$

with probability at least $1 - e^{-2(t+\sqrt{\log k})^2} \geq 1 - k^{-2}e^{-2t^2}$. Therefore, by the union bound,

$$\mathbf{P}\left[\exists f \in \mathcal{F} :\ L(f) > \inf_{k\geq 1}\left\{L_{\varphi_k,n}(f) + 4M_{\varphi_k}\mathbf{E}R_n(\mathcal{F}(X^n)) + \sqrt{\frac{\log k}{n}}\right\} + \frac{t}{\sqrt{n}}\right]$$
$$\leq \sum_{k\geq 1}\frac{e^{-2t^2}}{k^2} \leq 2e^{-2t^2},$$

where we have used the fact that $\sum_{k\geq 1} k^{-2} = \pi^2/6 \leq 2$. $\square$

The main consequence of Theorem 8.3 is the following result:

THEOREM 8.4. *Let $\varphi : \mathbb{R} \to [0,1]$ be a Lipschitz-continuous penalty function with constant M_φ, and such that, for all $0 < \gamma \leq \gamma'$,*

$$\varphi(u/\gamma) \leq \varphi(u/\gamma'), \qquad \forall u. \tag{8.11}$$

Then, for any n and any $t > 0$,

$$L(f) \leq \inf_{\gamma\in(0,1]} \left\{ A_{\varphi(\cdot/\gamma),n}(f) + \frac{8M_\varphi}{\gamma}\mathbf{E}R_n(\mathcal{F}(X^n)) + \sqrt{\frac{\log\log_2(2/\gamma)}{n}} \right\} + \frac{t}{\sqrt{n}}, \qquad \forall f \in \mathcal{F} \tag{8.12}$$

with probability at least $1 - 2e^{-2t^2}$.

REMARK 8.1. For example, the ramp function $\varphi(u) = \min\{1, (1+u)_+\}$ or the saturated exponential $\varphi(u) = \min\{1, e^u\}$ satisfy the monotonicity condition (8.11).

PROOF. For each $k \geq 0$, let $\gamma_k := 1/2^k$. For $k \geq 1$, $\varphi_k(u) := \varphi(u/\gamma_k)$ is a valid penalty function with Lipschitz constant $M_{\varphi_k} \leq \frac{M_\varphi}{\gamma_k}$. Applying Theorem 8.3 to the family $\{\varphi_k\}$, we see that, with probability at least $1 - 2e^{-2t^2}$,

$$L(f) \leq \inf_{k\geq 1} \left\{ A_{\varphi_k,n}(f) + \frac{4M_\varphi}{\gamma_k}\mathbf{E}R_n(\mathcal{F}(X^n)) + \sqrt{\frac{\log k}{n}} \right\} + \frac{t}{\sqrt{n}}, \qquad \forall \in \mathcal{F}. \tag{8.13}$$

Now, for any $\gamma \in (0,1]$, there exists k such that $\gamma \in (\gamma_k, \gamma_{k-1}]$. Then

$$\varphi_k(u) = \varphi(u/\gamma_k) \leq \varphi(u/\gamma)$$

by virtue of (8.11), and the following also hold:

$$\frac{1}{\gamma_k} \leq \frac{2}{\gamma} \qquad \text{and} \qquad \log k = \log\log_2 \frac{1}{\gamma_k} \leq \log\log_2(2/\gamma).$$

Therefore,

$$\begin{aligned} A_{\varphi(\cdot/\gamma),n}(f) &+ \frac{8M_\varphi}{\gamma}\mathbf{E}R_n(\mathcal{F}(X^n)) + \sqrt{\frac{\log\log_2(2/\gamma)}{n}} \\ &\geq A_{\varphi_k,n}(f) + \frac{4M_\varphi}{\gamma_k}\mathbf{E}R_n(\mathcal{F}(X^n)) + \sqrt{\frac{\log k}{n}} \qquad \forall \in \mathcal{F}. \end{aligned}$$

Consequently, (8.13) implies (8.12), and the proof is complete. $\square$

What the above theorems tell us is that the performance of the learned classifier $\operatorname{sgn} \widehat{f}_n$ is controlled by the Rademacher average of the class $\mathcal{F}$ of functions f, and we can always arrange it to be relatively small. In the next four sections of this chapter we look at several specific examples.

8.3. Weighted linear combination of classifiers

Let $\mathcal{G} = \{g : \mathbb{R}^d \to \{-1, 1\}\}$ be a class of *base classifiers* (not to be confused with *Bayes* classifiers!), and consider the class

$$\mathcal{F}_\lambda := \left\{ f = \sum_{j=1}^N c_j g_j : N \in \mathbb{N}, \sum_{j=1}^N |c_j| \le \lambda; g_1, \ldots, g_N \in \mathcal{G} \right\},$$

where $\lambda > 0$ is a tunable parameter. Then for each $f = \sum_{j=1}^N c_j g_j \in \mathcal{F}_\lambda$ the corresponding classifier g_f of the form (8.3) is given by

$$g_f(x) = \operatorname{sgn}\left(\sum_{j=1}^N c_j g_j(x) \right).$$

A useful way of thinking about g_f is that, upon receiving a feature $x \in \mathbb{R}^d$, it computes the outputs $g_1(x), \ldots, g_N(x)$ of the N base classifiers from $\mathcal{G}$ and then takes a weighted "majority vote" – indeed, if we had $c_1 = \ldots = c_N = \lambda/N$, then $\operatorname{sgn}(g_f(x))$ would precisely correspond to taking the majority vote among the N base classifiers. Note, by the way, that the number of base classifiers is not fixed, and can be learned from the data.

Now, Theorem 8.2 tells us that the performance of any learning algorithm that accepts a training sample Z^n and produces a function $\widehat{f}_n \in \mathcal{F}_\lambda$ is controlled by the Rademacher average $R_n(\mathcal{F}_\lambda(X^n))$. It turns out, moreover, that we can relate it to the Rademacher average of the base class $\mathcal{G}$. To start, note that

$$\mathcal{F}_\lambda = \lambda \cdot \operatorname{absconv} \mathcal{G},$$

where

$$\operatorname{absconv} \mathcal{G} = \left\{ \sum_{j=1}^N c_j g_j : N \in \mathbb{N}; \sum_{j=1}^N c = |c_j| \le 1; g_1, \ldots, g_N \in \mathcal{G} \right\}$$

is the absolute convex hull of $\mathcal{G}$. Therefore

$$R_n(\mathcal{F}_\lambda(X^n)) = \lambda \cdot R_n(\mathcal{G}(X^n)).$$

Now note that the functions in $\mathcal{G}$ are binary-valued. Therefore, assuming that the base class $\mathcal{G}$ is a VC class, we will have

$$R_n(\mathcal{G}(X^n)) \le C\sqrt{\frac{V(\mathcal{G})}{n}}.$$

Combining these bounds with the single version bound of Theorem 8.2, we conclude that for any $\widehat{f}_n$ selected from $\mathcal{F}_\lambda$ based on the training sample Z^n, the bound

$$L(\widehat{f}_n) \le A_{\varphi,n}(\widehat{f}_n) + C\lambda M_\varphi \sqrt{\frac{V(\mathcal{G})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

will hold with probability at least $1 - \delta$ (assuming, as before, that φ takes values in $[0, 1]$) and M_φ is the Lipschitz constant of the penalty function φ. The double version bound of Theorem 8.2 can be similarly specialized.

Note that the above bound involves only the VC dimension of the *base class*, which is typically small. On the other hand, the class $\mathcal{F}_\lambda$ obtained by forming weighted combinations of classifiers from $\mathcal{G}$ is extremely rich, and, when thresholded to yield binary valued functions, will generally have infinite VC dimension! But there is a price we pay: The first term is the empirical surrogate risk $A_{\varphi,n}(\widehat{f}_n)$, rather than the empirical classification error $L_n(\widehat{f}_n)$. However, it is possible to choose the penalty function φ in such a way that $A_{\varphi,n}(\cdot)$ can be bounded in terms of a quantity *related* to the number of misclassified training examples. Here is an example.

Fix a positive parameter $\gamma > 0$ and consider

$$\varphi(x) = \begin{cases} 0, & \text{if } x \le -\gamma \\ 1, & \text{if } x \ge 0 \\ 1 + x/\gamma, & \text{otherwise} \end{cases}$$

This is a valid penalty function that takes values in $[0,1]$ and is Lipschitz-continuous with constant $M_\varphi = 1/\gamma$. In addition, we have $\varphi(x) \le \mathbf{1}_{\{x > -\gamma\}}$, which implies that $\varphi(-yf(x)) \le \mathbf{1}_{\{yf(x) < \gamma\}}$. Therefore, for any f we have

$$A_{\varphi,n}(f) = \frac{1}{n}\sum_{i=1}^n \varphi(-Y_i f(X_i)) \le \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i f(X_i) < \gamma\}}. \tag{8.14}$$

The quantity

$$L_n^\gamma(f) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i f(X_i) < \gamma\}} \tag{8.15}$$

is called the *margin error* of f. Notice that:

- For any $\gamma > 0$, $L_n^\gamma(f) \ge L_n(f)$
- The function $\gamma \mapsto L_n^\gamma(f)$ is increasing.

Notice also that we can write

$$L_n^\gamma(f) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i f(X_i) < 0\}} + \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{0 \le Y_i f(X_i) < \gamma\}},$$

where the first term is just $L_n(f)$, while the second term is the number of training examples that were classified correctly, but only with small "margin" (the quantity $Yf(X)$ is often called the *margin* of the classifier f).

Theorem 8.5 (Margin-based risk bound for weighted linear combinations). *For any $\gamma > 0$, the bound*

$$L(\widehat{f}_n) \le L_n^\gamma(\widehat{f}_n) + \frac{C\lambda}{\gamma}\sqrt{\frac{V(\mathcal{G})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{8.16}$$

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

Remark 8.2. Note that the first term on the right-hand side of (8.16) increases with γ, while the second term decreases with γ. Hence, if the learned classifier $\widehat{f}_n$ has a small margin error for a large γ, i.e., it classifies the training samples well and with high "confidence," then its generalization error will be small.

8.4. AdaBoost

A particular strategy for combining classifiers is the so-called AdaBoost algorithm of Freund and Schapire [**FS97**]. Let a class $\mathcal{G}$ of classifiers $g : \mathbb{R}^d \to \{-1, +1\}$ be given; the elements $g \in \mathcal{G}$ are referred to as *weak learners*. Given training data $Z^n = (Z_1, \ldots, Z_n)$, where each $Z_i = (X_i, Y_i)$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \{-1, +1\}$, the AdaBoost algorithm works iteratively as follows:

- Initialize $w^{(1)} = (w_1^{(1)}, \ldots, w_n^{(1)})$ with $w_i^{(1)} = 1/n$ for all i.
- At each iteration $k = 1, \ldots, K$:
  - let $g_k \in \mathcal{G}$ be any weak learner that minimizes the weighted empirical error

$$e_k(g) := \sum_{i=1}^{n} w_i^{(k)} \mathbf{1}_{\{Y_i \neq g(X_i)\}} \tag{8.17}$$

    over $\mathcal{G}$. Let $e_k := e_k(g_k)$. The standing assumption is that $e_k \le 1/2$, i.e., there exists at least one weak learner with better-than-chance performance.
  - Update the weight vector $w^{(k)}$ to $w^{(k+1)}$, where, for each $i \in [n]$,

$$w_i^{(k+1)} = \frac{w_i^{(k)} \exp\left(-\alpha_k Y_i g_k(X_i)\right)}{\mathcal{Z}_k}, \tag{8.18}$$

    where $\alpha_k := \frac{1}{2} \log \frac{1-e_k}{e_k}$ and

$$\mathcal{Z}_k := \sum_{i=1}^{n} w_i^{(k)} \exp\left(-\alpha_k Y_i g_k(X_i)\right). \tag{8.19}$$

- After K iterations, output the classifier $\widehat{f}_n : \mathbb{R}^d \to \mathbb{R}$ given by

$$\widehat{f}_n(x) := \frac{\sum_{k=1}^{K} \alpha_k g_k(x)}{\sum_{k=1}^{K} \alpha_k}. \tag{8.20}$$

  Note that, since $\alpha_k \ge 0$ for all k, $\widehat{f}_n \in \mathrm{conv}(\mathcal{G})$. (The reason $\widehat{f}_n(x)$ is normalized in (8.20) is for theoretical purposes. The final output $\mathrm{sgn}(\widehat{f}_n(x))$ is unaffected.)

The following lemma is crucial in the analysis of AdaBoost:

Lemma 8.3.

$$\frac{1}{n} \sum_{i=1}^{n} \exp\left(-Y_i \sum_{k=1}^{K} \alpha_k g_k(X_i)\right) = \prod_{k=1}^{K} 2\sqrt{e_k(1-e_k)}. \tag{8.21}$$

Proof. From the form of the AdaBoost weight update rule (8.18), we see that, for any $i \in [n]$ and any $k \in [K]$,

$$\exp\left(-\alpha_k Y_i g_k(X_i)\right) = \frac{w_i^{(k+1)}}{w_i^{(k)}} \mathcal{Z}_k. \tag{8.22}$$

Therefore,

$$\begin{aligned}
\exp\left(-Y_i\sum_{k=1}^K \alpha_k g_k(X_i)\right) &= \prod_{k=1}^K \exp\left(-\alpha_k Y_i g_k(X_i)\right) \\
&= \prod_{k=1}^K \frac{w_i^{(k+1)}}{w_i^{(k)}} \mathcal{Z}_k \\
&= \frac{w_i^{(K+1)}}{w_i^{(1)}} \prod_{k=1}^K \mathcal{Z}_k \\
&= n w_i^{(K+1)} \prod_{k=1}^K \mathcal{Z}_k.
\end{aligned}$$

Averaging this over i gives

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \exp\left(-Y_i\sum_{k=1}^K \alpha_k g_k(X_i)\right) &= \prod_{k=1}^K \mathcal{Z}_k \sum_{i=1}^n w_i^{(K+1)} \\
&= \prod_{k=1}^K \mathcal{Z}_k.
\end{aligned}$$

The proof will be complete once we show that $\mathcal{Z}_k = 2\sqrt{e_k(1-e_k)}$ for each k. But this is rather simple:

$$\begin{aligned}
\mathcal{Z}_k &= \sum_{i=1}^n w_i^{(k)} \exp\left(-Y_i \alpha_k g_k(X_i)\right) \\
&= e^{\alpha_k} \sum_{i=1}^n w_i^{(k)} \mathbf{1}_{\{Y_i \neq g_k(X_i)\}} + e^{-\alpha_k} \sum_{i=1}^n w_i^{(k)} \mathbf{1}_{\{Y_i = g_k(X_i)\}} \\
&= e^{\alpha_k} e_k + e^{-\alpha_k}(1-e_k) \\
&= 2\sqrt{e_k(1-e_k)},
\end{aligned}$$

where we have used (8.19) and the definition of α_k. □

We can now state a performance guarantee for AdaBoost due to Koltchinskii and Panchenko [**KP02**], who improved upon the results of Schapire et al. [**SFBL98**]:

Theorem 8.6. *With probability at least $1-\delta$, the classifier $\widehat{f}_n$ generated by K iterations of AdaBoost satisfies*

$$\begin{aligned}
L(\widehat{f}_n) \le \prod_{k=1}^K 2\sqrt{e_k(1-e_k)} &+ 8\left(1 \vee \log \prod_{k=1}^K \sqrt{\frac{1-e_k}{e_k}}\right) \mathbf{E} R_n(\mathcal{G}(X^n)) \\
&+ \sqrt{\frac{1}{n}\log\log_2\left(2\left(1 \vee \log\prod_{k=1}^K \sqrt{\frac{1-e_k}{e_k}}\right)\right)} + \sqrt{\frac{\log(1/\delta)}{2n}}. \qquad (8.23)
\end{aligned}$$

Proof. Let $\varphi : \mathbb{R}^d \to [0,1]$ be a penalty function which is Lipschitz-continuous with constant 1 and satisfies (8.11) and the additional condition $\varphi(u) \le e^u$. For example, we could take $\varphi(u) = 1 \wedge e^u$. Define the data dependent constant γ by

$$\gamma := 1 \wedge \frac{1}{\sum_{k=1}^K \alpha_k}.$$

Then, since $\widehat{f}_n \in \mathrm{conv}(\mathcal{G})$ and $R_n(\mathrm{conv}(\mathcal{G})(X^n)) = R_n(\mathcal{G}(X^n))$, Theorem 8.4 guarantees that, with probability at least $1-\delta$,

$$L(\widehat{f}_n) \le A_{\varphi(\cdot/\gamma),n}(\widehat{f}_n) + \frac{8}{\gamma}\mathbf{E}R_n(\mathcal{G}(X^n)) + \sqrt{\frac{\log\log_2(2/\gamma)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{8.24}$$

Now, from the definition of γ and the assumption that $\varphi(u) \le e^u$, it follows that, for any $i \in [n]$,

$$\begin{aligned}\varphi\left(-\frac{Y_i\widehat{f}_n(X_i)}{\gamma}\right) &\le \varphi\left(-Y_i\sum_{k=1}^K \alpha_k g_k(X_i)\right)\\ &\le \exp\left(-Y_i\sum_{k=1}^K \alpha_k g_k(X_i)\right).\end{aligned}$$

Using this together with Lemma 8.3 gives

$$A_{\varphi(\cdot/\gamma),n}(\widehat{f}_n) \le \frac{1}{n}\sum_{i=1}^n \exp\left(-Y_i\sum_{k=1}^K \alpha_k g_k(X_i)\right) \le \prod_{k=1}^K 2\sqrt{e_k(1-e_k)}.$$

Moreover,

$$\frac{1}{\gamma} = 1 \vee \left(\sum_{k=1}^K \alpha_k\right) = 1 \vee \left(\prod_{k=1}^K \log\sqrt{\frac{1-e_k}{e_k}}\right).$$

Using these in (8.24), we see that (8.23) holds with probability at least $1-\delta$. □

8.5. Neural nets

The basic theorem for classification based on surrogate loss, Theorem 8.2, gives a performance guarantee for classifiers of the form $\widehat{y} = \mathrm{sgn}\, f(x)$ in terms of the expected Rademacher average $\mathbf{E}R_n(\mathcal{F}(X^n))$, which depends on the distribution of the X's and the class of functions $\mathcal{F}$. In this section we consider classes of functions $\mathcal{F}$ of the form arising in neural networks, and provide bounds on $\mathbf{E}R_n(\mathcal{F}(X^n))$. We begin by considering linear functions, of the form $f(x) = \langle w, x\rangle$, where both the feature vector x and the classifier weight vector w are elements of $\mathbb{R}^d$.

Let $\mathcal{F}$ be the collection of all such classifiers satisfying the norm constraint $\|w\| \le B$:

$$\mathcal{F} := \{\langle w, \cdot\rangle : \|w\| \le B\}.$$

Then, for any realization of $X_1, \ldots, X_n$,

$$\begin{aligned}
R_n(\mathcal{F}(X^n)) &= \frac{1}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{w\in\mathbb{R}^d\colon \|w\|\le B}\left|\sum_{i=1}^n \varepsilon_i\langle w, X_i\rangle\right|\right] \\
&= \frac{1}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{\|w\|\le B}\left|\left\langle w, \sum_{i=1}^n \varepsilon_i X_i\right\rangle\right|\right] \\
&= \frac{B}{n}\mathbf{E}_{\varepsilon^n}\left\|\sum_{i=1}^n \varepsilon_i X_i\right\| \\
&\le \frac{B}{n}\cdot\sqrt{\sum_{i=1}^n \|X_i\|^2},
\end{aligned}$$

where the third step is by the Cauchy–Schwarz inequality, and the last step follows from the following calculation: for any collection of vectors $v_1, \ldots, v_n \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbf{E}_{\varepsilon^n}\left\|\sum_{i=1}^n \varepsilon_i v_i\right\| &= \mathbf{E}_{\varepsilon^n}\sqrt{\sum_{i=1}^n\sum_{j=1}^n \varepsilon_i\varepsilon_j\langle v_i, v_j\rangle} \\
&\le \sqrt{\mathbf{E}_{\varepsilon^n}\left[\sum_{i=1}^n\sum_{j=1}^n \varepsilon_i\varepsilon_j\langle v_i, v_j\rangle\right]} \\
&= \sqrt{\sum_{i=1}^n \|v_i\|^2},
\end{aligned}$$

where we have used Jensen's inequality and the fact that, by independence of the ε_i's, $\mathbf{E}[\varepsilon_i\varepsilon_j] = \mathbf{1}_{\{i=j\}}$. In particular, if the common distribution of the X_i's is supported on the radius-R ball centered at the origin, then

$$\mathbf{E}R_n(\mathcal{F}(X^n)) \le \frac{BR}{\sqrt{n}}. \tag{8.25}$$

Note that this bound is completely dimension-free, in contrast to the bound derived in Section 8.1.1, namely, that for a family of Dudley classifiers with dimension m, Theorem 8.1 holds with $V(\mathcal{C}) = m$. This is due to the fact the latter was obtained under no restrictions on either the classifier weight vector or the feature vector.

Using the contraction principle, we can also cover the case of nonlinear classifiers of the form

$$f(x) = \sigma(\langle w, x\rangle), \tag{8.26}$$

where, as before, $\|w\| \le B$, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a fixed Lipschitz-continuous function with Lipschitz constant L that satisfies $\sigma(0) = 0$. Then

$$\begin{aligned} R_n(\mathcal{F}(X^n)) &= \frac{1}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{\|w\|\le B}\left|\sum_{i=1}^n \varepsilon_i \sigma(\langle w, X_i\rangle)\right|\right] \\ &\le \frac{2L}{n}\cdot \mathbf{E}_{\varepsilon^n}\left[\sup_{\|w\|\le B}\left|\sum_{i=1}^n \varepsilon_i \langle w, X_i\rangle\right|\right] \\ &\le \frac{2LB}{n}\cdot\sqrt{\sum_{i=1}^n \|X_i\|^2}. \end{aligned}$$

Again, the bound is dimension-free — it depends only on the Lipschitz constant of σ and on the maximal ℓ_2 norm of the weight w. If the features $X_1, \dots, X_n$ are each supported on a ball of radius R, then

$$\mathbf{E}R_n(\mathcal{F}(X^n)) \le \frac{2LBR}{\sqrt{n}}.$$

The classifier in Eq. (8.26) is a basic building block of *neural nets*. In a very rough analogy with biological neurons, its output is a nonlinear function of some linear combination of its inputs. The nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$ is known as the *activation function*.

Next we consider feedforward neural networks. First, given a nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$ and a vector $w \in \mathbb{R}^m$ for some m, define the function $N_{\sigma,w} : \mathbb{R}^m \to \mathbb{R}$ by

$$N_{\sigma,w}(u_1, \dots, u_m) := \sigma\left(\sum_{j=1}^m w_j u_j\right),$$

and denote $N_{\sigma,w}$ composed with m real-valued functions $(h_k)_{1\le k\le m}$ on $\mathbb{R}^d$ by

$$\begin{aligned} \big[N_{\sigma,w} \circ (h_1, \dots, h_m)\big](x) &:= N_{\sigma,w}\left(h_1(x), \dots, h_m(x)\right) \\ &= \sigma\left(\sum_{j=1}^m w_j h_j(x)\right). \end{aligned}$$

Let $\mathcal{G}$ be a family of base classifiers $g : \mathsf{X} \to \mathbb{R}$, where X is a subset of $\mathbb{R}^d$. For $\ell \ge 1$, let $\sigma_1, \dots, \sigma_\ell : \mathbb{R} \to \mathbb{R}$ be a sequence of nonlinearities, such that, for each $j \in [\ell]$, σ_j is Lipschitz-continuous with Lipschitz constant L_j, and $\sigma_j(0) = 0$. Finally, let a sequence of positive reals $B_1, \dots, B_\ell$ be given. We then define function classes $\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_\ell$ recursively as follows:

$$\mathcal{F}_0 := \mathcal{G},$$

and, for $1 \le j \le \ell$,

$$\mathcal{F}_j := \left\{N_{\sigma_j,w} \circ (f_1, \dots, f_m) : m \in \mathbb{N};\ |w_1| + \ldots + |w_m| \le B_j;\ f_1, \dots, f_m \in \mathcal{F}_{j-1}\right\}.$$

In other words, the functions in $\mathcal{F}_1$ are of the form

$$f(x) = \sigma_1\left(\sum_{j=1}^{m} w_j g_j(x)\right)$$

for all $m \in \mathbb{N}$, all vectors $w \in \mathbb{R}^m$ satisfying $\|w\|_1 := |w_1| + \ldots + |w_m| \le B_1$, and all choices of base classifiers $g_1, \ldots, g_m \in \mathcal{G}$; the functions in $\mathcal{F}_2$ are all functions of the form

$$f(x) = \sigma_2\left(\sum_{j=1}^{m} w_j f_j(x)\right),$$

for all $m \in \mathbb{N}$, all $w \in \mathbb{R}^m$ satisfying $\|w\|_1 \le B_2$, and all choices of $f_1, \ldots, f_m \in \mathcal{F}_1$; and so on. The integer ℓ is called the number of layers.

Now let us upper-bound the Rademacher average $\mathbf{E}R_n(\mathcal{F}_\ell(X^n))$. To that end, we first observe the following structural property:

$$\mathcal{F}_j = \sigma_j \circ (B_j \cdot \mathrm{absconv}(\mathcal{F}_{j-1}))\,. \tag{8.27}$$

In other words, each $f_j \in \mathcal{F}_j$ has the form $\sigma_j \circ \widetilde{f}_j$ for some $\widetilde{f}_j \in B_j \cdot \mathrm{absconv}(\mathcal{F}_{j-1})$. Armed with this, we start at the last (i.e., ℓth) layer and proceed recursively. Since $\mathcal{F}_\ell = \sigma_\ell \circ (B_\ell \cdot \mathrm{absconv}(\mathcal{F}_{\ell-1}))$, the same holds for the families of vectors obtained by applying the families of functions to the n data samples, X^n. Therefore,

$$\begin{aligned} R_n(\mathcal{F}_\ell(X^n)) &= R_n(\sigma_\ell \circ (B_\ell \cdot \mathrm{absconv}(\mathcal{F}_{\ell-1}(X^n)))) \\ &\le 2L_\ell \cdot R_n(B_\ell \cdot \mathrm{absconv}(\mathcal{F}_{\ell-1}(X^n))) \\ &= 2L_\ell B_\ell \cdot R_n(\mathcal{F}_{\ell-1}(X^n)), \end{aligned}$$

where the second line uses the contraction principle, and the last line uses the properties of Rademacher averages. Thus, we have "peeled off" the last layer. Proceeding inductively, we arrive at the bound

$$R_n(\mathcal{F}_\ell(X^n)) \le \prod_{j=1}^{\ell}(2L_j B_j) \cdot R_n(\mathcal{G}(X^n)). \tag{8.28}$$

Apart from the Rademacher average of the "base" class $\mathcal{G}$, the bound (8.28) involves only the number of layers ℓ, the Lipschitz constants $L_1, \ldots, L_\ell$ of the activation functions in each layer, and the weight constraints $B_1, \ldots, B_n$. In particular, the number of neurons does not appear explicitly anywhere in the bound. The first bound of this sort was obtained by Bartlett [**Bar98**] (see also [**BM02**]). The bound in (8.28) is due to Koltchinskii and Panchenko [**KP02**]. However, observe that, by invoking the contraction principle, we gain a factor of 2 at each layer. Thus, the bound is very loose unless $\prod_{j=1}^{\ell}(L_j B_j) \le 2^{-\ell}$, which is quite a tall order — for example, for the so-called *rectified linear unit* (or ReLU) activation function $\sigma(u) := u \vee 0$, then $L_j = 1$ for all j, and therefore we must have $\prod_{j=1}^{\ell} B_j \le 2^{-\ell}$.

Deep neural nets, i.e., neural nets with ℓ very large, have become popular recently due to their remarkable empirical performance in a variety of domains, such as computer vision, speech processing, and natural language processing. However, theoretical understanding of their performance is still incomplete. One worrisome issue is the explicit exponential dependence of the bound (8.28) on the number of layers. Bartlett, Foster, and Telgarsky

[**BFT17**] have removed this dependence using rather delicate recursive covering number estimates. In a more recent paper, Golowich, Rakhlin, and Shamir [**GRS17**] showed that the factor of 2^ℓ can be reduced to $\sqrt{\ell}$ using a simple but effective log-exp device. In order to state and prove their result, we need two technical lemmas:

LEMMA 8.4 ([**GRS17**]). *Let $\mathcal{F}$ and $\mathcal{F}'$ be two classes of real-valued functions on X, such that*

$$\mathcal{F} = \sigma \circ (B \cdot \mathrm{absconv}(\mathcal{F}')) \tag{8.29}$$

for some Lipschitz-continuous nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$ satisfying $\sigma(0) = 0$. Let $G : \mathbb{R} \to \mathbb{R}$ be a convex nondecreasing function. Then

$$\mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}} G\left(\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right)\right] \le 2\cdot \mathbf{E}_{\varepsilon^n}\left[\sup_{f'\in\mathcal{F}'} G\left(LB\cdot\left|\sum_{i=1}^n \varepsilon_i f'(X_i)\right|\right)\right], \tag{8.30}$$

where L is the Lipschitz constant of σ.

PROOF. Since G is nondecreasing, $G(|u|) = G(u \vee (-u)) \le G(u) + G(-u)$. Then

$$\begin{aligned}
&\mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}} G\left(\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right)\right] \\
&\le \mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}} G\left(\sum_{i=1}^n \varepsilon_i f(X_i)\right)\right] + \mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}} G\left(-\sum_{i=1}^n \varepsilon_i f(X_i)\right)\right] \\
&= 2\cdot \mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}} G\left(\sum_{i=1}^n \varepsilon_i f(X_i)\right)\right],
\end{aligned} \tag{8.31}$$

where the last step uses the fact that Rademacher random variables are symmetric. We now invoke the following generalization of the contraction principle (see [**LT91**, Eq. (4.20)]): For any $\mathcal{A} \subset \mathbb{R}^n$ and for G, σ satisfying the conditions of the lemma,

$$\mathbf{E}_{\varepsilon^n}\left[\sup_{a\in\mathcal{A}} G\left(\sum_{i=1}^n \varepsilon_i \sigma(a_i)\right)\right] \le \mathbf{E}_{\varepsilon^n}\left[\sup_{a\in\mathcal{A}} G\left(L\cdot\sum_{i=1}^n \varepsilon_i a_i\right)\right], \tag{8.32}$$

where L is the Lipschitz constant of σ. Now, by (8.29), any $f \in \mathcal{F}$ has the form $f = \sigma \circ \widetilde{f}$ for some $\widetilde{f} \in B \cdot \mathrm{absconv}(\mathcal{F}')$. Using this fact together with (8.32), we get

$$\mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}} G\left(\sum_{i=1}^n \varepsilon_i f(X_i)\right)\right] \le \mathbf{E}_{\varepsilon^n}\left[\sup_{\widetilde{f}\in B\cdot\mathrm{absconv}(\mathcal{F}')} G\left(L\cdot\sum_{i=1}^n \varepsilon_i \widetilde{f}(X_i)\right)\right]. \tag{8.33}$$

Now, by Hölder's inequality and by the monotonicity of G,

$$\sup_{\widetilde{f}\in B\cdot\mathrm{absconv}(\mathcal{F}')} G\left(L\cdot\sum_{i=1}^n \varepsilon_i \widetilde{f}(X_i)\right) \le \sup_{f'\in\mathcal{F}'} G\left(LB\cdot\left|\sum_{i=1}^n \varepsilon_i f'(X_i)\right|\right). \tag{8.34}$$

Combining Eqs. (8.31), (8.33), and (8.34), we get (8.30). $\square$

LEMMA 8.5. *Let $\mathcal{A}$ be a bounded subset of $\mathbb{R}^n$. Then, for any $\lambda > 0$,*

$$\mathbf{E}_{\varepsilon^n}\left[\exp\left(\lambda \sup_{a\in\mathcal{A}}\left|\sum_{i=1}^n \varepsilon_i a_i\right|\right)\right] \le \exp\left(\frac{\lambda^2}{2}\sum_{i=1}^n \sup_{a\in\mathcal{A}}|a_i|^2\right)\exp\left(\lambda n R_n(\mathcal{A})\right). \tag{8.35}$$

PROOF. The random variable $U := \sup_{a\in\mathcal{A}}|\sum_{i=1}^n \varepsilon_i a_i|$ is a deterministic function of $\varepsilon_1,\dots,\varepsilon_n$, and, for each i,

$$U(\varepsilon_1,\dots,\varepsilon_i,\dots,\varepsilon_n) - U(\varepsilon_1,\dots,-\varepsilon_i,\dots,\varepsilon_n) \le 2\sup_{a\in\mathcal{A}}|a_i|.$$

Thus, mimicking the proof of McDiarmid's inequality, we arrive at

$$\begin{aligned}\mathbf{E}[e^{\lambda U}] &= e^{\lambda\,\mathbf{E}U}\mathbf{E}[e^{\lambda(U-\mathbf{E}U)}]\\ &\le e^{\lambda\,\mathbf{E}U}\cdot\exp\left(\frac{\lambda^2}{2}\sum_{i=1}^n\sup_{a\in\mathcal{A}}|a_i|^2\right).\end{aligned}$$

□

We are now ready to prove the following result, due to Golowich, Rakhlin, and Shamir [**GRS17**]:

THEOREM 8.7. *For any realization $X_1,\dots,X_n$,*

$$R_n(\mathcal{F}_\ell(X^n)) \le \prod_{j=1}^{\ell}(L_jB_j)\cdot\left(R_n(\mathcal{G}(X^n)) + \frac{2}{n}\sqrt{\ell\log 2\cdot\sum_{i=1}^n\sup_{g\in\mathcal{G}}|g(X_i)|^2}\right). \tag{8.36}$$

PROOF. Fix some $\lambda > 0$. Then

$$\begin{aligned}R_n(\mathcal{F}_\ell(X^n)) &= \frac{1}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}_\ell}\left|\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right]\\ &= \frac{1}{\lambda n}\mathbf{E}_{\varepsilon^n}\left[\log\exp\sup_{f\in\mathcal{F}_\ell}\lambda\left|\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right]\\ &\le \frac{1}{\lambda n}\log\mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}_\ell}\exp\left(\lambda\left|\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right)\right], \qquad (8.37)\end{aligned}$$

where the last step is by Jensen's inequality. Now let $G(u) := e^{\lambda u}$. Then, taking into account (8.27) and invoking Lemma 8.4, we can write

$$\begin{aligned}&\mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}_\ell}\exp\left(\lambda\left|\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right)\right]\\ &\le 2\cdot\mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}_{\ell-1}}\exp\left(\lambda L_\ell B_\ell\cdot\left|\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right)\right].\end{aligned}$$

Continuing inductively in this manner, we arrive at

$$\mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}_\ell}\exp\left(\lambda\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right)\right]$$

$$\le 2^\ell\cdot\mathbf{E}_{\varepsilon^n}\left[\exp\left(\lambda\prod_{j=1}^\ell (L_jB_j)\cdot\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^n\varepsilon_i g(X_i)\right|\right)\right]. \tag{8.38}$$

Substituting (8.38) into (8.37), we have

$$R_n(\mathcal{F}_\ell(X^n)) \le \frac{1}{\lambda n}\left\{\ell\cdot\log 2+\log\mathbf{E}_{\varepsilon^n}\left[\exp\left(\lambda\prod_{j=1}^\ell (L_jB_j)\cdot\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^n\varepsilon_i g(X_i)\right|\right)\right]\right\}. \tag{8.39}$$

Let $M := \prod_{j=1}^\ell (L_jB_j)$. Then, by Lemma 8.5,

$$\mathbf{E}_{\varepsilon^n}\left[\exp\left(\lambda\prod_{j=1}^\ell (L_jB_j)\cdot\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^n\varepsilon_i g(X_i)\right|\right)\right]$$

$$\le \exp\left(\frac{\lambda^2M^2}{2}\sum_{i=1}^n\sup_{g\in\mathcal{G}}|g(X_i)|^2\right)\exp\left(M\lambda nR_n(\mathcal{G}(X^n))\right). \tag{8.40}$$

Using (8.40) iin (8.39), we get

$$\begin{aligned} R_n(\mathcal{F}_\ell(X^n)) &\le \frac{1}{\lambda n}\left\{\ell\log 2+\frac{\lambda^2M^2}{2}\sum_{i=1}^n\sup_{g\in\mathcal{G}}|g(X_i)|^2+M\lambda nR_n(\mathcal{G}(X^n))\right\}\\ &= MR_n(\mathcal{G}(X^n))+\frac{\ell\log 2}{\lambda n}+\frac{\lambda M^2}{2n}\sum_{i=1}^n\sup_{g\in\mathcal{G}}|g(X_i)|^2. \end{aligned}$$

Using the identity

$$\inf_{\lambda\ge 0}\left\{\frac{a}{\lambda}+b\lambda\right\}=2\sqrt{ab}$$

for $a,b\ge 0$, we finally obtain

$$R_n(\mathcal{F}(X^n)) \le MR_n(\mathcal{G}(X^n))+\frac{2M}{n}\sqrt{\ell\log 2\cdot\sum_{i=1}^n\sup_{g\in\mathcal{G}}|g(X_i)|^2}.$$

□

8.6. Kernel machines

As we have seen, a powerful way of building complex classifiers from simple ones is by using functions that are linear combinations of simple functions. The norms induced by kernels, as described in Chapter 4, offer an effective way to control the complexity of the linear combinations. Kernel methods are popular in machine learning for a variety of reasons, not the least of which is that any algorithm that operates in a Euclidean space and relies only on the computation of inner products between feature vectors can be modified to work with any suitably well-behaved kernel.

Let us describe empirical risk minimization in an RKHS. Pick a kernel K on our feature space X, where X is a closed subset of $\mathbb{R}^d$, and consider classifiers of the form

$$g_f(x) = \operatorname{sgn} f(x) \equiv \begin{cases} 1, & \text{if } f(x) \ge 0 \\ -1, & \text{otherwise} \end{cases}$$

with the underlying f taken from a suitable subset of the RKHS $\mathcal{H}_K$. One choice, which underlies such things as the Support Vector Machine, is to take a ball in $\mathcal{H}_K$: given some $\lambda > 0$, let

$$\mathcal{F}_\lambda := \{f \in \mathcal{H}_K : \|f\|_K \le \lambda\}.$$

This set is the closure (in the $\|\cdot\|_K$ norm) of the convex set

$$\left\{ \sum_{j=1}^N c_j K_{x_j} : N \in \mathbb{N}; c_1, \dots, c_N \in \mathbb{R}; x_1, \dots, x_N \in \mathsf{X}; \sum_{i,j=1}^N c_i c_j K(x_i, x_j) \le \lambda^2 \right\} \subset \mathcal{L}_K(\mathsf{X}),$$

and is itself convex.

As we already know, the performance of any learning algorithm that chooses an element $\widehat{f}_n \in \mathcal{F}_\lambda$ in a data-dependent way is controlled by the Rademacher average $R_n(\mathcal{F}_\lambda(X^n))$. An advantage of using kernels is that this Rademacher average can be estimated using one of the bounds in the following proposition (in this section we use the second bound in the proposition):

Proposition 8.1. *(i) Let $C_K = \sqrt{\sup_{x \in X} K(x,x)}$. For any $x_1, \dots, x_n \in \mathsf{X}$,*

$$R_n(\mathcal{F}_\lambda(x^n)) \le \frac{C_K \lambda}{\sqrt{n}}. \tag{8.41}$$

(ii) Suppose $X_1, \dots, X_n$ are independent, and each having the distribution of a random variable X with values in X. Then

$$\mathbf{E} R_n(\mathcal{F}_\lambda(X^n)) \le \frac{\lambda \sqrt{\mathbf{E} K(X,X)}}{\sqrt{n}}. \tag{8.42}$$

Proof. By the reproducing kernel property (4.13) and then the linearity of the inner product $\langle \cdot, \cdot \rangle_K$,

$$\begin{aligned} R_n(\mathcal{F}_\lambda(x^n)) &= \frac{1}{n} \mathbf{E}_{\varepsilon^n} \sup_{f : \|f\|_K \le \lambda} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \\ &= \frac{1}{n} \mathbf{E}_{\varepsilon^n} \sup_{f : \|f\|_K \le \lambda} \left| \sum_{i=1}^n \varepsilon_i \langle f, K_{x_i} \rangle_K \right| \\ &= \frac{1}{n} \mathbf{E}_{\varepsilon^n} \sup_{f : \|f\|_K \le \lambda} \left| \left\langle f, \sum_{i=1}^n \varepsilon_i K_{x_i} \right\rangle_K \right| \end{aligned}$$

Using the Cauchy–Schwarz inequality (4.2), it is not hard to show that

$$\sup_{f : \|f\|_K \le \lambda} |\langle f, g \rangle_K| = \lambda \|g\|_K$$

for any $g \in \mathcal{H}_K$. Therefore,

$$R_n(\mathcal{F}_\lambda(x^n)) = \frac{\lambda}{n}\mathbf{E}_{\varepsilon^n}\left\|\sum_{i=1}^n \varepsilon_i K_{x_i}\right\|_K.$$

Next we prove that for any n functions $g_1, \ldots, g_n \in \mathcal{H}_K$,

$$\mathbf{E}_{\varepsilon^n}\left\|\sum_{i=1}^n \varepsilon_i g_i\right\|_K \le \sqrt{\sum_{i=1}^n \|g_i\|_K^2}. \tag{8.43}$$

The proof is in two steps: First, by concavity of the square root and Jensen's inequality:

$$\mathbf{E}_{\varepsilon^n}\sqrt{\left\|\sum_{i=1}^n \varepsilon_i g_i\right\|_K^2} \le \sqrt{\mathbf{E}\left\|\sum_{i=1}^n \varepsilon_i g_i\right\|_K^2}.$$

Then we expand the squared norm:

$$\left\|\sum_{i=1}^n \varepsilon_i g_i\right\|_K^2 = \left\langle \sum_{i=1}^n \varepsilon_i g_i, \sum_{i=1}^n \varepsilon_i g_i \right\rangle_K = \sum_{i,j=1}^n \varepsilon_i \varepsilon_j \langle g_i, g_j\rangle_K.$$

And finally we take the expectation over ε^n and use the fact that $\mathbf{E}[\varepsilon_i\varepsilon_j] = 1$ if $i = j$ and 0 otherwise to get

$$\mathbf{E}\left\|\sum_{i=1}^n \varepsilon_i g_i\right\|_K^2 = \sum_{i=1}^n \langle g_i, g_i\rangle_K = \sum_{i=1}^n \|g_i\|_K^2.$$

Hence, we obtain

$$R_n(\mathcal{F}_\lambda(x^n)) \le \frac{\lambda}{n}\sqrt{\sum_{i=1}^n \langle K_{x_i}, K_{x_i}\rangle_K} = \frac{\lambda}{n}\sqrt{\sum_{i=1}^n K(x_i, x_i)}. \tag{8.44}$$

Equation (8.41) follows from (8.44). Replacing x_i by X_i in (8.44) for each i, taking the expectation w.r.t. X^n over each side, and once more using concavity of the square root and Jensen's inequality, yields (8.42). $\square$

REMARK 8.3. *If $K(x, y) = \langle x, y\rangle$ then $\mathcal{H}_K$ consists of functions of the form $f(x) = \langle w, x\rangle$, and $\|f\|_K^2 = \|w\|^2$. Hence, (8.42) implies (8.25).*

With the bound (8.42) in hand, we can specialize Theorem 8.2 to get the following more explicit bound.

COROLLARY 8.2. *(Performance bound for RKHS using surrogate loss) Suppose $\mathcal{F}_\lambda$ is the closed ball of radius $\lambda > 0$ in an RKHS of functions on a closed set $\mathsf{X} \subset \mathbb{R}^d$ with associated Mercer kernal K. Let φ be any penalty function such that $\varphi(x) \ge \min\{1, (1+x)_+\}$, i.e. , φ is greater than or equal to the ramp penalty function. Then*
(Single version) For any n and $\delta \in (0, 1)$, and any learning algorithm, the following bound holds with probability at least $1 - \delta$:

$$L(\widehat{f}_n) \le A_{\varphi,n}(\widehat{f}_n) + 4\lambda\sqrt{\frac{\mathbf{E}\left[K(X,X)\right]}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{8.45}$$

(Double version) For any n and $\delta \in (0,1)$, and for the ERM algorithm for surrogate loss, the following bound holds with probability at least $1-\delta$:

$$L(\widehat{f}_n) \le A^*_\varphi(\mathcal{F}_\lambda) + 8\lambda\sqrt{\frac{\mathbf{E}\left[K(X,X)\right]}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{8.46}$$

Proof. Theorem 8.2 and (8.42) imply (8.45) and (8.46) in case φ is the ramp penalty function, because the ramp penalty function is $1-$*Lipschitz* continuous and takes values in $[0,1]$. Therefore (8.45) and (8.46) hold for any choice of φ that is greater than or equal to the ramp penalty function, because the right-hand sides are increasing in φ (in the pointwise ordering of functions on $\mathbb{R}$. □

Another advantage of working with kernels is that, in many cases, a minimizer of empirical risk over a sufficiently regular subset of an RKHS will have the form of a linear combination of kernels centered at the training feature points. The results ensuring this are often referred to in the literature as *representer theorems*. Here is one such result (due, in a slightly different form, to Schölkopf, Herbrich, and Smola [**SHS01**]), sufficiently general for our purposes:

Theorem 8.8 (The generalized representer theorem). *Let X be a closed subset of $\mathbb{R}^d$ and let Y be a subset of the reals. Consider a nonnegative loss function $\ell : \mathsf{Y}\times\mathsf{Y} \to \mathbb{R}^+$. Let K be a Mercer kernel on X, and let $\mathcal{H}_K$ be the corresponding RKHS.*

Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be an i.i.d. sample from some distribution $P = P_{XY}$ on $\mathsf{X}\times\mathsf{Y}$, let $\mathcal{H}_n$ be the closed linear subspace of $\mathcal{H}_K$ spanned by $\{K_{X_i} : 1 \le i \le n\}$, and let Π_n denote the orthogonal projection onto $\mathcal{H}_n$. Let $\mathcal{F}$ be a subset of $\mathcal{H}_K$, such that $\Pi_n(\mathcal{F}) \subseteq \mathcal{F}$. Then

$$\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i)) = \inf_{f\in\Pi_n(\mathcal{F})} \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i)), \tag{8.47}$$

and if a minimizer of the left-hand side of (8.47) exists, then it can be taken to have the form

$$\widehat{f}_n = \sum_{i=1}^n c_i K_{X_i} \tag{8.48}$$

for some $c_1,\dots,c_n \in \mathbb{R}$.

Remark 8.4. Note that both the subspace $\mathcal{H}_n$ and the corresponding orthogonal projection Π_n are *random objects*, since they depend on the random features X^n.

Proof. Since $K_{X_i} \in \mathcal{H}_n$ for every i, by Proposition 4.1 we have

$$\langle f, K_{X_i}\rangle_K = \langle \Pi_n f, K_{X_i}\rangle_K, \qquad \forall f \in \mathcal{H}_K.$$

Moreover, from the reproducing kernel property (4.13) we deduce that

$$f(X_i) = \langle f, K_{X_i}\rangle_K = \langle \Pi_n f, K_{X_i}\rangle_K = \Pi_n f(X_i).$$

Therefore, for every $f \in \mathcal{F}$ we can write

$$\frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i)) = \frac{1}{n}\sum_{i=1}^n \ell\left(Y_i, \Pi_n f(X_i)\right).$$

This implies that

$$(8.49)\qquad \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) = \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i, \Pi_n f(X_i)\right) = \inf_{g\in\Pi_n(\mathcal{F})} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, g(X_i)).$$

Now suppose that $f_n \in \mathcal{F}$ achieves the infimum on the left-hand side of (8.49). Then its projection $\widehat{f}_n = \Pi_n f_n$ onto $\mathcal{H}_n$ achieves the infimum on the right-hand side. Moreover, since $\Pi_n(\mathcal{F}) \subseteq \mathcal{F}$ by hypothesis, we may conclude that $\widehat{f}_n \in \mathcal{H}_n$. Since every element of $\mathcal{H}_n$ has the form (8.48), the theorem is proved. □

We now discuss how the representer theorem leads to a computationally efficient ERM algorithm in the classification setting. Consider $\mathsf{Y} = \{-1, 1\}$, a penalty function φ, and classifiers of the form $g(x) = \mathrm{sgn}(f(x))$ for predictors f in some family of real-valued functions $\mathcal{F}$. The empirical φ risk is given by

$$A_{\varphi,n}(f) = \frac{1}{n}\sum_{i=1}^{n} \varphi(-Y_i f(X_i)).$$

Suppose the predictors are taken from $\mathcal{F} = \mathcal{H}_K$, the RKHS generated by some Mercer kernel K. The ERM optimization problem using surrogate loss for a penalty function φ and the RKHS norm for regularization can be formulated as a constrained problem:

$$(8.50)\qquad \min_{f\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^{n} \varphi(-Y_i f(X_i)) \text{ subject to: } \|f\|_K \le \lambda,$$

or as a closely related unconstrained problem with an additive regularization term:

$$(8.51)\qquad \min_{f\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^{n} \varphi(-Y_i f(X_i)) + \tau\|f\|_K^2.$$

Then by the representer theorem, in seeking the ERM defined by (8.50) or the ERM defined by (8.51), we can assume without loss of generality that $\widehat{f}_n \in \mathcal{H}_n \triangleq \mathrm{span}(K_{X_1}, \ldots, K_{X_n})$. In other words,

$$\widehat{f}_n(\cdot) = \sum_{j=1}^{n} c_j K_{X_j}(\cdot) = \sum_{j=1}^{n} c_j K(\cdot, X_j).$$

A solution to the constrained ERM problem (8.50) can be found by solving

$$(8.52)\qquad \min_{c\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} \varphi\left(-Y_i \sum_{j=1}^{n} c_j K(X_i, X_j)\right) \text{ subject to: } \sum_{i,j=1}^{n} c_i c_j K(X_i, X_j) \le \lambda^2,$$

and a solution to the unconstrained regularized ERM problem (8.51) can be found by solving

$$(8.53)\qquad \min_{c\in\mathbb{R}^n} \left\{ \frac{1}{n}\sum_{i=1}^{n} \varphi\left(-Y_i \sum_{j=1}^{n} c_j K(X_i, X_j)\right) + \tau \sum_{i,j=1}^{n} c_i c_j K(X_i, X_j) \right\}.$$

There are n variables involved in either the minimization problem (8.52) or (8.53), where n is the number of labeled data samples. Here n can be much smaller than the dimension of the space $\mathcal{H}_K$ of classifiers used, which could even be infinite. The reduction to n variables is achieved through the use of the representer theorem. Moreover, if φ is a convex function, such as $\varphi(x) = (1+x)_+$, then (8.50) has a convex objective function and quadratic constraints, and (8.53) has a convex objective function, allowing for efficient numerical solution methods such as interior point algorithms. For detailed background of such algorithms see the text of Boyd and Vandenberghe [**BV04**].

An alternative way to go about solving (8.50) or (8.51) is to use a basis that does not depend on the data. Recall from Section 4.3 that if the kernel K has the Mercer expansion $K(x,x') = \sum_i w_i \widetilde{\psi}_i(x)\widetilde{\psi}_i(x')$, then the RKHS norm of a function $f(x) = \sum_j a_j \widetilde{\psi}_j(x)$ satisfies $\|f\|_K^2 = \sum_j a_j^2/w_j$. Inserting this expression for f into the ERM minimization problems (8.50) and (8.51) result in the following constrained ERM minimization problem

$$\min_a \frac{1}{n}\sum_{i=1}^n \varphi\left(-Y_i \sum_j a_j \widetilde{\psi}_j(X_i)\right) \text{ subject to: } \sum_j a_j^2/w_j \le \lambda^2, \tag{8.54}$$

and the related unconstrained problem with additive regularization term:

$$\min_a \frac{1}{n}\sum_{i=1}^n \varphi\left(-Y_i \sum_j a_j \widetilde{\psi}_j(X_i)\right) + \tau \sum_j a_j^2/w_j. \tag{8.55}$$

Thus, (8.54) and (8.55) are equivalent to (8.50) and (8.51), up to a change in coordinates.

Note that the dimension of the vector (a_j) sought in the optimization problems (8.54) and (8.55) does not depend on the number of samples. The dimension of a could be much smaller than the number of samples, for example if d linear functions are used on $\mathsf{X} = \mathbb{R}^d$ and the number of samples is much larger than d. Or the dimension of a could be very large compared to n, or even infinite. In such cases, the optimization problems (8.52) and (8.53) derived from the representer theorem would seem more efficient. However, even if there is a very large number of basis functions, or even infinitely many basis functions, in the representation of K as a series, in practice, only the ten to thirty most heavily weighted basis functions would typically play a role in the performance of a classifier. Thus, in practice, there is often a smaller difference between using basis functions of the form K_x for a Mercer kernel as in (8.52) and (8.53), and directly using basis functions ψ_j as in (8.54) and (8.55).

8.7. Convex risk minimization

Choosing a convex penalty function φ has many advantages in general. First of all, we may arrange things in such a way that the function f^* that minimizes the surrogate loss $A_\varphi(f)$ over all measurable $f : \mathsf{X} \to \mathbb{R}$ induces the Bayes classifier:

$$\operatorname{sgn} f^*(x) \equiv \begin{cases} 1, & \text{if } \eta(x) > 1/2 \\ -1, & \text{otherwise} \end{cases} \tag{8.56}$$

THEOREM 8.9. *Let $P = P_{XY}$ be the joint distribution of the feature $X \in \mathbb{R}^d$ and the binary label $Y \in \{-1,+1\}$, and let $\eta(x) = \mathbf{P}[Y = 1|X = x]$ be the corresponding regression function. Consider a penalty function φ, which is strictly convex and differentiable. Then the unique minimizer of the surrogate loss $A_\varphi(f) = \mathbf{E}[\varphi(-Yf(X))]$ over all (measurable) functions $f : \mathsf{X} \to \mathbb{R}$ has the form*

$$f^*(x) = \arg\min_{u \in \mathbb{R}} h_{\eta(x)}(u),$$

where for each $\eta \in [0,1]$ we have $h_\eta(u) := \eta\varphi(-u) + (1-\eta)\varphi(u)$. Moreover, $f^*(x)$ is positive if and only if $\eta(x) > 1/2$, i.e, the induced sign classifier $g_{f^*}(x) = \operatorname{sgn}(f^*(x))$ is the Bayes classifier (1.2).

PROOF. By the law of iterated expectation,

$$A_\varphi(f) = \mathbf{E}\left[\varphi(-Yf(X))\right] = \mathbf{E}[\mathbf{E}[\varphi(-Yf(X))|X]].$$

Hence,

$$\begin{aligned}\inf_f A_\varphi(f) &= \inf_f \mathbf{E}\left[\mathbf{E}\left[\varphi(-Yf(X))|X\right]\right] \\ &= \mathbf{E}\left[\inf_{u\in\mathbb{R}} \mathbf{E}[\varphi(-Yu)|X=x]\right].\end{aligned}$$

For every $x \in \mathsf{X}$, we have

$$\begin{aligned}\mathbf{E}[\varphi(-Yu)|X=x] &= \mathbf{P}[Y=1|X=x]\varphi(-u) + \mathbf{P}[Y=-1|X=x]\varphi(u) \\ &= \eta(x)\varphi(-u) + (1-\eta(x))\varphi(u) \\ &\equiv h_{\eta(x)}(u).\end{aligned}$$

Since φ is strictly convex and differentiable, so is h_η for every $\eta \in [0,1]$. Therefore, $\inf_{u\in\mathbb{R}} h_\eta(u)$ exists, and is achieved by a unique u^*; in particular,

$$f^*(x) = \operatorname*{arg\,min}_{u\in\mathbb{R}} h_{\eta(x)}(u).$$

To find the u^* minimizing h_η, we differentiate h_η w.r.t. u and set the derivative to zero. Since

$$h'_\eta(u) = -\eta\varphi'(-u) + (1-\eta)\varphi'(u),$$

the point of minimum u^* is the solution to the equation

$$\frac{\varphi'(u)}{\varphi'(-u)} = \frac{\eta}{1-\eta}.$$

Suppose $\eta > 1/2$; then

$$\frac{\varphi'(u)}{\varphi'(-u)} > 1.$$

Since φ is strictly convex, its derivative φ' is strictly increasing. Hence, $u^* > -u^*$ which implies that $u^* > 0$. Conversely, if $u^* \le 0$, then $u^* \le -u^*$, so $\varphi'(u^*) \le \varphi'(-u^*)$, which means that $\eta/(1-\eta) \le 1$, i.e., $\eta \le 1/2$. Thus, we conclude that $f^*(x)$, which is the minimizer of $h_{\eta(x)}$, is positive if and only if $\eta(x) > 1/2$, i.e., $\operatorname{sgn}(f^*(x))$ is the Bayes classifier. $\square$

Secondly, under some additional regularity conditions it is possible to relate the minimum surrogate loss

$$A^*_\varphi := \inf_f A_\varphi(f)$$

to the Bayes rate

$$L^* = \inf_f \mathbf{P}(Y \neq \operatorname{sgn} f(X)),$$

where in both expressions the infimum is over *all* measurable functions $f : \mathsf{X} \to \mathbb{R}$:

THEOREM 8.10. *Assume that the penalty function φ satisfies the usual conditions of our basic surrogate bound, and that there exist positive constants $s \ge 1$ and c, such that the inequality*

$$L(f) - L^* \le c\left(A_\varphi(f) - A^*_\varphi\right)^{1/s} \tag{8.57}$$

holds for any measurable function $f : \mathsf{X} \to \mathbb{R}$. Consider the learning algorithm that minimizes empirical surrogate risk over some class $\mathcal{F}$:

$$\widehat{f}_n = \operatorname*{arg\,min}_{f\in\mathcal{F}} A_{\varphi,n}(f) = \operatorname*{arg\,min}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varphi(-Y_i f(X_i)). \tag{8.58}$$

Then

$$L(\widehat{f}_n) - L^* \le 2^{1/s} c\left(4M_\varphi \mathbf{E}R_n(\mathcal{F}(X^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}}\right)^{1/s} + c\left(\inf_{f\in\mathcal{F}} A_\varphi(f) - A^*_\varphi\right)^{1/s} \tag{8.59}$$

with probability at least $1-\delta$.

PROOF. We have the following:

$$\begin{aligned}
L(\widehat{f}_n) - L^* &\le c\left(A_\varphi(\widehat{f}_n) - A^*_\varphi\right)^{1/s} && (8.60)\\
&= c\left(A_\varphi(\widehat{f}_n) - \inf_{f\in\mathcal{F}} A_\varphi(f) + \inf_{f\in\mathcal{F}} A_\varphi(f) - A^*_\varphi\right)^{1/s} && (8.61)\\
&\le c\left(A_\varphi(\widehat{f}_n) - \inf_{f\in\mathcal{F}} A_\varphi(f)\right)^{1/s} + c\left(\inf_{f\in\mathcal{F}} A_\varphi(f) - A^*_\varphi\right)^{1/s} && (8.62)\\
&\le 2^{1/s} c\left(\sup_{f\in\mathcal{F}} |A_{\varphi,n}(f) - A_\varphi(f)|\right)^{1/s} + c\left(\inf_{f\in\mathcal{F}} A_\varphi(f) - A^*_\varphi\right)^{1/s} && (8.63)\\
&\le 2^{1/s} c\left(4M_\varphi \mathbf{E}R_n(\mathcal{F}(X^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}}\right)^{1/s} \\
&\qquad + c\left(\inf_{f\in\mathcal{F}} A_\varphi(f) - A^*_\varphi\right)^{1/s} \qquad \text{w.p. } \ge 1-\delta, && (8.64)
\end{aligned}$$

where:

- (8.60) follows from (8.57);
- (8.62) follows from the inequality $(a+b)^{1/s} \le a^{1/s} + b^{1/s}$ that holds for all $a, b \ge 0$ and all $s \ge 1$
- (8.63) and (8.64) follow from the same argument as the one used in the proof of the basic surrogate bound.

This completes the proof. □

Remark 8.5. Condition (8.57) is often easy to check. For instance, Zhang [**Zha04**] proved that it is satisfied, provided the inequality

$$\left|\frac{1}{2}-\eta\right|^s \leq (2c)^s \left(1-\inf_u h_\eta(u)\right) \tag{8.65}$$

holds for all $\eta \in [0,1]$. For instance, (8.65) holds for the exponential loss $\varphi(u) = e^u$ and the logit loss $\varphi(u) = \log_2(1+e^u)$ with $s = 2$ and $c = 2\sqrt{2}$; for the hinge loss $\varphi(u) = (u+1)_+$, (8.65) holds with $s = 1$ and $c = 4$.

What Theorem 8.10 says is that, assuming the expected Rademacher average $\mathbf{E}R_n(\mathcal{F}(X^n)) = O(1/\sqrt{n})$, the difference between the generalization error of the Convex Risk Minimization algorithm (8.58) and the Bayes rate L^* is, with high probability, bounded by the combination of two terms: the $O(n^{-1/2s})$ "estimation error" term and the $(\inf_{f\in\mathcal{F}} A_\varphi(f) - A^*_\varphi)^{1/s}$ "approximation error" term. If the hypothesis space $\mathcal{F}$ is rich enough, so that $\inf_{f\in\mathcal{F}} A_\varphi(f) = A^*_\varphi$, then the difference between $L(\widehat{f}_n)$ and L^* is, with high probability, bounded as $O(1/n^{-2s})$, *independently* of the dimension d of the feature space.

CHAPTER 9

Regression with quadratic loss

Regression with quadratic loss is another basic problem studied in statistical learning theory. We have a random couple $Z = (X, Y)$, where, X is an $\mathbb{R}^d$-valued feature vector (or input vector) and Y is the *real-valued* response (or output). We assume that the unknown joint distribution $P = P_Z = P_{XY}$ of (X, Y) belongs to some class $\mathcal{P}$ of probability distributions over $\mathbb{R}^d \times \mathbb{R}$. The learning problem, then, is to produce a *predictor* of Y given X on the basis of an i.i.d. training sample $Z^n = (Z_1, \ldots, Z_n) = ((X_1, Y_1), \ldots, (X_n, Y_n))$ from P. A predictor is just a (measurable) function $f : \mathbb{R}^d \to \mathbb{R}$, and we evaluate its performance by the *expected quadratic loss*

$$L(f) := \mathbf{E}[(Y - f(X))^2].$$

As we have seen before, the smallest expected loss is achieved by the *regression function* $f^*(x) := \mathbf{E}[Y|X = x]$, i.e.,

$$L^* := \inf_f L(f) = L(f^*) = \mathbf{E}[(Y - \mathbf{E}[Y|X])^2].$$

Moreover, for any other f we have

$$L(f) = L^* + \|f - f^*\|^2_{L^2(P_X)},$$

where

$$\|f - f^*\|^2_{L^2(P_X)} = \int_{\mathbb{R}^d} |f(x) - f^*(x)|^2 P_X(dx).$$

An observation we have made many times by now is that when the joint distribution of the input-output pair $(X, Y) \in \mathsf{X} \times \mathbb{R}$ is unknown, there is no hope in general to learn the optimal predictor f^* from a finite training sample. So, as before, instead we aim to find a good approximation to the best predictor in some class $\mathcal{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$, i.e., to use the training data Z^n to construct a predictor $\widehat{f}_n \in \mathcal{F}$, such that

$$L(\widehat{f}_n) \approx L^*(\mathcal{F}) := \inf_{f \in \mathcal{F}} L(f)$$

with high probability. Restricting our attention to some hypothesis space $\mathcal{F}$, which is a proper subset of the class of *all* measurable functions $f : \mathsf{X} \to \mathbb{R}$, is a form of *insurance*: If we do not do this, then we can find some function f that attains zero empirical risk (if no two samples have same X but different Y's) , yet performs spectacularly badly on the inputs outside the training set. When this happens, we say that our learned predictor *overfits*. If $\mathcal{F}$ consists of well-behaved functions, then it is possible to learn a predictor that achieves a graceful balance between in-sample data fit and out-of-sample generalization. The price we pay is the *approximation error*

$$L^*(\mathcal{F}) - L^* \equiv \inf_{f \in \mathcal{F}} L(f) - \inf_{f : \mathsf{X} \to \mathbb{R}} L(f) \geq 0.$$

In the regression setting with mean square error, the approximation error is given by:

$$L^*(\mathcal{F}) - L^* = \inf_{f \in \mathcal{F}} \|f - f^*\|^2_{P_X},$$

where $f^*(x) = \mathbf{E}[Y|X = x]$ is the regression function (the MMSE predictor of Y given X).

We will assume that the marginal distribution P_X of the feature vector is supported on a closed subset $\mathsf{X} \subseteq \mathbb{R}^d$, and that the joint distribution P of (X, Y) is such that, with probability one,

$$|Y| \le M \tag{9.1}$$

for some constant $0 < M < \infty$. Thus we can assume that the training samples belong to the set $\mathsf{Z} = \mathsf{X} \times [-M, M]$. We also assume that the class $\mathcal{F}$ is a subset of a suitable reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ induced by some Mercer kernel $K : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$, such that C_K, defined by

$$C_K := \sup_{x \in \mathsf{X}} \sqrt{K(x, x)}, \tag{9.2}$$

is finite. By Lemma 4.1, for any $f \in \mathcal{H}_K$, $\|f\|_\infty := \sup_{x \in \mathsf{X}} |f(x)| \le C_K \|f\|_K$.

9.1. Constraint regularized least squares in RKHS

First, we will look at the simplest case: ERM over a ball in $\mathcal{H}_K$. Thus, we pick the radius $\lambda > 0$ and take

$$\mathcal{F} = \mathcal{F}_\lambda = \{f \in \mathcal{H}_K : \|f\|_K \le \lambda\}.$$

We have introduced the inequality constraint $\|f\|_K \le \lambda$ because in many cases, only assuming $f \in \mathcal{H}_K$ is not a strong enough regularization assumption. The ERM algorithm outputs the predictor

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}_\lambda} L_n(f) \equiv \arg\min_{f \in \mathcal{F}_\lambda} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2,$$

where $L_n(f)$ denotes, as usual, the empirical risk (in this case, empirical average quadratic loss) of f.

Theorem 9.1. *With probability at least $1 - \delta$, the ERM algorithm output $\widehat{f}$ satisfies:*

$$L(\widehat{f}_n) \le L^*(\mathcal{F}_\lambda) + \frac{16(M + C_K\lambda)^2}{\sqrt{n}} + (M^2 + C_K^2\lambda^2)\sqrt{\frac{8\log(1/\delta)}{n}} \tag{9.3}$$

Proof. First let us introduce some notation. Let us denote the quadratic loss function $(y, u) \mapsto (y - u)^2$ by $\ell(y, u)$, and for any $f : \mathbb{R}^d \to \mathbb{R}$ let

$$\ell \bullet f(x, y) := \ell(y, f(x)) = (y - f(x))^2$$

Let $\ell \bullet \mathcal{F}_\lambda$ denote the function class $\{\ell \bullet f : f \in \mathcal{F}_\lambda\}$.

Let f^*_λ denote any minimizer of $L(f)$ over $\mathcal{F}_\lambda$, i.e., $L(f^*_\lambda) = L^*(\mathcal{F}_\lambda)$. As usual, we write

$$\begin{aligned}
L(\widehat{f}_n) - L^*(\mathcal{F}_\lambda) &= L(\widehat{f}_n) - L^*(\mathcal{F}_\lambda) \\
&= L(\widehat{f}_n) - L_n(\widehat{f}_n) + L_n(\widehat{f}_n) - L_n(f^*_\lambda) + L_n(f^*_\lambda) - L(f^*_\lambda) \\
&\le 2 \sup_{f \in \mathcal{F}_\lambda} |L_n(f) - L(f)| \\
&= 2 \sup_{f \in \mathcal{F}_\lambda} |P_n(\ell \bullet f) - P(\ell \bullet f)| \\
&= 2\Delta_n(\ell \bullet \mathcal{F}_\lambda), \qquad (9.4)
\end{aligned}$$

where we have defined the uniform deviation

$$\Delta_n(\ell \bullet \mathcal{F}_\lambda) := \sup_{f \in \mathcal{F}_\lambda} |P_n(\ell \bullet f) - P(\ell \bullet f)|.$$

Note that (9.4) can also be viewed as a consequence of the double version of the mismatched minimization lemma, Lemma 5.1. Next we show that, as a function of the training sample Z^n, $g(Z^n) = \Delta_n(\ell \bullet \mathcal{F}_\lambda)$ has bounded differences. Indeed, for any $1 \le i \le n$, any $z^n \in \mathsf{Z}^n$, and any $z'_i \in \mathsf{Z}$, let $z^n_{(i)}$ denote z^n with the ith coordinate replaced by z'_i. Then

$$\begin{aligned}
\left|g(z^n) - g(z^n_{(i)})\right| &\le \frac{1}{n} \sup_{f \in \mathcal{F}_\lambda} \left|(y_i - f(x_i))^2 - (y'_i - f(x'_i))^2\right| \\
&\le \frac{1}{n} \sup_{x \in \mathsf{X}} \sup_{|y| \le M} \sup_{f \in \mathcal{F}_\lambda} |y - f(x)|^2 \\
&\overset{(a)}{\le} \frac{2}{n} \left(M^2 + \sup_{f \in \mathcal{F}_\lambda} \|f\|_\infty^2 \right) \\
&\overset{(b)}{\le} \frac{2}{n} \left(M^2 + C_K^2 \lambda^2 \right),
\end{aligned}$$

where (a) holds by the fact $(a+b)^2 \le 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$, and (b) holds by Lemma 4.1. Thus, $\Delta_n(\ell \bullet \mathcal{F}_\lambda)$ has the bounded difference property with $c_1 = \ldots = c_n = 2(M^2 + C_K^2\lambda^2)/n$, so McDiarmid's inequality says that, for any $t > 0$,

$$\mathbf{P}\Big(\Delta_n(\ell \bullet \mathcal{F}_\lambda) \ge \mathbf{E}\Delta_n(\ell \bullet \mathcal{F}_\lambda) + t\Big) \le \exp\left(- \frac{nt^2}{2(M^2 + C_K^2\lambda^2)^2} \right).$$

Therefore, letting

$$t = (M^2 + C_K^2\lambda^2)\sqrt{\frac{2\log(1/\delta)}{n}},$$

we see that

$$\Delta_n(\ell \bullet \mathcal{F}_\lambda) \le \mathbf{E}\Delta_n(\ell \bullet \mathcal{F}_\lambda) + (M^2 + C_K^2\lambda^2)\sqrt{\frac{2\log(1/\delta)}{n}} \qquad (9.5)$$

with probability at least $1 - \delta$. Moreover, by symmetrization (i.e. Theorem 6.1), we have

$$\mathbf{E}\Delta_n(\ell \bullet \mathcal{F}_\lambda) \le 2\mathbf{E}R_n(\ell \bullet \mathcal{F}_\lambda(Z^n)), \qquad (9.6)$$

where

$$R_n(\ell \bullet \mathcal{F}_\lambda(Z^n)) = \frac{1}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}_\lambda}\left|\sum_{i=1}^n \varepsilon_i \cdot \ell \bullet f(Z_i)\right|\right]$$

is the Rademacher average of the (random) set

$$\begin{aligned}\ell \bullet \mathcal{F}_\lambda(Z^n) &= \{(\ell \bullet f(Z_1), \dots, \ell \bullet f(Z_n)) : f \in \mathcal{F}_\lambda\} \\ &= \left\{((Y_1 - f(X_1)^2), \dots, (Y_n - f(X_n))^2) : f \in \mathcal{F}_\lambda\right\}.\end{aligned}$$

To bound the Rademacher average, we will use the contraction principle. To that end, consider the function $\varphi(t) = t^2$. For $A > 0$, $|\varphi'(t)| \le 2A$ on the interval $[-A, A]$, so φ is Lipschitz continuous on $[-A, A]$ with Lipschitz constant $2A$, i.e.,

$$|s^2 - t^2| \le 2A|s - t|, \qquad -A \le s, t \le A.$$

The fact $|Y_i| \le M$ and $|f(X_i)| \le C_K\lambda$ implies $|Y_i - f(X_i)| \le M + C_K\lambda$ for all $1 \le i \le n$, so taking $A = M + C_K\lambda$ and using the contraction principle yields:

$$R_n(\ell \bullet \mathcal{F}_\lambda(Z^n)) \le \frac{4(M + C_K\lambda)}{n}\mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}_\lambda}\left|\sum_{i=1}^n \varepsilon_i\left(Y_i - f(X_i)\right)\right|\right]. \tag{9.7}$$

Moreover

$$\begin{aligned}\mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}_\lambda}\left|\sum_{i=1}^n \varepsilon_i\left(Y_i - f(X_i)\right)\right|\right] &\overset{(a)}{\le} \mathbf{E}_{\varepsilon^n}\left|\sum_{i=1}^n \varepsilon_i Y_i\right| + \mathbf{E}_{\varepsilon^n}\left[\sup_{f\in\mathcal{F}_\lambda}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \\ &\overset{(b)}{\le} \sqrt{\sum_{i=1}^n Y_i^2} + nR_n(\mathcal{F}_\lambda(X^n)) \\ &\overset{(c)}{\le} (M + C_K\lambda)\sqrt{n}, \end{aligned} \tag{9.8}$$

where (a) holds by the triangle inequality, (b) holds Jensen's inequality and the definition of Rademacher average, and (c) holds by the assumption $|Y| \le M$ and the bound on the Rademacher average for the projection of a ball in an RKHS onto a set of n samples, given in Proposition 8.1.

Combining (9.5) through (9.8), we conclude that

$$\Delta_n(\ell \bullet \mathcal{F}_\lambda) \le \frac{8(M + C_K\lambda)^2}{\sqrt{n}} + (M^2 + C_K^2\lambda^2)\sqrt{\frac{2\log(1/\delta)}{n}} \tag{9.9}$$

with probability at least $1 - \delta$. Finally, combining this with (9.4), we get (9.3). $\square$

9.2. Penalty regularized least squares in an RKHS

The use of the constraint $\|f\|_{\mathcal{H}_K} \le \lambda$ in the previous section is a form of *regularization* — a way of guaranteeing that the learned predictor performs well outside the training sample. This section illustrates a closely related method: *penalty regularization*. The idea is to use an additive penalty term with some parameter γ in the cost function instead of the hard constraint. To illustrate the method we derive a variation of Theorem 9.1. The learning problem is denoted by $(\mathsf{X} = \mathbb{R}^d, \mathsf{Y} = [-M, M], \mathcal{P}, \mathcal{F} = \mathcal{H}_K, \ell(y, u) = (y - u)^2)$ such that $\mathcal{P}$ is a set of probability measures for random variables in $\mathbb{R}^d \times \mathbb{R}$. We again consider predictors $f \in \mathcal{H}_K$ where $\mathcal{H}_K$ is the RKHS generated by some Mercer kernel K. Given training data $Z^n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ consisting of n independent labeled samples with distribution P, a learning algorithm produces a predictor $f \in \mathcal{H}_K$. We use the following definitions.

- Generalization risk for predictor f: $L(f) = \mathbf{E}[(Y - f(X))^2]$
- Empirical risk for predictor f: $L_n(f) = \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2$
- Generalization regularized risk for predictor f: $J_\gamma(f) = L(f) + \gamma\|f\|_K^2$
- Empirical regularized risk for predictor f: $J_{n,\gamma}(f) = L_n(f) + \gamma\|f\|_K^2$
- Minimum risk for unconstrained predictors: $L^* = \mathbf{E}[(Y - \mathbf{E}[Y|X])^2]$
- Minimum regularized risk for class of predictors $\mathcal{H}_K$: $J_\gamma^*(\mathcal{H}_K) = \inf_{f\in\mathcal{H}_K} J_\gamma(f)$
- Increase in minimum risk due to regularization term: $A(\gamma) \triangleq J_\gamma^*(\mathcal{H}_K) - L^*$
- Regularized ERM predictor: $\widehat{f}_{n,\gamma} = \arg\min_{f\in\mathcal{F}} J_{n,\gamma}(f)$.

Theorem 9.2. *Consider the regression problem with quadratic loss,* $(\mathsf{X} = \mathbb{R}^d, \mathsf{Y} = [-M, M], \mathcal{P}, \mathcal{F} = \mathcal{H}_K, \ell(y, u) = (y - u)^2)$, *where* $\mathcal{H}_K$ *is the RKHS generated by some Mercer kernel* K *with* $C_K < \infty$. *For any* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$,

$$L(\widehat{f}_{n.\gamma}) - L^* \le A(\gamma) + \frac{16M^2\left(1 + \frac{C_K}{\sqrt{\gamma}}\right)^2}{\sqrt{n}} + M^2\left(1 + \frac{C_K^2}{\gamma}\right)\sqrt{\frac{8\log(1/\delta)}{n}}. \tag{9.10}$$

Proof. The infimum defining $J_\gamma^*(\mathcal{H}_K)$ can be restricted to $f \in \mathcal{F}_\lambda$, where $\lambda = \frac{M}{\sqrt{\gamma}}$, because if $\|f\|_K > \lambda$, then the identically zero predictor has a smaller regularized risk: $J_\gamma(f) \ge \gamma\|f\|_K^2 > M^2 \ge J_\gamma(0)$. For the same reason, the infimum defining the minimum regularized empirical risk, $J_{\gamma,n}(f)$, can also be restricted to $\mathcal{F}_\lambda$. In particular, $\widehat{f}_{n,\gamma} = \arg\min_{f\in\mathcal{F}_\lambda} J_{n,\gamma}(f)$.

By the double version of the mismatched minimization lemma, Lemma 5.1, $J_\gamma(\widehat{f}_{\gamma,n}) \le J_\gamma^*(\mathcal{H}_K) + 2\widetilde{\triangle}_{\lambda,\gamma,n}(Z^n)$ where $\widetilde{\triangle}_{\lambda,\gamma,n}(Z^n) \triangleq \sup_{f\in\mathcal{F}_\lambda} |J_{n,\gamma}(f) - J_\gamma(f)|$. However, both $J_\gamma(f)$ and $J_{n,\gamma}(f)$ include the additive term $\gamma\|f\|_K^2$, so that $\widetilde{\triangle}_{\lambda,\gamma,n}(Z^n) = \Delta_n(\ell \bullet \mathcal{F}_\lambda)$, where $\Delta_n(\ell \bullet \mathcal{F}_\lambda) = \sup_{f\in\mathcal{F}_\lambda} |L_n(f) - L(f)|$, as used in the proof of Theorem 9.1. The proof of Theorem 9.1 establishes that with probability at least $1 - \delta$,

$$2\Delta_n(\ell \bullet \mathcal{F}_\lambda) \le \frac{16(M + C_K\lambda)^2}{\sqrt{n}} + (M^2 + C_K^2\lambda^2)\sqrt{\frac{8\log(1/\delta)}{n}}.$$

Combining these observations with the trivial inequality $L(\widehat{f}_{n,\gamma}) \le J_\gamma(\widehat{f}_{n,\gamma})$, yields that with probability at least $1 - \delta$,

$$L(\widehat{f}_{n,\gamma}) \le J_\gamma^*(\mathcal{H}_K) + \frac{16(M + C_K\lambda)^2}{\sqrt{n}} + (M^2 + C_K^2\lambda^2)\sqrt{\frac{8\log(1/\delta)}{n}}. \tag{9.11}$$

Subtracting L^* from each side of (9.11) and making the substitution $\lambda^2 = \frac{M^2}{\gamma}$ yields (9.10). □

Remark 9.1. *The value of* λ *used for the purpose of the proof of Theorem 9.2 satisfies* $\lambda^2\gamma = M^2$. *If, instead,* λ *is given first, then the first part of the proof of Theorem 9.2 shows that if* γ *is such that* $\lambda^2\gamma \ge M^2$ *then the solution of the regularized ERM problem in this section meets the constraint of Section 9.1. Typically, however, smaller values of* γ *would give solutions such that the constraint of Section 9.1 is satisfied with near equality.*

Part 3

Some Applications

CHAPTER 10

Empirical vector quantization

Now that we have safely made our way through the combinatorial forests of Vapnik–Chervonenkis classes, we will look at an interesting application of the VC theory to a problem in communications engineering: empirical design of vector quantizers. Vector quantization is a technique for *lossy data compression* (or *source coding*), so we will first review, at a very brisk pace, the basics of source coding, and then get to business. The presentation will closely follow an excellent survey by Tamás Linder [**Lin01**].

10.1. A brief introduction to source coding

It's trite but true: we live in a digital world. We store, exchange, and manipulate vast quantities of binary data. While a lot of the data are inherently discrete (e.g., text), most are *compressed representations* of continuous-valued (analog) sources, such as audio, speech, images, or video. The process of mapping source data from their "native" format to binary representations and back is known in the information theory and the communications engineering communities as *source coding*.

There are two types of source coding: *lossless* and *lossy*. The former pertains to constructing compact binary representations of discrete data, such as text, and the objective is to map any sequence of symbols emitted by the source of interest into a binary file which is as short as possible and which will permit *exact* (i.e., error-free) reconstruction (decompression) of the data. The latter, on the other hand, deals with continuous-valued sources (such as images), and the objective is to map any source realization to a compact binary representation that would, upon decompression, differ from the original source as little as possible. We will focus on lossy source coding. Needless to say, we will only be able to give a very superficial overview of this rich subject. A survey article by Gray and Neuhoff [**GN98**] does a wonderful job of tracing both the historical development and the state of the art in lossy source coding; for an encyclopedic treatment, see the book by Gersho and Gray [**GG92**].

One of the simpler models of an analog source is a stationary stochastic process $Z_1, Z_2, \ldots$ with values in $\mathbb{R}^d$. For example, if d is a perfect square, then each Z_i could represent a $\sqrt{d} \times \sqrt{d}$ image patch. The compression process consists of two stages. First, each Z_i is mapped to a binary string b_i. Thus, the entire data stream $\{Z_i\}_{i=1}^{\infty}$ is represented by the sequence of binary strings $\{b_i\}_{i=1}^{\infty}$. The source data are reconstructed by mapping each b_i into a vector $\widehat{Z}_i \in \mathbb{R}^d$. Since each Z_i takes on a continuum of values, the mapping $Z_i \mapsto b_i$ is inherently *many-to-one*, i.e., noninvertible. This is the reason why this process is called *lossy* source coding — in going from the analog data $\{Z_i\}$ to the digital representation $\{b_i\}$ and then to the reconstruction $\widehat{Z}_i$, we lose information needed to recover each Z_i exactly. The overall mapping $Z_i \mapsto b_i \mapsto \widehat{Z}_i$ is called a *vector quantizer*, where the term "vector" refers to the vector-valued nature of the source $\{Z_i\}$, while the term "quantizer" indicates the process of representing a continuum by a discrete set. We assume that the mappings comprising the quantizer are time-invariant, i.e., do not depend on the time index $i \in \mathbb{N}$.

There are two figures of merit for a given quantizer: the compactness of the binary representation $Z_i \mapsto b_i$ and the accuracy of the reconstruction $b_i \mapsto \widehat{Z}_i$. The former is given by the *rate* of the quantizer, i.e., the expected length of b_i in bits. Since the source $\{Z_i\}$ is assumed to be stationary and the quantizer is assumed to be time-invariant, we have

$$\mathbf{E}[\mathrm{len}(b_i)] = \mathbf{E}[\mathrm{len}(b_1)], \qquad \forall i \in \mathbb{N},$$

where, for a binary string b, $\mathrm{len}(b)$ denotes its length in bits. If the length of $b_i \equiv b_i(Z_i)$ depends on Z_i, then we say that the quantizer is *variable-rate*; otherwise, we say that the quantizer is *fixed-rate*. The latter is how well the reconstruction $\widehat{Z}_i$ approximates the source Z_i on average. In order to measure that, we pick a nonnegative *distortion measure* $d : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$, so that $d(z, \widehat{z}) \geq 0$ quantifies how well one vector $z \in \mathbb{R}^d$ is approximated by another $\widehat{z} \in \mathbb{R}^d$. Then we look at the expected value $\mathbf{E}[d(Z_i, \widehat{Z}_i)]$, which is the same for all i, again owing to the stationarity of $\{Z_i\}$ and the time invariance of the quantizer. A typical distortion measure is the squared Euclidean norm

$$d(z, \widehat{z}) = \|z - \widehat{z}\|^2 = \sum_{j=1}^{d} |z(j) - \widehat{z}(j)|^2,$$

where $z(j)$ denotes the jth coordinate of z. We will focus only on this distortion measure from now on.

Now, using the fact that the rate and the expected distortion of a quantizer do not depend on the time index i, we can just consider the problem of quantizing a single $\mathbb{R}^d$-valued random variable Z with the same distribution as that of Z_1. From now on, we will refer to such a Z as the source. Thus, the rate of a given quantizer $Z \mapsto b \mapsto \widehat{Z}$ is given by $\mathbf{E}[\mathrm{len}(b)]$ and the expected distortion $\mathbf{E}\|Z - \widehat{Z}\|^2$. Naturally, one would like to keep both of these as low as possible: low rate means that it will take less memory space to store the compressed data and that it will be possible to transmit the compressed data over low-capacity digital channels; low expected distortion means that the reconstructed source will be a very accurate approximation of the true source. However, these two quantities are in conflict: if we make the rate too low, we will be certain to incur a lot of loss in reconstructing the data; if we insist on very accurate reconstruction, the binary representation must use a large number of bits. For this reason, the natural question is as follows: what is the smallest distortion achievable on a given source by any quantizer with a given rate?

10.2. Fixed-rate vector quantization

Let $\mathsf{Z} = \mathbb{R}^d$.

DEFINITION 10.1. *Let $k \in \mathbb{N}$. A (d-dimensional) k*-point vector quantizer *is a (measurable) mapping $q : \mathsf{Z} \to \mathcal{C} = \{y_1, \ldots, y_k\} \subset \mathsf{Z}$, where the set $\mathcal{C}$ is called the* codebook *and its elements are called the* codevectors.

The source is a random vector $Z \in \mathbb{R}^d$ with some probability distribution P_Z. A given k-point quantizer q represents Z by the quantized output $\widehat{Z} = q(Z)$. Since $q(Z)$ can take only k possible values, it is possible to represent it uniquely by a binary string of $\lceil \log_2 k \rceil$ bits. The number

$$R(q) := \lceil \log_2 k \rceil$$

is called the *rate* of q (in bits), where we follow standard practice and ignore the integer constraint on the length of the binary representation. The rate is often normalized by the dimension d to give $r(q) = d^{-1}R(q)$ (measured in bits per coordinate); however, since we assume d fixed, there is no need to worry about the normalization. The fidelity of q in representing $Z \sim P_Z$ is measured by the *expected distortion*

$$D(P_Z, q) := \mathbf{E}\|Z - q(Z)\|^2 = \int_{\mathbb{R}^d} \|z - q(z)\|^2 P_Z(dz).$$

We assume throughout that Z has finite second moment, $\mathbf{E}\|Z\|^2 < \infty$, so $D(P_Z, q) < \infty$.

The main objective in vector quantization is to minimize the expected distortion subject to a constraint on the rate (or, equivalently, on the codebook size). Thus, if we denote by $\mathcal{Q}_k$ the set of all k-point vector quantizers, then the optimal performance on a given source distribution P_Z is defined by

$$D_k^*(P_Z) := \inf_{q \in \mathcal{Q}_k} D(P_Z, q) \equiv \inf_{q \in \mathcal{Q}_k} \mathbf{E}\|Z - q(Z)\|^2. \tag{10.1}$$

Definition 10.2. *We say that a quantizer $q^* \in \mathcal{Q}_k$ is* optimal *for P_Z if*

$$D(P_Z, q^*) = D_k^*(P_Z).$$

As we will soon see, it turns out that an optimal quantizer always exists — in other words, the infimum in (10.1) is actually a minimum — and it can always be chosen to have a particularly useful structural property:

Definition 10.3. *A quantizer $q \in \mathcal{Q}_k$ with codebook $\mathcal{C} = \{y_1, \ldots, y_k\}$ is called* nearest-neighbor *if, for all $z \in \mathsf{Z}$,*

$$\|z - q(z)\|^2 = \min_{1 \le j \le k} \|z - y_j\|^2.$$

Let $\mathcal{Q}_k^{\mathrm{NN}}$ denote the set of all k-point nearest-neighbor quantizers. We have the following simple but important result:

Lemma 10.1. *For any $q \in \mathcal{Q}_k$ we can always find some $q' \in \mathcal{Q}_k^{\mathrm{NN}}$, such that $D(P_Z, q') \le D(P_Z, q)$.*

Proof. Given a quantizer $q \in \mathcal{Q}_k$ with codebook $\mathcal{C} = \{y_1, \ldots, y_k\}$, define q' by

$$q'(z) := \operatorname*{arg\,min}_{y_j \in \mathcal{C}} \|z - y_j\|^2,$$

where ties are broken by going with the lowest index. Then q' is clearly a nearest-neighbor quantizer, and

$$\begin{aligned}
D(P_Z, q') &= \mathbf{E}\|Z - q'(Z)\|^2 \\
&= \mathbf{E}\left[\min_{1 \le j \le k} \|Z - y_j\|^2\right] \\
&\le \mathbf{E}\|Z - q(Z)\|^2 \\
&\equiv D(P_Z, q).
\end{aligned}$$

The lemma is proved. $\square$

In light of this lemma, we can rewrite (10.1) as

$$D_k^*(P_Z) = \inf_{q \in \mathcal{Q}_k^{\mathrm{NN}}} \mathbf{E}\|Z - q(Z)\|^2 = \inf_{\mathcal{C}=\{y_1,\dots,y_k\}\subset \mathsf{Z}} \mathbf{E}\left[\min_{1\le j\le k} \|Z - y_j\|^2\right]. \tag{10.2}$$

An important result due to Pollard [**Pol82**], which we state here without proof, then says the following:

Theorem 10.1. *If Z has a finite second moment, $\mathbf{E}\|Z\|^2 < \infty$, then there exists a nearest-neighbor quantizer $q^* \in \mathcal{Q}_k^{\mathrm{NN}}$ such that $D(P_Z, q^*) = D_k^*(P_Z)$.*

10.3. Learning an optimal quantizer

Unfortunately, finding an optimal q^* is a very difficult problem. Indeed, the optimization problem in (10.2) has a *combinatorial search* component to it, since we have to optimize over all k-point sets $\mathcal{C}$ in $\mathbb{R}^d$. Moreover, the source distribution P_Z is often not known exactly, especially for very complex sources, such as natural images. For these reasons, we have to resort to *empirical* methods for quantizer design, which rely on the availability of a large number of independent samples from the source distribution of interest.

Assuming that such samples are easily available, we can formulate the empirical quantizer design problem as follows. Let us fix the desired codebook size k. For each $n \in \mathbb{N}$, let $Z^n = (Z_1, \dots, Z_n)$ be an i.i.d. sample from P_Z. We seek an algorithm that would take Z^n and produce a quantizer $\widehat{q}_n \in \mathcal{Q}_k$ that would approximate, as closely as possible, an optimal quantizer $q^* \in \mathcal{Q}_k$ that achieves $D_k^*(P_Z)$. In other words, we hope to *learn* an (approximately) optimal quantizer for P_Z based on a sufficiently long training sample.

The first thing to note is that the theory of quantization outlined in the preceding section applies to the *empirical distribution* of the training sample Z^n,

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}.$$

In particular, given a quantizer $q \in \mathcal{Q}_k$, we can compute its expected distortion

$$D(P_n, q) = \mathbf{E}_{P_n}\|Z - q(Z)\|^2 = \frac{1}{n}\sum_{i=1}^n \|Z_i - q(Z_i)\|^2.$$

Moreover, the minimum achievable distortion is given by

$$D_k^*(P_n) = \min_{q \in \mathcal{Q}_k} \frac{1}{n}\sum_{i=1}^n \|Z_i - q(Z_i)\|^2 = \min_{q \in \mathcal{Q}_k^{\mathrm{NN}}} \|Z_i - q(Z_i)\|^2.$$

Note that we have replaced the infimum with the minimum, since an optimal quantizer always exists and can be assumed to have the nearest-neighbor property. Moreover, since P_n is a discrete distribution, the existence of an optimal nearest-neighbor quantizer can be proved directly, without recourse to Pollard's theorem. Thus, we can restrict our attention to nearest-neighbor k-point quantizers.

Definition 10.4. *We say that a quantizer* $\widehat{q}_n \in \mathcal{Q}_k^{\mathrm{NN}}$ *is* empirically optimal *for* Z^n *if*

$$D(P_n, \widehat{q}_n) = D_k^*(P_n) = \min_{q \in \mathcal{Q}_k^{\mathrm{NN}}} D(P_n, q) = \min_{q \in \mathcal{Q}_k^{\mathrm{NN}}} \frac{1}{n} \sum_{i=1}^n \|Z_i - q(Z_i)\|^2.$$

Note that, by the nearest-neighbor property,

$$D_k^*(P_n) = \min_{\mathcal{C}=\{y_1,\ldots,y_k\} \subset \mathsf{Z}} \frac{1}{n} \sum_{i=1}^n \min_{1 \le j \le k} \|Z_i - y_j\|^2.$$

Thus, let $\widehat{q}_n \in \mathcal{Q}_k^{\mathrm{NN}}$ be an empirically optimal nearest-neighbor quantizer. Let $Z \sim P_Z$ be a new source realization, independent of the training data Z^n. If we apply $\widehat{q}_n$ to Z, the resulting quantized output $\widehat{Z} = \widehat{q}_n(Z)$ will depend on both the input Z and on the training data Z^n. Moreover, the expected distortion of $\widehat{q}_n$, given by

$$D(P_Z, \widehat{q}_n) = \mathbf{E}\Big[\|Z - \widehat{q}_n(Z)\|^2 \Big| Z^n\Big] = \int_{\mathsf{Z}} \|z - \widehat{q}_n(z)\|^2 P_Z(dz),$$

is a random variable, since it depends (through $\widehat{q}_n$) on the training data Z^n. In the next section we will show that, under certain assumptions on the source P_Z, the empirically optimal quantizer $\widehat{q}_n$ is nearly optimal on P_Z as well, in the sense that

$$\mathbf{E}\left[D(P_Z, \widehat{f}_n) - D_k^*(P_Z)\right] \le \frac{C}{\sqrt{n}}, \tag{10.3}$$

where the expectation is w.r.t. the distribution Z^n and $C > 0$ is some constant that depends on d, k, and a certain characteristic of P_Z. More generally, it is possible to show that empirically optimal quantizers are *strongly consistent* in the sense that

$$D(P_Z, \widehat{q}_n) - D_k^*(P_Z) \xrightarrow{n \to \infty} 0 \qquad \text{almost surely}$$

provided the source P_Z has a finite second moment (see Linder's survey [**Lin01**] for details).

Remark 10.1. It should be pointed out that the problem of finding an *exact* minimizer of $D(P_n, q)$ over $q \in \mathcal{Q}_k^{\mathrm{NN}}$ is NP-complete. Instead, various approximation techniques are used. The most popular one is the Lloyd algorithm, known in the computer science community as the method of k-means. There, one starts with an initial codebook $\mathcal{C}^{(0)} = \{y_1^{(0)}, \ldots, y_k^{(0)}\}$ and then iteratively recomputes the quantizer partition and the new codevectors until convergence.

10.4. Finite sample bound for empirically optimal quantizers

In this section, we will show how the VC theory can be used to establish (10.3) for any source supported on a ball of finite radius. This result was proved by Linder, Lugosi and Zeger [**LLZ94**], and since then refined and extended by multiple authors. Some recent works even remove the requirement that Z be finite-dimensional and consider more general coding schemes in Hilbert spaces [**MP10**].

For a given $r > 0$ and $z \in \mathbb{R}^d$, let $B_r(z)$ denote the ℓ_2 ball of radius r centered at z:

$$B_r(z) := \left\{ y \in \mathbb{R}^d : \|y - z\| \le r \right\}.$$

Let $\mathcal{P}(r)$ denote the set of all probability distributions P_Z on $\mathsf{Z} = \mathbb{R}^d$, such that

$$P_Z\left(B_r(0)\right) = 1.$$

Here is the main result we will prove in this section:

THEOREM 10.2. *There exists some absolute constant $C > 0$, such that*

$$\sup_{P_Z \in \mathcal{P}(r)} \mathbf{E}\left[D(P_Z, \widehat{q}_n) - D_k^*(P_Z)\right] \le Cr^2 \sqrt{\frac{k(d+1)\log(k(d+1))}{n}}.$$

Here, as before, $\widehat{q}_n$ denotes an empirically optimal quantizer based on an i.i.d. sample Z^n.

Before launching into the proof, we state and prove a useful lemma:

LEMMA 10.2. *Let $\mathcal{Q}_k^{\rm NN}(r)$ denote the set of all nearest-neighbor k-point quantizers whose codewords lie in $B_r(0)$. Then for any $P_Z \in \mathcal{P}(r)$,*

$$D(P_Z, \widehat{q}_n) - D_k^*(P_Z) \le 2 \sup_{q \in \mathcal{Q}_k^{\rm NN}(r)} \left| D(P_n, q) - D(P_Z, q) \right|.$$

PROOF. Fix P_Z and let $q^* \in \mathcal{Q}_k^{\rm NN}$ denote an optimal quantizer, i.e., $D(P_Z, q^*) = D_k^*(P_Z)$. Then, using our old trick of adding and subtracting the right empirical quantities, we can write

$$D(P_Z, \widehat{q}_n) - D_k^*(P_Z) = D(P_Z, \widehat{q}_n) - D(P_n, \widehat{q}_n) + D(P_n, \widehat{q}_n) - D(P_n, q^*) + D(P_n, q^*) - D(P_Z, q^*).$$

Since $\widehat{q}_n$ minimizes the empirical distortion $D(P_n, q)$ over all $q \in \mathcal{Q}_k^{\rm NN}$, we have $D(P_n, \widehat{q}_n) \le D(P_n, q^*)$, which leads to

$$D(P_Z, \widehat{q}_n) - D_k^*(P_Z) \le D(P_Z, \widehat{q}_n) - D(P_n, \widehat{q}_n) + D(P_n, q^*) - D(P_Z, q^*). \tag{10.4}$$

Now, since $B_r(0)$ is a convex set, for any point $y \notin B_r(0)$ we can compute its *projection* y' onto $B_r(0)$, namely $y' = ry/\|y\|$. Then y' is strictly closer to all $z \in B_r(0)$ than y, i.e.,

$$\|z - y'\| < \|z - y\|, \qquad \forall z \in B_r(0).$$

Thus, if we take an arbitrary quantizer $q \in \mathcal{Q}_k$ and replace all of its codevectors outside $B_r(0)$ by their projections, we will obtain another quantizer q', such that $\|z - q'(z)\| \le \|z - q(z)\|$ for all $z \in B_r(0)$. (The $\le$ sign is due to the fact that some of the codevectors of q may already be in $B_r(0)$, so the projection will not affect them). But then for any $P_Z \in \mathcal{P}(r)$ we will have $D(P_Z, q') \le D(P_Z, q)$. Moreover, if Z^n is an i.i.d. sample from P_Z and P_n is the corresponding empirical distribution, then $P_n \in \mathcal{P}(r)$ with probability one. Hence, we can assume that both $\widehat{q}_n$ and q^* have all their codevectors in $B_r(0)$, and therefore from (10.4) we obtain

$$\begin{aligned} D(P_Z, \widehat{q}_n) - D_k^*(P_Z) &\le D(P_Z, \widehat{q}_n) - D(P_n, \widehat{q}_n) + D(P_n, q^*) - D(P_Z, q^*) \\ &\le \left| D(P_n, \widehat{q}_n) - D(P_Z, \widehat{q}_n) \right| + \left| D(P_Z, q^*) - D(P_n, q^*) \right| \\ &\le 2 \sup_{q \in \mathcal{Q}_k^{\rm NN}(r)} \left| D(P_n, q) - D(P_Z, q) \right|. \end{aligned}$$

This finishes the proof. $\square$

Now we can get down to business:

PROOF (OF THEOREM 10.2). For a given quantizer $q \in \mathcal{Q}_k^{\rm NN}(r)$, define the function

$$f_q(z) := \|z - q(z)\|^2,$$

which is just the squared Euclidean distortion between z and $q(z)$. In particular, for any $P \in \mathcal{P}(r)$ the expected distortion $D(P, q)$ is equal to $P(f_q)$. Since $q \in \mathcal{Q}_k^{\mathrm{NN}}(r)$, we have $\|q(z)\| \le r$ for all z. Therefore, for any $z \in B_r(0)$ we will have

$$0 \le f_q(z) \le 2\|z\|^2 + 2\|q(z)\|^2 \le 4r^2.$$

Therefore, using the fact that the expectation of any nonnegative random variable U can be written as

$$\mathbf{E}U = \int_0^\infty \mathbf{P}(U > u)du,$$

we can write

$$D(P_Z, q) = P_Z(f_q) = \int_0^{4r^2} P_Z(f_q(Z) > u)du$$

and

$$D(P_n, q) = P_n(f_q) = \int_0^{4r^2} P_n(f_q(Z) > u)du = \int_0^{4r^2} \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{f_q(Z_i)>u\}}du \qquad \text{a.s.}$$

Therefore

$$\begin{aligned}
&\sup_{q \in \mathcal{Q}_k^{\mathrm{NN}}(r)} |D(P_n, q) - D(P_Z, q)| \\
&\quad = \sup_{q \in \mathcal{Q}_k^{\mathrm{NN}}(r)} |P_n(q) - P_Z(q)| \\
&\quad = \sup_{q \in \mathcal{Q}_k^{\mathrm{NN}}(r)} \left| \int_0^{4r^2} \left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{f_q(Z_i)>u\}} - P_Z(f_q(Z) > u) \right) du \right| \\
&\quad \le 4r^2 \sup_{q \in \mathcal{Q}_k^{\mathrm{NN}}(r)} \sup_{0 \le u \le 4r^2} \left| \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{f_q(Z_i)>u\}} - P_Z(f_q(Z) > u) \right| \qquad \text{a.s.}
\end{aligned} \tag{10.5}$$

where the last step uses the fact that

$$\int_a^b h(u)du \le |b - a| \sup_{a \le u \le b} |h(u)|.$$

Now, for a given $q \in \mathcal{Q}_k^{\mathrm{NN}}(r)$ and a given $u > 0$ let us define the set

$$A_{u,q} := \left\{ z \in \mathbb{R}^d : f_q(z) > u \right\},$$

and let $\mathcal{A}$ denote the class of all such sets: $\mathcal{A} := \{A_{u,q} : u > 0, q \in \mathcal{Q}_k^{\mathrm{NN}}(r)\}$. Then $\mathbf{1}_{\{f_q(z)>u\}} = \mathbf{1}_{\{z \in A_{u,q}\}}$, so from (10.5) we can write

$$\sup_{q \in \mathcal{Q}_k^{\mathrm{NN}}(r)} |D(P_n, q) - D(P_Z, q)| \le 4r^2 \sup_{A \in \mathcal{A}} |P_n(A) - P_Z(A)|. \tag{10.6}$$

Therefore,

$$\mathbf{E}\left[D(P_Z, \widehat{q}_n) - D_k^*(P_Z)\right] \leq 2\mathbf{E}\left[\sup_{q \in \mathcal{Q}_k^{\mathrm{NN}}(r)} |D(P_n, q) - D(P_Z, q)|\right]$$
$$\leq 8r^2 \mathbf{E}\left[\sup_{A \in \mathcal{A}} |P_n(A) - P_Z(A)|\right],$$

where the first step follows from Lemma 10.2 and the second step follows from (10.6). To finish the proof, we will show that $\mathcal{A}$ is a VC class with $V(\mathcal{A}) \leq 4k(d+1)\log(k(d+1))$, so that

$$\mathbf{E}\left[\sup_{A \in \mathcal{A}} |P_n(A) - P_Z(A)|\right] \stackrel{(a)}{\leq} C\sqrt{\frac{V(\mathcal{A})}{n}} \leq 2C\sqrt{\frac{k(d+1)\log(k(d+1))}{n}},$$

where (a) follows from Theorem 6.1 (based on symmetrization trick) and Theorem 7.2 (Dudley's theorem using chaining technique). In order to bound the VC dimension of $\mathcal{A}$, let us consider a typical set $A_{u,q}$. Let $\{y_1, \ldots, y_k\}$ denote the codevectors of q. Since q is a nearest-neighbor quantizer, a point z will be in $A_{u,q}$ if and only if

$$f_q(z) = \min_{1 \leq j \leq k} \|z - y_j\|^2 > u,$$

which is equivalent to

$$\|z - y_j\| > \sqrt{u}, \qquad \forall 1 \leq j \leq k.$$

In other words, we can write

$$A_{u,q} = \bigcap_{j=1}^{k} B_{\sqrt{u}}(y_j)^c.$$

Since this can be done for every $u > 0$ and every $q \in \mathcal{Q}_k^{\mathrm{NN}}(r)$, we conclude that the class $\mathcal{A}$ is contained in another class $\widetilde{\mathcal{A}}$, defined by

$$\widetilde{\mathcal{A}} := \left\{\bigcap_{j=1}^{k} B_j^c : B_j \in \mathcal{B}, \forall j\right\},$$

where $\mathcal{B}$ denotes the class of all closed balls in $\mathbb{R}^d$. Therefore, $V(\mathcal{A}) \leq V(\widetilde{\mathcal{A}})$. To bound $V(\widetilde{\mathcal{A}})$, we must examine its shatter coefficients. We will need the following facts[1]:

(1) For any class of sets $\mathcal{M}$, let $\overline{\mathcal{M}}$ denote the class $\{M^c : M \in \mathcal{M}\}$ formed by taking the complements of all sets in $\mathcal{M}$. Then for any n

$$\mathbb{S}_n(\overline{\mathcal{M}}) = \mathbb{S}_n(\mathcal{M}).$$

(2) For any class of sets $\mathcal{N}$, let $\mathcal{N}_k$ denote the class $\{N_1 \cap N_2 \cap \ldots \cap N_k : N_j \in \mathcal{N}, 1 \leq j \leq k\}$, formed by taking intersections of all possible choices of k sets from $\mathcal{N}$. Then

$$\mathbb{S}_n(\mathcal{N}_k) \leq \mathbb{S}_n^k(\mathcal{N}).$$

[1]Exercise: prove them!

In the above notation, $\widetilde{\mathcal{A}} = (\overline{\mathcal{B}})_k$, so

$$\mathbb{S}_n(\widetilde{\mathcal{A}}) \leq \mathbb{S}_n^k(\mathcal{B}),$$

where $\mathcal{B}$ is the class of all closed balls in $\mathbb{R}^d$. Section 7.2.5, based on Dudley classses of binary classifiers, shows that $V(\mathcal{B}) = d + 1$, and so the Sauer–Shelah lemma gives

$$\mathbb{S}_n(\widetilde{\mathcal{A}}) \leq \left(\frac{ne}{d+1}\right)^{k(d+1)}, \qquad \text{for } n \geq d+1. \tag{10.7}$$

We can now upper-bound $V(\widetilde{\mathcal{A}})$ by finding an n for which the right-hand side of (10.7) is less than 2^n. It is easy to check that, for $d \geq 2$, $n = 4k(d+1)\log(k(d+1))$ does the job; for $d = 1$ it's clear that $V(\widetilde{\mathcal{A}}) \leq 2k$. Thus,

$$V(\mathcal{A}) \leq V(\widetilde{\mathcal{A}}) \leq 4k(d+1)\log(k(d+1)),$$

as claimed. The proof is finished. $\square$

CHAPTER 11

Dimensionality reduction in Hilbert spaces

Dimensionality reduction is a generic name for any procedure that takes a complicated object living in a high-dimensional (or possibly even infinite-dimensional) space and approximates it in some sense by a finite-dimensional vector. We are interested in a particular class of dimensionality reduction methods. Consider a data source that generates vectors in some Hilbert space $\mathcal{H}$, which is either infinite-dimensional or has a finite but extremely large dimension (think $\mathbb{R}^d$ with the usual Euclidean norm, where d is huge). We will assume that the vectors of interest lie in the unit ball of $\mathcal{H}$,

$$B(\mathcal{H}) := \Big\{ x \in \mathcal{H} : \|x\| \le 1 \Big\},$$

where $\|x\| = \sqrt{\langle x, x\rangle}$ is the norm on $\mathcal{H}$. We wish to represent each $x \in B(\mathcal{H})$ by a vector $\widehat{y} \in \mathbb{R}^k$ for some fixed k (if $\mathcal{H}$ is d-dimensional, then of course we must have $d \gg k$). For instance, k may represent some storage limitation, such as a device that can store no more than k real numbers (or, more realistically, k double-precision floating-point numbers, which for all practical purposes can be thought of as real numbers). The mapping $x \mapsto \widehat{y}$ can be thought of as an *encoding* rule. In addition, given $\widehat{y} \in \mathbb{R}^k$, we need a *decoding* rule that takes $\widehat{y}$ and outputs a vector $\widehat{x} \in \mathcal{H}$ that will serve as an approximation of x. In general, the cascade of mappings

$$x \xrightarrow{\text{encoding}} \widehat{y} \xrightarrow{\text{decoding}} \widehat{x}$$

will be lossy, i.e., $x \neq \widehat{x}$. So, the goal is to ensure that the squared norm error $\|x - \widehat{x}\|^2$ is as small as possible. In this lecture, we will see how Rademacher complexity techniques can be used to characterize the performance of a particular fairly broad class of dimensionality reduction schemes in Hilbert spaces. Our exposition here is based on a beautiful recent paper of Maurer and Pontil [**MP10**].

We will consider a particular type of dimensionality reduction schemes, where the encoder is a (nonlinear) projection, whereas the decoder is a linear operator from $\mathbb{R}^k$ into $\mathcal{H}$ (the Appendix contains some basic facts pertaining to linear operators between Hilbert spaces). To specify such a scheme, we fix a pair (Y, T) consisting of a closed set $\mathsf{Y} \subseteq \mathbb{R}^k$ and a linear operator $T : \mathbb{R}^k \to \mathcal{H}$. We call Y the *codebook* and use the encoding rule

$$\widehat{y} = \arg\min_{y \in \mathsf{Y}} \|x - Ty\|^2. \tag{11.1}$$

Unless Y is a closed subspace of $\mathbb{R}^k$, this encoding map will be nonlinear. The decoding, on the other hand, is linear: $\widehat{x} = T\widehat{y}$. With these definitions, the reconstruction error is given by

$$\|x - \widehat{x}\|^2 = f_T(x) := \min_{y \in \mathsf{Y}} \|x - Ty\|^2.$$

Now suppose that the input to our dimensionality reduction scheme is a *random vector* $X \in B(\mathcal{H})$ with some unknown distribution P. Then we measure the performance of the coding scheme (Y, T) by its expected reconstruction error

$$L(T) := \mathbf{E}_P[f_T(X)] \equiv \mathbf{E}_P\left[\min_{y \in \mathsf{Y}} \|X - Ty\|^2\right]$$

(note that, even though the reconstruction error depends on the codebook Y, we do not explicitly indicate this dependence, since the choice of Y will be fixed by a particular application). Now let $\mathcal{T}$ be some fixed class of admissible linear decoding maps $T : \mathbb{R}^k \to \mathcal{H}$. So, if we knew P, we could find the best decoder $\widetilde{T} \in \mathcal{T}$ that achieves

$$L^*(\mathcal{T}) := \inf_{T \in \mathcal{T}} L(T)$$

(assuming, of course, that the infimum exists and is achieved by at least one $T \in \mathcal{T}$).

By now, you know the drill: We don't know P, but we have access to a large set of samples $X_1, \ldots, X_n$ drawn i.i.d. from P. So we attempt to learn $\widetilde{T}$ via ERM:

$$\begin{aligned}\widehat{T}_n &:= \arg\min_{T \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^n f_T(X_i) \\ &= \arg\min_{T \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^n \min_{y \in \mathsf{Y}} \|X_i - Ty\|^2.\end{aligned}$$

Our goal is to establish the following result:

THEOREM 11.1. *Assume that Y is a closed subset of the unit ball $B_2^k = \{y \in \mathbb{R}^k : \|y\|_2 \le 1\}$, and that every $T \in \mathcal{T}$ satisfies*

$$\begin{aligned}\|Te_j\| &\le \alpha, \qquad 1 \le j \le k \\ \|T\|_{\mathsf{Y}} &:= \sup_{y \in \mathsf{Y},\, y \neq 0} \|Ty\| \le \alpha\end{aligned}$$

for some finite $\alpha \ge 1$, where $e_1, \ldots, e_k$ is the standard basis of $\mathbb{R}^k$. Then

$$L(\widehat{T}_n) \le L^*(\mathcal{T}) + \frac{60\alpha^2 k^2}{\sqrt{n}} + 4\alpha^2 \sqrt{\frac{2\log(1/\delta)}{n}} \tag{11.2}$$

with probability at lest $1 - \delta$. In the special case when $\mathsf{Y} = \{e_1, \ldots, e_k\}$, the standard basis in $\mathbb{R}^k$, the event

$$L(\widehat{T}_n) \le L^*(\mathcal{T}) + \frac{40\alpha^2 k}{\sqrt{n}} + 4\alpha^2 \sqrt{\frac{2\log(1/\delta)}{n}} \tag{11.3}$$

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

REMARK 11.1. The above result is slightly weaker than the one from [**MP10**]; as a consequence, the constants in Eqs. (11.2) and (11.3) are slightly worse than they could otherwise be.

11.1. Examples

Before we get down to business and prove the theorem, let's look at a few examples.

11.1.1. Principal component analysis (PCA). The objective of PCA is, given k, construct a projection Π onto a k-dimensional closed subspace of $\mathcal{H}$ to maximize the average "energy content" of the projected vector:

$$\begin{aligned} &\text{maximize } \mathbf{E}\|\Pi X\|^2 \\ &\text{subject to } \dim \Pi(\mathcal{H}) = k \\ &\qquad\qquad \Pi^2 = \Pi \end{aligned} \tag{11.4}$$

For any $x \in \mathcal{H}$,

$$\|\Pi x\|^2 = \|x\|^2 - \|(I - \Pi)x\|^2, \tag{11.5}$$

where I is the identity operator on $\mathcal{H}$. To prove (11.5), expand the right-hand side:

$$\begin{aligned} \|x\|^2 - \|(I-\Pi)x\|^2 &= \|x\|^2 - \|x - \Pi x\|^2 \\ &= 2\langle x, \Pi x\rangle - \|\Pi x\|^2 \\ &= \|\Pi x\|^2, \end{aligned}$$

where the last step is by the properties of projections. Thus,

$$\begin{aligned} \|\Pi x\|^2 &= \|x\|^2 - \|x - \Pi x\|^2 \\ &= \|x\|^2 - \min_{x' \in \mathcal{K}} \|x - x'\|^2, \end{aligned} \tag{11.6}$$

where $\mathcal{K}$ is the range of Π (the closure of the linear span of all vectors of the form $\Pi x, x \in \mathcal{H}$). Moreover, any projection operator $\Pi : \mathcal{H} \to \mathcal{K}$ with $\dim(\mathcal{H}) = k$ can be factored as TT^*, where $T : \mathbb{R}^k \to \mathcal{H}$ is an isometry (see Appendix for definitions and the proof of this fact). Using this fact, we can write

$$\mathcal{K} = \Pi(\mathcal{H}) = \left\{ Ty : y \in \mathbb{R}^k \right\}.$$

Using this in (11.6), we get

$$\|\Pi x\|^2 = \|x\|^2 - \min_{y \in \mathbb{R}^k} \|x - Ty\|^2.$$

Hence, solving the optimization problem (11.4) is equivalent to finding the best linear decoding map $\widetilde{T}$ for the pair $(\mathsf{Y}, \mathcal{T})$, where $\mathsf{Y} = \mathbb{R}^k$ and $\mathcal{T}$ is the collection of all isometries $T : \mathbb{R}^k \to \mathcal{H}$. Moreover, if we recall our assumption that $X \in B(\mathcal{H})$ with probability one, then we see that there is no loss of generality if we take

$$\mathsf{Y} = B_2^k := \left\{ y \in \mathbb{R}^k : \|y\|_2 \le 1 \right\},$$

i.e., the unit ball in $(\mathbb{R}^k, \|\cdot\|_2)$. This follows from the fact that $\|\Pi x\| \le \|x\|$ for any projection Π, so, in particular, for $\Pi = TT^*$ the encoding $\widehat{y}$ in (11.1) can be written as $\widehat{y} = T^* x$, and

$$\|\widehat{y}\|_2 = \|T\widehat{y}\| = \|TT^*x\| = \|\Pi x\| \le \|x\| \le 1.$$

Thus, Theorem 11.1 applies with $\alpha = 1$. That said, there are much tighter bounds for PCA that rely on deeper structural results pertaining to finite-dimensional subspaces of Hilbert spaces, but that is beside the point. The key idea here is that we can already get nice bounds using the tools already at our fingertips.

11.1.2. Vector quantization or k-means clustering. Vector quantization (or k-means clustering) is a procedure that takes a vector $x \in \mathcal{H}$ and maps it to its nearest neighbor in a finite set $\mathcal{C} = \{\xi_1, \dots, \xi_k\} \subset \mathcal{H}$, where k is a given positive integer:

$$\widehat{x} = \arg\min_{\xi \in \mathcal{C}} \|x - \xi\|^2.$$

If X is random with distribution P, then the optimal k-point quantizer is given a size-k set $\widetilde{\mathcal{C}} = \{\widetilde{\xi}_1, \dots, \widetilde{\xi}_k\}$ that minimizes the reconstruction error

$$\mathbf{E}_P \left[\min_{\xi \in \mathcal{C}} \|X - \xi\|^2 \right]$$

over all $\mathcal{C} \subset \mathcal{H}$ with $|\mathcal{C}| = k$. We can cast the problem of finding $\widetilde{\mathcal{C}}$ in our framework by taking $\mathsf{Y} = \{e_1, \dots, e_k\}$ (the standard basis in $\mathbb{R}^k$) and letting $\mathcal{T}$ be the set of all linear operators $T : \mathbb{R}^k \to \mathcal{H}$. It is easy to see that any $\mathcal{C} \subset \mathcal{H}$ with $|\mathcal{C}| = k$ can be obtained as an image of the standard basis $\{e_1, \dots, e_k\}$ under some linear operator $T : \mathbb{R}^k \to \mathcal{H}$. Indeed, for any $\mathcal{C} = \{\xi_1, \dots, \xi_k\}$, we can just *define* a linear operator $T : \mathbb{R}^k \to \mathcal{H}$ by

$$Te_j := \xi_j, \qquad 1 \le j \le k$$

and then extending it to all of $\mathbb{R}^k$ by linearity:

$$T \left(\sum_{j=1}^k y_j e_j \right) = \sum_{j=1}^k y_j T e_j = \sum_{j=1}^k y_j \xi_j.$$

So, another way to interpret the objective of vector quantization is as follows: given a distribution P supported on $B(\mathcal{H})$, we seek a k-element set $\mathcal{C} = \{\xi_1, \dots, \xi_k\} \subset \mathcal{H}$, such that the random vector $X \sim P$ can be well-approximated on average by linear combinations of the form

$$\sum_{j=1}^k y_j \xi_j,$$

where the vector of coefficients $y = (y_1, \dots, y_k)$ can have only one nonzero component, which is furthermore required to be equal to 1. In fact, there is no loss of generality in assuming that $\mathcal{C} \subset B(\mathcal{H})$ as well. This is a consequence of the fact that, for any $x \in B(\mathcal{H})$ and any $x' \in \mathcal{H}$, we can always find some $x'' \in B(\mathcal{H})$ such that

$$\|x - x''\| \le \|x - x'\|.$$

Indeed, it suffices to take $x'' = \arg\min_{z \in B(\mathcal{H})} \|x' - z\|^2$, and it is not hard to show that $x'' = x'/\|x'\|$.

Thus, Theorem 11.1 applies with $\alpha = 1$. Moreover, the excess risk grows *linearly* with dimension k, cf. Eq. (11.3). It is not known whether this linear dependence on k is optimal — there are $\Omega(\sqrt{k/n})$ lower bounds for vector quantization, but it is still an open question whether these lower bounds are tight [**MP10**].

11.1.3. Nonnegative matrix factorization. Consider approximating the random vector $X \sim P$, where P is supported on the unit ball $B(\mathcal{H})$, by linear combinations of the form

$$\sum_{j=1}^{k} y_j \xi_j,$$

where the real vector $y = (y_1, \dots, y_k)$ is constrained to lie in the nonnegative orthant

$$\mathbb{R}^k_+ := \left\{ y = (y_1, \dots, y_k) \in \mathbb{R}^k : y_j \ge 0, 1 \le j \le k \right\},$$

while the unit vectors $\xi_1, \dots, \xi_k \in B(\mathcal{H})$ are constrained by the positivity condition

$$\langle \xi_j, \xi_\ell \rangle_{\mathcal{H}} \ge 0, \qquad 1 \le j, \ell \le k.$$

This is a generalization of the *nonnegative matrix factorization* (NMF) problem, originally posed by Lee and Seung [**LS99**].

To cast NMF in our framework, let $\mathsf{Y} = \mathbb{R}^k_+$, and let $\mathcal{T}$ be the set of all linear operators $T : \mathbb{R}^k \to \mathcal{H}$ such that (i) $\|Te_j\| = 1$ for all $1 \le j \le k$ and (ii) $\langle Te_j, Te_\ell \rangle \ge 0$ for all $1 \le j, \ell \le k$. Then the choice of T is equivalent to the choice of $\xi_1, \dots, \xi_k \in B(\mathcal{H})$, as above. Moreover, it can be shown that, for any $x \in B(\mathcal{H})$ and any $T \in \mathcal{T}$, the minimum of $\|x - Ty\|^2$ over all $y \in \mathbb{R}^k_+$ is achieved at some $\widehat{y} \in \mathbb{R}^k_+$ with $\|\widehat{y}\|_2 \le 1$. Thus, there is no loss of generality if we take $\mathsf{Y} = \mathbb{R}^k_+ \cap B^k_2$. In this case, the conditions of Theorem 11.1 are satisfied with $\alpha = 1$.

11.1.4. Sparse coding. Take Y to be the ℓ_1 unit ball

$$B^k_1 := \left\{ y = (y_1, \dots, y_k) \in \mathbb{R}^k : \|y\|_1 = \sum_{j=1}^{k} |y_j| \le 1 \right\},$$

and let $\mathcal{T}$ be the collection of all linear operators $T : \mathbb{R}^k \to \mathcal{H}$ with $\|Te_j\| \le 1$ for all $1 \le j \le k$. In this case, the dimensionality reduction problem is to approximate a random $X \in B(\mathcal{H})$ by a linear combination of the form

$$\sum_{j=1}^{k} y_j \xi_j,$$

where $y = (y_1, \dots, y_k) \in \mathbb{R}^k$ satisfies the constraint $\|y\|_1 \le 1$, while the vectors $\xi_1, \dots, \xi_k$ belong to the unit ball $B(\mathcal{H})$. Then for any $y = \sum_{j=1}^{k} y_j e_j \in \mathsf{Y}$ we have

$$\begin{aligned} \|Ty\| &= \left\| \sum_{j=1}^{k} y_j Te_j \right\| \\ &\le \sum_{j=1}^{k} |y_j| \|Te_j\| \\ &\le \|y\|_1 \cdot \max_{1 \le j \le k} \|Te_j\| \\ &\le 1, \end{aligned}$$

where the third line is by Hölder's inequality. Then the conditions of Theorem 11.1 are satisfied with $\alpha = 1$.

11.2. Proof of Theorem 11.1

Now we turn to the proof of Theorem 11.1. The format of the proof is the familiar one: if we consider the empirical reconstruction error

$$\begin{aligned} L_n(T) &:= \frac{1}{n}\sum_{i=1}^{n} f_T(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \min_{y \in \mathsf{Y}} \|X_i - Ty\|^2 \end{aligned}$$

for every $T \in \mathfrak{T}$ and define the uniform deviation

$$\Delta_n(X^n) := \sup_{T \in \mathfrak{T}} |L_n(T) - L(T)|, \tag{11.7}$$

then

$$L(\widehat{T}_n) \le L^*(\mathfrak{T}) + 2\Delta_n(X^n).$$

Now, for any $x \in B(\mathcal{H})$, any $y \in \mathsf{Y}$, and any $T \in \mathfrak{T}$, we have

$$0 \le \|x - Ty\|^2 \le 2\|x\|^2 + 2\|Ty\|^2 \le 4\alpha^2.$$

Thus, the uniform deviation $\Delta_n(X^n)$ has bounded differences with $c_1 = \ldots = c_n = 4\alpha^2/n$, so by McDiarmid's inequality,

$$L(\widehat{T}_n) \le L^*(\mathfrak{T}) + 2\mathbf{E}\Delta_n(X^n) + 4\alpha^2\sqrt{\frac{2\log(1/\delta)}{n}}, \tag{11.8}$$

with probability at least $1-\delta$. By the usual symmetrization argument, we obtain the bound $\mathbf{E}\Delta_n(X^n) \le 2\mathbf{E}R_n(\mathcal{F}(X^n))$, where $\mathcal{F}$ is the class of functions f_T for all $T \in \mathfrak{T}$. Now, the whole affair hinges on getting a good upper bound on the Rademacher averages $R_n(\mathcal{F}(X^n))$. We will do this in several steps, and we need to introduce some additional machinery along the way.

11.2.1. Gaussian averages. Let $\gamma_1, \ldots, \gamma_n$ be i.i.d. standard normal random variables. In analogy to the Rademacher average of a bounded set $\mathcal{A} \subset \mathbb{R}^n$, we can define the *Gaussian average* of $\mathcal{A}$ [**BM02**] as

$$G_n(\mathcal{A}) := \mathbf{E}_{\gamma^n} \sup_{a \in \mathcal{A}} \left| \frac{1}{n}\sum_{i=1}^{n} \gamma_i a_i \right|.$$

Lemma 11.1 (Gaussian averages vs. Rademacher averages).

$$R_n(\mathcal{A}) \le \sqrt{\frac{\pi}{2}} G_n(\mathcal{A}). \tag{11.9}$$

PROOF. Let $\sigma^n = (\sigma_1, \ldots, \sigma_n)$ be an n-tuple of i.i.d. Rademacher random variables independent of γ^n. Since each γ_i is a symmetric random variable, it has the same distribution as $\sigma_i|\gamma_i|$. Therefore,

$$
\begin{aligned}
G_n(\mathcal{A}) &= \frac{1}{n}\mathbf{E}_{\gamma^n} \sup_{a\in\mathcal{A}} \left|\sum_{i=1}^n \gamma_i a_i\right| \\
&= \frac{1}{n}\mathbf{E}_{\sigma^n}\mathbf{E}_{\gamma^n} \sup_{a\in\mathcal{A}} \left|\sum_{i=1}^n \sigma_i|\gamma_i| a_i\right| \\
&\geq \frac{1}{n}\mathbf{E}_{\sigma^n} \sup_{a\in\mathcal{A}} \left|\sum_{i=1}^n \sigma_i a_i \mathbf{E}_{\gamma_i}|\gamma_i|\right| \\
&= \mathbf{E}|\gamma_1| \cdot \frac{1}{n}\mathbf{E}_{\sigma^n} \sup_{a\in\mathcal{A}} \left|\sum_{i=1}^n \sigma_i a_i\right| \\
&= \mathbf{E}|\gamma_1| R_n(\mathcal{A}),
\end{aligned}
$$

where the second step is by convexity, while in the last step we have used the fact that $\gamma_1, \ldots, \gamma_n$ are i.i.d. random variables. Now, if γ is a standard normal random variable, then

$$
\begin{aligned}
\mathbf{E}|\gamma| &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |t| e^{-t^2/2} \mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} t e^{-t^2/2} dt - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} t e^{-t^2/2} \mathrm{d}t \\
&= \sqrt{\frac{2}{\pi}} \int_0^{\infty} t e^{-t^2/2} \mathrm{d}t \\
&= \sqrt{\frac{2}{\pi}}.
\end{aligned}
$$

Rearranging, we get (11.9). □

Gaussian averages are often easier to work with than Rademacher averages. The reason for this is that, for any n real constants $a_1, \ldots, a_n$, the sum $W_a := a_1\gamma_1 + \ldots + a_n\gamma_n$ is a Gaussian random variable with mean 0 and variance $a_1^2 + \ldots + a_n^2$. Moreover, for any finite collection of vectors $a^{(1)}, \ldots, a^{(m)} \in \mathcal{A}$, the random variables $W_{a^{(1)}}, \ldots, W_{a^{(m)}}$ are jointly Gaussian. Thus, the collection of random variables $(W_a)_{a\in\mathcal{A}}$ is a zero-mean *Gaussian process*, where we say that a collection of real-valued random variables $(W_a)_{a\in\mathcal{A}}$ is a Gaussian process if all finite linear combinations of the W_a's are Gaussian random variables. In particular, we can compute covariances: for any $a, a' \in \mathcal{A}$,

$$\begin{aligned}
\mathbf{E}[W_a W_{a'}] &= \mathbf{E}\left[\sum_{i=1}^n \sum_{j=1}^n \gamma_i \gamma_j a_i a'_j\right] \\
&= \sum_{i=1}^n \sum_{j=1}^n \mathbf{E}[\gamma_i \gamma_j] a_i a'_j \\
&= \sum_{i=1}^n a_i a'_i \\
&= \langle a, a' \rangle
\end{aligned}$$

and things like

$$\begin{aligned}
\mathbf{E}[(W_a - W_{a'})^2] &= \mathbf{E}\left[\sum_{i=1}^n \sum_{j=1}^n \gamma_i \gamma_j (a_i - a'_i)(a_j - a'_j)\right] \\
&= \sum_{i=1}^n \sum_{j=1}^n \mathbf{E}\left[\gamma_i \gamma_j\right] (a_i - a'_i)(a_j - a'_j) \\
&= \sum_{i=1}^n (a_i - a'_i)^2 \\
&= \|a - a'\|^2.
\end{aligned}$$

The latter quantities are handy because of a very useful result called *Slepian's lemma* [**Sle62, LT91**]:

Lemma 11.2. *Let $(W_a)_{a\in\mathcal{A}}$ and $(V_a)_{a\in\mathcal{A}}$ be two zero-mean Gaussian processes with some index set $\mathcal{A}$ (not necessarily a subset of $\mathbb{R}^n$), such that*

$$\mathbf{E}[(W_a - W_{a'})^2] \le \mathbf{E}[(V_a - V_{a'})^2], \qquad \forall a, a' \in \mathcal{A}. \tag{11.10}$$

Then

$$\mathbf{E} \sup_{a\in\mathcal{A}} W_a \le \mathbf{E} \sup_{a\in\mathcal{A}} V_a. \tag{11.11}$$

Remark 11.2. The Gaussian processes $(W_a), (V_a)$ that appear in Slepian's lemma are not necessarily of the form $W_a = \langle a, \gamma^n \rangle$ with $\gamma^n = (\gamma_1, \dots, \gamma_n)$ a vector of independent Gaussians. They can be arbitrarily collections of random variables indexed by the elements of $\mathcal{A}$, such that any finite linear combination of W_a's or of V_a's is Gaussian.

Slepian's lemma is typically used to obtain upper bounds on the expected supremum of one Gaussian process in terms of another, which is hopefully easier to handle. The only wrinkle is that we can't apply Slepian's lemma to the problem of estimating the Gaussian average $G_n(\mathcal{A})$ because of the absolute value. However, if all $a \in \mathcal{A}$ are uniformly bounded in norm, the absolute value makes little difference:

LEMMA 11.3. *Let $\mathcal{A} \subset \mathbb{R}^n$ be a set of vectors uniformly bounded in norm, i.e., there exists some $L < \infty$ such that $\|a\| \le L$ for all $a \in \mathcal{A}$. Let*

$$\widetilde{G}_n(\mathcal{A}) := \frac{1}{n}\mathbf{E}\left[\sup_{a\in\mathcal{A}} \sum_{i=1}^n \gamma_i a_i\right]. \tag{11.12}$$

Then

$$\widetilde{G}_n(\mathcal{A}) \le G_n(\mathcal{A}) \le 2\widetilde{G}_n(\mathcal{A}) + \sqrt{\frac{2}{\pi}}\frac{L}{n}. \tag{11.13}$$

PROOF. The first inequality in (11.13) is obvious. For the second inequality, pick an arbitrary $a' \in \mathcal{A}$, let $W_a = \sum_{i=1}^n \gamma_i a_i$ for any $a \in \mathcal{A}$, and write

$$\begin{aligned} G_n(\mathcal{A}) &= \frac{1}{n}\mathbf{E}\left[\sup_{a\in\mathcal{A}} |W_a|\right] \\ &\le \frac{1}{n}\mathbf{E}\left[\sup_{a\in\mathcal{A}} |W_a - W_{a'}|\right] + \frac{1}{n}\mathbf{E}|W_{a'}|. \end{aligned}$$

Since a' was arbitrary, this gives

$$\begin{aligned} G_n(\mathcal{A}) &\le \sup_{a'\in\mathcal{A}}\left\{\frac{1}{n}\mathbf{E}\left[\sup_{a\in\mathcal{A}} |W_a - W_{a'}|\right] + \frac{1}{n}\mathbf{E}|W_{a'}|.\right\} \\ &\le \frac{1}{n}\mathbf{E}\left[\sup_{a,a'\in\mathcal{A}} |W_a - W_{a'}|\right] + \frac{1}{n}\sup_{a'\in\mathcal{A}} \mathbf{E}|W_{a'}|. \end{aligned} \tag{11.14}$$

For any two a, a', the random variable $W_a - W_{a'}$ is symmetric, so

$$\mathbf{E}\left[\sup_{a,a'\in\mathcal{A}} |W_a - W_{a'}|\right] = 2\,\mathbf{E}\left[\sup_{a\in\mathcal{A}} W_a\right].$$

Moreover, for any $a' \in \mathcal{A}$, $W_{a'}$ is Gaussian with zero mean and variance $\|a'\|^2 \le L^2$. Thus,

$$\sup_{a'\in\mathcal{A}} \mathbf{E}|W_{a'}| \le L\,\mathbf{E}|\gamma| = \sqrt{\frac{2}{\pi}}L.$$

Using the two above formulas in (11.14), we get the second inequality in (11.13), and the lemma is proved. $\square$

Armed with this lemma, we can work with the quantity $\widetilde{G}_n(\mathcal{A})$ instead of the Gaussian average $G_n(\mathcal{A})$. The advantage is that now we can rely on tools like Slepian's lemma.

11.2.2. Bounding the Rademacher average. Now everything hinges on bounding the Gaussian average $G_n(\mathcal{F}(x^n))$ for a fixed sample $x^n = (x_1, \ldots, x_n)$, which in turn will give us a bound on the Rademacher average $R_n(\mathcal{F}(x^n))$, by Lemmas 11.1 and 11.3. Let $(\gamma_i)_{1\le i\le n}$, $(\gamma_{ij})_{1\le i\le n, 1\le j\le k}$, and $(\gamma_{ij\ell})_{1\le i\le n, 1\le j,\ell\le k}$ be mutually independent sequences of i.i.d. standard Gaussian random variables. Define the following zero-mean Gaussian processes, indexed by

$T \in \mathfrak{T}$:

$$
\begin{aligned}
W_T &:= \sum_{i=1}^n \gamma_i f_T(x_i), \\
V_T &:= \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij} \langle x_i, Te_j \rangle, \\
U_T &:= \sum_{i=1}^n \sum_{j=1}^k \sum_{\ell=1}^k \gamma_{ij\ell} \langle Te_j, Te_\ell \rangle, \\
\Upsilon_T &:= \sqrt{8} V_T + \sqrt{2} U_T.
\end{aligned}
$$

By definition,

$$
\begin{aligned}
G_n(\mathcal{F}(x^n)) &= \mathbf{E} \sup_{T \in \mathfrak{T}} \left| \frac{1}{n} \sum_{i=1}^n \gamma_i f_T(x_i) \right| \\
&= \frac{1}{n} \mathbf{E} \sup_{T \in \mathfrak{T}} |W_T|,
\end{aligned}
$$

and we define $\widetilde{G}_n(\mathcal{F}(x^n))$ similarly. We will use Slepian's lemma to upper-bound $\widetilde{G}_n(\mathcal{F}(x^n))$ in terms of expected suprema of $(V_T)_{T \in \mathfrak{T}}$ and $(U_T)_{T \in \mathfrak{T}}$. To that end, we start with

$$
\begin{aligned}
\mathbf{E}\left[(W_T - W_{T'})^2\right] &= \sum_{i=1}^n \left(f_T(x_i) - f_{T'}(x_i)\right)^2 \\
&= \sum_{i=1}^n \left(\min_{y \in \mathsf{Y}} \|x_i - Ty\|^2 - \min_{y \in \mathsf{Y}} \|x_i - T'y\|^2 \right)^2 \\
&\le \sum_{i=1}^n \left(\max_{y \in \mathsf{Y}} \left| \|x_i - Ty\|^2 - \|x_i - T'y\|^2 \right| \right)^2 \\
&= \sum_{i=1}^n \left(\max_{y \in \mathsf{Y}} \left| 2\langle x_i, Ty - T'y \rangle + \|Ty\|^2 - \|T'y\|^2 \right| \right)^2 \\
&\le 8 \sum_{i=1}^n \max_{y \in \mathsf{Y}} \left| \langle x_i, Ty - T'y \rangle \right|^2 + 2 \sum_{i=1}^n \max_{y \in \mathsf{Y}} \left(\|Ty\|^2 - \|T'y\|^2 \right)^2, \qquad (11.15)
\end{aligned}
$$

where in the third line we have used properties of inner products, and the last line is by the inequality $(a+b)^2 \le 2a^2 + 2b^2$. Now, for each i,

$$\begin{aligned}
\max_{y \in \mathsf{Y}} |\langle x_i, Ty - T'y\rangle|^2 &= \max_{y \in \mathsf{Y}} \left| \sum_{j=1}^{k} y_j \langle x_i, Te_j - T'e_j\rangle \right|^2 \\
&\le \max_{y \in \mathsf{Y}} \|y\|_2^2 \sum_{j=1}^{k} |\langle x_i, Te_j - T'e_j\rangle|^2 \\
&\le \sum_{j=1}^{k} |\langle x_i, Te_j - T'e_j\rangle|^2 ,
\end{aligned}$$

where in the second step we have used Cauchy–Schwarz. Summing over $1 \le i \le n$, we see that

$$\begin{aligned}
\sum_{i=1}^{n} \max_{y \in \mathsf{Y}} |\langle x_i, Ty - T'y\rangle|^2 &\le \sum_{i=1}^{n} \sum_{j=1}^{k} |\langle x_i, Te_j - T'e_j\rangle|^2 \\
&= \mathbf{E}\left[(V_T - V_{T'})^2\right]. \qquad (11.16)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\max_{y \in \mathsf{Y}} \left(\|Ty\|^2 - \|T'y\|^2\right)^2 &= \max_{y \in \mathsf{Y}} \left(\sum_{j=1}^{k} \sum_{\ell=1}^{k} y_j y_\ell \langle Te_j, Te_\ell\rangle - \langle T'e_j, T'e_\ell\rangle \right)^2 \\
&\le \max_{y \in \mathsf{Y}} \sum_{j=1}^{k} \sum_{\ell=1}^{k} y_j^2 y_\ell^2 \cdot \sum_{j=1}^{k} \sum_{\ell=1}^{k} \left(\langle Te_j, Te_\ell\rangle - \langle T'e_j, T'e_\ell\rangle\right)^2 \\
&= \max_{y \in \mathsf{Y}} \|y\|_2^4 \cdot \sum_{j=1}^{k} \sum_{\ell=1}^{k} \left(\langle Te_j, Te_\ell\rangle - \langle T'e_j, T'e_\ell\rangle\right)^2 \\
&\le \sum_{j=1}^{k} \sum_{\ell=1}^{k} \left(\langle Te_j, Te_\ell\rangle - \langle T'e_j, T'e_\ell\rangle\right)^2 .
\end{aligned}$$

Therefore,

$$\sum_{i=1}^{n} \max_{y \in \mathsf{Y}} \left(\|Ty\|^2 - \|T'y\|^2\right) \le \mathbf{E}\left[(U_T - U_{T'})^2\right]. \qquad (11.17)$$

Using (11.16) and (11.17) in (11.15), we have

$$\begin{aligned}
\mathbf{E}\left[(W_T - W_{T'})^2\right] &\le 8\,\mathbf{E}\left[(V_T - V_{T'})^2\right] + 2\,\mathbf{E}\left[(U_T - U_{T'})^2\right] \\
&= \mathbf{E}\left[(\Upsilon_T - \Upsilon_{T'})^2\right].
\end{aligned}$$

We can therefore apply Slepian's lemma (Lemma 11.2) to $(W_T)_{T\in\mathfrak{T}}$ and $(\Upsilon_T)_{T\in\mathfrak{T}}$ to write

$$\begin{aligned}
\widetilde{G}_n(\mathcal{F}(x^n)) &= \frac{1}{n}\mathbf{E}\sup_{T\in\mathfrak{T}} W_T \\
&\le \frac{1}{n}\mathbf{E}\sup_{T\in\mathfrak{T}} \Upsilon_T \\
&\le \frac{\sqrt{8}}{n}\,\mathbf{E}\sup_{T\in\mathfrak{T}} V_T + \frac{\sqrt{2}}{n}\,\mathbf{E}\sup_{T\in\mathfrak{T}} U_T. \qquad (11.18)
\end{aligned}$$

We now upper-bound the expected suprema of V_T and U_T. For the former,

$$\begin{aligned}
\mathbf{E}\sup_{T\in\mathfrak{T}} V_T &= \mathbf{E}\sup_{T\in\mathfrak{T}} \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij}\langle x_i, Te_j\rangle \\
&= \mathbf{E}\sup_{T\in\mathfrak{T}} \sum_{j=1}^k \left\langle \sum_{i=1}^n \gamma_{ij} x_i, Te_j \right\rangle && \text{(linearity)} \\
&\le \mathbf{E}\sup_{T\in\mathfrak{T}} \sum_{j=1}^k \left\| \sum_{i=1}^n \gamma_{ij} x_i \right\| \|Te_j\| && \text{(Cauchy–Schwarz)} \\
&\le \mathbf{E} \sum_{j=1}^k \left\| \sum_{i=1}^n \gamma_{ij} x_i \right\| \sup_{T\in\mathfrak{T}} \|Te_j\| \\
&\le \alpha \sum_{j=1}^k \mathbf{E} \left\| \sum_{i=1}^n \gamma_{ij} x_i \right\| && \text{(assumption on } \|T\|\text{)} \\
&\le \alpha \sum_{j=1}^k \mathbf{E} \sqrt{\sum_{i=1}^n \sum_{i'=1}^n \gamma_{ij}\gamma_{i'j}\langle x_i, x_i'\rangle} && \text{(linearity)} \\
&\le \alpha \sum_{j=1}^k \sqrt{\sum_{i=1}^n \sum_{i'=1}^n \mathbf{E}\left[\gamma_{ij}\gamma_{i'j}\right] \langle x_i, x_i'\rangle} && \text{(Jensen)} \\
&= \alpha \sum_{j=1}^k \sqrt{\sum_{i=1}^n \|x_i\|^2} && \text{(properties of i.i.d. Gaussians)} \\
&\le \alpha k\sqrt{n}. && (x_i \in B(\mathcal{H}) \text{ for all } i)
\end{aligned}$$

Similarly, for the latter,

$$\begin{aligned}
\mathbf{E} \sup_{T \in \mathcal{T}} U_T &= \mathbf{E} \sup_{T \in \mathcal{T}} \sum_{i=1}^n \sum_{j=1}^k \sum_{\ell=1}^k \gamma_{ij\ell} \langle Te_j, Te_\ell \rangle \\
&\le \sum_{j=1}^k \sum_{\ell=1}^k \mathbf{E} \sup_{T \in \mathcal{T}} \sum_{i=1}^n \gamma_{ij\ell} \langle Te_j, Te_\ell \rangle \\
&\le \sum_{j=1}^k \sum_{\ell=1}^k \mathbf{E} \left| \sum_{i=1}^n \gamma_{ij\ell} \right| \sup_{T \in \mathcal{T}} \|Te_j\| \|Te_\ell\| \\
&\le \alpha^2 k^2 \sqrt{\frac{2n}{\pi}}.
\end{aligned}$$

Substituting these bounds into (11.18), we have

$$\widetilde{G}_n(\mathcal{F}(x^n)) \le \frac{1}{n} \left(\alpha k \sqrt{8n} + \alpha^2 k^2 \frac{2\sqrt{n}}{\sqrt{\pi}} \right) \le \frac{5\alpha^2 k^2}{\sqrt{n}}.$$

Thus, applying Lemmas 11.1 and 11.3, we have

$$\begin{aligned}
R_n(\mathcal{F}(x^n)) &\le \sqrt{\frac{\pi}{2}} G_n(\mathcal{F}(x^n)) \\
&\le \sqrt{\frac{\pi}{2}} \left[2\widetilde{G}_n(\mathcal{F}(x^n)) + \sqrt{\frac{2}{\pi}} \frac{\max_{T \in \mathcal{T}} \sqrt{\sum_{i=1}^n |f_T(x_i)|^2}}{n} \right] \\
&\le \sqrt{\frac{\pi}{2}} \left[\frac{10\alpha^2 k^2}{\sqrt{n}} + \sqrt{\frac{2}{\pi}} \frac{2\alpha}{\sqrt{n}} \right] \\
&\le \frac{15\alpha^2 k^2}{\sqrt{n}}
\end{aligned}$$

Recalling (11.8), we see that the event (11.2) holds with probability at least $1 - \delta$.

For the special case of k-means clustering, i.e., when $\mathsf{Y} = \{e_1, \dots, e_k\}$, we follow a slightly different strategy. Define a zero-mean Gaussian process

$$\Xi_T := \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij} \|x_i - Te_j\|^2, \qquad T \in \mathcal{T}.$$

Then

$$
\begin{aligned}
\mathbf{E}\left[(W_T - W_{T'})^2\right] &= \sum_{i=1}^n \left(\min_{1\le j\le k} \|x_i - Te_j\|^2 - \min_{1\le j\le k} \|_i - T'e_j\|^2 \right)^2 \\
&\le \sum_{i=1}^n \max_{1\le j\le k} \left(\|x_i - Te_j\|^2 - \|x_i - T'e_j\|^2 \right)^2 \\
&\le \sum_{i=1}^n \sum_{j=1}^k \left(\|x_i - Te_j\|^2 - \|x_i - T'e_j\|^2 \right)^2 \\
&= \mathbf{E}\left[(\Xi_T - \Xi_{T'})^2\right].
\end{aligned}
$$

For the process (Ξ_T), we have

$$
\begin{aligned}
\mathbf{E}\sup_{T\in\mathcal{T}} \Xi_T &= \mathbf{E}\sup_{T\in\mathcal{T}} \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij} \|x_i - Te_j\|^2 \\
&= \mathbf{E}\sup_{T\in\mathcal{T}} \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij} \left\{ \|x_i\|^2 - 2\langle x_i, Te_j\rangle + \|Te_j\|^2 \right\} \\
&\le \sum_{j=1}^k \mathbf{E}\sup_{T\in\mathcal{T}} \left\{ 2\sum_{i=1}^n \gamma_{ij} |\langle x_i, Te_j\rangle| + \sum_{i=1}^n \gamma_{ij} \|Te_j\|^2 \right\} \\
&\le 3k\alpha^2\sqrt{n},
\end{aligned}
$$

where the methods used to obtain this bound are similar to what we did for (V_T) and (U_T). Using Lemmas 11.1–11.3, we have

$$
\begin{aligned}
R_n(\mathcal{F}(x^n)) &\le \sqrt{\frac{\pi}{2}} G_n(\mathcal{F}(x^n)) \\
&\le \sqrt{\frac{\pi}{2}} \left[2\widetilde{G}_n(\mathcal{F}(x^n)) + \sqrt{\frac{2}{\pi}} \frac{\max_{T\in\mathcal{T}} \sqrt{\sum_{i=1}^n |f_T(x_i)|^2}}{n} \right] \\
&\le \sqrt{\frac{\pi}{2}} \left[\frac{6\alpha^2 k}{\sqrt{n}} + \sqrt{\frac{2}{\pi}} \frac{2\alpha}{\sqrt{n}} \right] \\
&\le \frac{10\alpha^2 k}{\sqrt{n}}.
\end{aligned}
$$

Again, recalling (11.8), we see that the event (11.3) occurs with probability at least $1-\delta$. The proof of Theorem 11.1 is complete.

11.3. Linear operators between Hilbert spaces

We assume, for simplicity, that all Hilbert spaces $\mathcal{H}$ of interest are *separable*. By definition, a Hilbert space $\mathcal{H}$ is separable if it has a countable dense subset: there exists a countable set $\{h_1, h_2, \ldots\} \subset \mathcal{H}$, such that for any $h \in \mathcal{H}$ and any $\varepsilon > 0$ there exists some $j \in \mathbb{N}$, for which $\|h - h_j\|_{\mathcal{H}} < \varepsilon$. Any separable Hilbert space $\mathcal{H}$ has a countable complete and orthonormal basis, i.e., a countable set $\{\varphi_1, \varphi_2, \ldots\} \subset \mathcal{H}$ with the following properties:

(1) **Orthonormality** — $\langle \varphi_i, \varphi_j \rangle_{\mathcal{H}} = \delta_{ij}$;
(2) **Completeness** — if there exists some $h \in \mathcal{H}$ which is orthogonal to all φ_j's, i.e., $\langle h, \varphi_j \rangle = 0$ for all j, then $h = 0$.

As a consequence, any $h \in \mathcal{H}$ can be uniquely represented as an infinite linear combination

$$h = \sum_{j=1}^{\infty} c_j \varphi_j, \qquad \text{where } c_j = \langle h, \varphi_j \rangle_{\mathcal{H}},$$

where the infinite series converges in norm, i.e., for any $\varepsilon > 0$ there exists some $n \in \mathbb{N}$, such that

$$\left\| \varphi - \sum_{j=1}^{n} c_j \varphi_j \right\|_{\mathcal{H}} < \varepsilon.$$

Moreover, $\|h\|^2_{\mathcal{H}} = \sum_{j=1}^{\infty} |c_j|^2$.

Let $\mathcal{H}$ and $\mathcal{K}$ be two Hilbert spaces. A *linear operator* from $\mathcal{H}$ into $\mathcal{K}$ is a mapping $T : \mathcal{H} \to \mathcal{K}$, such that (i) $T(\alpha h + \alpha' h') = \alpha T h + \alpha' T h'$ for any two $h, h' \in \mathcal{H}$ and $\alpha, \alpha' \in \mathbb{R}$. A linear operator $T : \mathcal{H} \to \mathcal{K}$ is *bounded* if

$$\|T\|_{\mathcal{H} \to \mathcal{K}} := \sup_{h \in \mathcal{H},\, h \neq 0} \frac{\|Th\|_{\mathcal{K}}}{\|h\|_{\mathcal{H}}} < \infty.$$

We will denote the space of all bounded linear operators $T : \mathcal{H} \to \mathcal{K}$ by $\mathcal{L}(\mathcal{H}, \mathcal{K})$. When $\mathcal{H} = \mathcal{K}$, we will write $\mathcal{L}(\mathcal{H})$ instead. For any operator $T \in \mathcal{L}(\mathcal{H}, \mathcal{K})$, we have the *adjoint operator* $T^* \in \mathcal{L}(\mathcal{K}, \mathcal{H})$, which is characterized by

$$\langle g, Th \rangle_{\mathcal{K}} = \langle T^* g, h \rangle_{\mathcal{H}}, \qquad \forall g \in \mathcal{K}, h \in \mathcal{H}.$$

If $T \in \mathcal{L}(\mathcal{H})$ has the property that $T = T^*$, we say that T is *self-adjoint*.

Some examples:

- The *identity operator* on $\mathcal{H}$, denoted by $I_{\mathcal{H}}$, maps each $h \in \mathcal{H}$ to itself. $I_{\mathcal{H}}$ is a self-adjoint operator with $\|I_{\mathcal{H}}\| \equiv \|I_{\mathcal{H}}\|_{\mathcal{H} \to \mathcal{H}} = 1$. We will often omit the index $\mathcal{H}$ and just write I.
- A *projection* is an operator $\Pi \in \mathcal{L}(\mathcal{H})$ satisfying $\Pi^2 = \Pi$, i.e., $\Pi(\Pi h) = \Pi h$ for any $h \in \mathcal{H}$. This is a bounded operator with $\|\Pi\| = 1$. Any projection is self-adjoint.
- An *isometry* is an operator $T \in \mathcal{L}(\mathcal{H}, \mathcal{K})$, such that $\|Th\|_{\mathcal{K}} = \|h\|_{\mathcal{H}}$ for all $h \in \mathcal{H}$, i.e., T preserves norms. If T is an isometry, then $T^* T = I_{\mathcal{H}}$, while $TT^* \in \mathcal{L}(\mathcal{K})$ is a projection. This is easy to see:

  $$(TT^*)(TT^*) = T(T^*T)T^* = TT^*.$$

  If $T \in \mathcal{L}(\mathcal{H})$ and $T^* \in \mathcal{L}(\mathcal{H})$ are both isometries, then T is called a *unitary operator*.
- If $\Pi \in \mathcal{L}(\mathcal{H})$ is a projection whose range $\mathcal{K} \subseteq \mathcal{H}$ is a closed k-dimensional subspace, then there exists an isometry $T \in \mathcal{L}(\mathbb{R}^k, \mathcal{K})$, such that $\Pi = TT^*$. Here, $\mathbb{R}^k$ is a Hilbert space with the usual $\|\cdot\|_2$ norm. To see this, let $\{\psi_1, \ldots, \psi_k\} \subset \mathcal{H}$ be an orthonormal basis of $\mathcal{K}$, and complete it to a countable basis $\{\psi_1, \psi_2, \ldots, \psi_k, \psi_{k+1}, \psi_{k+2}, \ldots\}$ for the entire $\mathcal{H}$. Here, the elements of $\{\psi_j\}_{j=k+1}^{\infty}$

are mutually orthonormal and orthogonal to $\{\psi_j\}_{j=1}^k$. Any $h \in \mathcal{H}$ has a unique representation

$$h = \sum_{j=1}^{\infty} \alpha_j \psi_j$$

for some real coefficients $\alpha_1, \alpha_2, \ldots$. With this, we can write out the action of Π explicitly as

$$\Pi h = \sum_{j=1}^{k} \alpha_j \psi_j.$$

Now consider the map $T : \mathbb{R}^k \to \mathcal{K}$ that takes

$$\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k \longmapsto \sum_{j=1}^{k} \alpha_j \psi_j.$$

It is easy to see that T is an isometry. Indeed,

$$\|T\alpha\|_{\mathcal{H}} = \left\| \sum_{j=1}^{k} \alpha_j \psi_j \right\|_{\mathcal{H}} = \sqrt{\sum_{j=1}^{k} \alpha_j^2} = \|\alpha\|_2.$$

The adjoint of T is easily computed: for any $\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k$ and any $h' = \sum_{j=1}^{\infty} \alpha'_j \psi_j \in \mathcal{H}$,

$$\begin{aligned}
\langle h', T\alpha \rangle_{\mathcal{H}} &= \langle \Pi h', T\alpha \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{j=1}^{k} \alpha'_j \psi_j, \sum_{j=1}^{k} \alpha_j \psi_j \right\rangle \\
&= \sum_{j=1}^{k} \alpha'_j \alpha_j \\
&= \langle T^* h', \alpha \rangle.
\end{aligned}$$

Since this must hold for arbitrary $\alpha \in \mathbb{R}^k$ and $h' \in \mathcal{H}'$, we must have $T^* h' = T^* \left(\sum_{j=1}^{\infty} \alpha'_j \psi_j \right) = (\alpha'_1, \ldots, \alpha'_j)$. Now let's compute $T^* h$ for any $h = \sum_j \alpha_j \psi_j$:

$$\begin{aligned}
TT^* h &= T(T^* h) \\
&= T \left(T^* \left(\sum_{j=1}^{\infty} \alpha_j \psi_j \right) \right) \\
&= T\left((\alpha_1, \ldots, \alpha_k)\right) \\
&= \sum_{j=1}^{k} \alpha_j \psi_j \\
&= \Pi h.
\end{aligned}$$

CHAPTER 12

Stochastic simulation via Rademacher bootstrap

In this chapter, we will look at an application of statistical learning theory to the problem of efficient stochastic simulation, which arises frequently in engineering design. The basic question is as follows. Suppose we have a system with input space Z. The system has a tunable parameter θ that lies in some set Θ. We have a *performance index* $\ell : \mathsf{Z} \times \Theta \to [0,1]$, where we assume that the lower the value of ℓ, the better the performance. Thus, if we use the parameter setting $\theta \in \Theta$ and apply input $z \in \mathsf{Z}$, the performance of the corresponding system is given by the scalar $\ell(z,\theta) \in [0,1]$. Now let's suppose that the input to the system is actually a *random variable* $Z \in \mathsf{Z}$ with some distribution $P \in \mathcal{P}(\mathsf{Z})$. Then we can define the *operating characteristic*

$$L(\theta) := \mathbf{E}_P[\ell(Z,\theta)] \equiv \int_{\mathsf{Z}} \ell(z,\theta) P_Z(dz), \qquad \theta \in \Theta. \tag{12.1}$$

The goal is to find an *optimal operating point* $\theta^* \in \Theta$ that achieves (or comes arbitrarily close to) $\inf_{\theta\in\Theta} L(\theta)$.

In practice, the problem of minimizing $L(\theta)$ is quite difficult for large-scale systems. First of all, computing the integral in (12.1) may be a challenge. Secondly, we may not even know the distribution P_Z. Thirdly, there may be more than one distribution of the input, each corresponding to different operating regimes and/or environments. For this reason, engineers often resort to Monte Carlo simulation techniques: Assuming we can efficiently sample from P_Z, we draw a large number of independent samples $Z_1, Z_2, \ldots, Z_n$ and compute

$$\widehat{\theta}_n = \operatorname*{arg\,min}_{\theta\in\Theta} L_n(\theta) \equiv \operatorname*{arg\,min}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell(Z_i,\theta),$$

where $L_n(\cdot)$ denotes the empirical version of the operating characteristic (12.1). Given an accuracy parameter $\varepsilon > 0$ and a confidence parameter $\delta \in (0,1)$, we simply need to draw enough samples, so that

$$L(\widehat{\theta}_n) \le \inf_{\theta\in\Theta} L(\theta) + \varepsilon$$

with probability at least $1-\delta$, regardless of what the true distribution P_Z happens to be.

This is, of course, just another instance of the ERM algorithm we have been studying extensively. However, there are two issues. One is how many samples we need to guarantee that the empirically optimal operating point will be good. The other is the complexity of actually computing an empirical minimizer.

The first issue has already come up in the course under the name of *sample complexity* of learning. The second issue is often handled by relaxing the problem a bit: We choose a probability distribution Q over Θ (assuming it can be equipped with an appropriate σ-algebra) and, instead of minimizing $L(\theta)$ over $\theta \in \Theta$, set some *level parameter* $\alpha \in (0,1)$, and seek any $\widehat{\theta} \in \Theta$, for which there exists some *exceptional set* $\Lambda \subset \Theta$ with $Q(\Lambda) \le \alpha$, such that

$$\inf_{\theta} L(\theta) - \varepsilon \le L(\widehat{\theta}) \le \inf_{\theta \in \Theta \setminus \Lambda} L(\theta) + \varepsilon \tag{12.2}$$

with probability at least $1-\delta$. Unless the actual optimal operating point θ^* happens to lie in the exceptional set Λ, we will come to within ε of the optimum with confidence at least $1-\delta$. Then we just need to draw a large enough number n of samples $Z_1, \dots, Z_n$ from P_Z and a large enough number m of samples $\theta_1, \dots, \theta_m$ from Q, and then compute

$$\widehat{\theta} = \operatorname*{arg\,min}_{\theta \in \{\theta_1, \dots, \theta_m\}} L_n(\theta).$$

In the next several lectures, we will see how statistical learning theory can be used to develop such simulation procedures. Moreover, we will learn how to use Rademacher averages[1] to determine how many samples we need *in the process of learning*. The use of statistical learning theory for simulation has been pioneered in the context of control by M. Vidyasagar [**Vid98, Vid01**]; the refinement of his techniques using Rademacher averages is due to Koltchinskii et al. [**KAA+00a, KAA+00b**]. We will essentially follow their presentation, but with slightly better constants.

We will follow the following plan. First, we will revisit the abstract ERM problem and its sample complexity. Then we will introduce a couple of refined tools pertaining to Rademacher averages. Next, we will look at *sequential* algorithms for empirical approximation, in which the sample complexity is not set *a priori*, but is rather determined by a data-driven *stopping rule*. And, finally, we will see how these sequential algorithms can be used to develop robust and efficient stochastic simulation strategies.

12.1. Empirical Risk Minimization: a quick review

Recall the *abstract Empirical Risk Minimization problem*: We have a space Z, a class $\mathcal{P}$ of probability distributions over Z, and a class $\mathcal{F}$ of measurable functions $f : \mathsf{Z} \to [0,1]$. Given an i.i.d. sample Z^n drawn according to some unknown $P \in \mathcal{P}$, we compute

$$\widehat{f}_n := \operatorname*{arg\,min}_{f \in \mathcal{F}} P_n(f) \equiv \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(Z_i).$$

We would like for $P(\widehat{f}_n)$ to be close to $\inf_{f \in \mathcal{F}} P(f)$ with high probability. To that end, we have derived the bound

$$P(\widehat{f}_n) - \inf_{f \in \mathcal{F}} P(f) \le 2\|P_n - P\|_{\mathcal{F}},$$

where, as before, we have defined the *uniform deviation*

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |P_n(f) - P(f)| = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(Z_i) - \mathbf{E}_P f(Z) \right|.$$

Hence, if n is sufficiently large so that, for every $P \in \mathcal{P}$, $\|P_n - P\|_{\mathcal{F}} \le \varepsilon/2$ with P-probability at least $1-\delta$, then $P(\widehat{f}_n)$ will be ε-close to $\inf_{f \in \mathcal{F}} P(f)$ with probability at least $1-\delta$. This motivates the following definition:

[1] More precisely, their stochastic counterpart, in which we do not take the expectation over the Rademacher sequence, but rather use it as a *resource* to aid the simulation.

Definition 12.1. *Given the pair* $(\mathcal{F}, \mathcal{P})$*, an accuracy parameter* $\varepsilon > 0$*, and a confidence parameter* $\delta \in (0,1)$*, the* sample complexity of empirical approximation *is*

$$N(\varepsilon;\delta) := \min\left\{ n \in \mathbb{N} : \sup_{P \in \mathcal{P}} \mathbf{P}\left\{ \|P_n - P\|_{\mathcal{F}} \ge \varepsilon \right\} \le \delta \right\}. \tag{12.3}$$

In other words, for any $\varepsilon > 0$ and any $\delta \in (0,1)$, $N(\varepsilon/2;\delta)$ is an *upper bound* on the number of samples needed to guarantee that $P(\widehat{f}_n) \le \inf_{f \in \mathcal{F}} P(f) + \varepsilon$ with probability (confidence) at least $1-\delta$.

12.2. Empirical Rademacher averages

As before, let Z^n be an i.i.d. sample of length n from some $P \in \mathcal{P}(\mathsf{Z})$. On multiple occasions we have seen that the performance of the ERM algorithm is controlled by the *Rademacher average*

$$R_n(\mathcal{F}(Z^n)) := \frac{1}{n}\mathbf{E}_{\sigma^n}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| \right], \tag{12.4}$$

where $\sigma^n = (\sigma_1, \ldots, \sigma_n)$ is an n-tuple of i.i.d. Rademacher random variables independent of Z^n. More precisely, we have stablished the fundamental *symmetrization inequality*

$$\mathbf{E}\|P_n - P\|_{\mathcal{F}} \le 2\mathbf{E}R_n(\mathcal{F}(Z^n)), \tag{12.5}$$

as well as the concentration bounds

$$\mathbf{P}\left\{ \|P_n - P\|_{\mathcal{F}} \ge \mathbf{E}\|P_n - P\|_{\mathcal{F}} + \varepsilon \right\} \le e^{-2n\varepsilon^2} \tag{12.6}$$

$$\mathbf{P}\left\{ \|P_n - P\|_{\mathcal{F}} \le \mathbf{E}\|P_n - P\|_{\mathcal{F}} - \varepsilon \right\} \le e^{-2n\varepsilon^2} \tag{12.7}$$

These results show two things:

1. The uniform deviation $\|P_n - P\|_{\mathcal{F}}$ tightly concentrates around its expected value.
2. The expected value $\mathbf{E}\|P_n - P\|_{\mathcal{F}}$ is bounded from above by $\mathbf{E}R_n(\mathcal{F}(Z^n))$.

It turns out that the expected Rademacher average $\mathbf{E}R_n(\mathcal{F}(Z^n))$ also furnishes a *lower bound* on $\mathbf{E}\|P_n - P\|_{\mathcal{F}}$:

Lemma 12.1 (Desymmetrization inequality). *For any class* $\mathcal{F}$ *of measurable functions* $f : \mathsf{Z} \to [0,1]$*, we have*

$$\frac{1}{2}\mathbf{E}R_n(\mathcal{F}(Z^n)) - \frac{1}{2\sqrt{n}} \le \frac{1}{2n}\mathbf{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i [f(Z_i) - P(f)] \right| \right] \le \mathbf{E}\|P_n - P\|_{\mathcal{F}}. \tag{12.8}$$

Proof. We will first prove the second inequality in (12.8). To that end, for each $1 \le i \le n$ and each $f \in \mathcal{F}$, let us define $U_i(f) := f(Z_i) - P(f)$. Then $\mathbf{E}U_i(f) = 0$. Let $\overline{Z}_1, \ldots, \overline{Z}_n$ be an independent copy of $Z_1, \ldots, Z_n$. Then we can define $\overline{U}_i(f), 1 \le i \le n$, similarly.

Moreover, since $\mathbf{E}U_i(f) = 0$, we can write

$$\begin{aligned}
\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i[f(Z_i)-P(f)]\right|\right] &= \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i U_i(f)\right|\right]\\
&= \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i\left[U_i(f)-\mathbf{E}\overline{U}_i(f)\right]\right|\right]\\
&\le \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i[U_i(f)-\overline{U}_i(f)]\right|\right].
\end{aligned}$$

Since, for each i, $U_i(f)$ and $\overline{U}_i(f)$ are i.i.d., the difference $U_i(f)-\overline{U}_i(f)$ is a symmetric random variable. Therefore,

$$\left\{\sigma_i[U_i(f)-\overline{U}_i(f)] : 1\le i\le n\right\} \stackrel{\text{(d)}}{=} \left\{U_i(f)-\overline{U}_i(f) : 1\le i\le n\right\}.$$

Using this fact and the triangle inequality, we get

$$\begin{aligned}
\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i[U_i(f)-\overline{U}_i(f)]\right|\right] &= \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}[U_i(f)-\overline{U}_i(f)]\right|\right]\\
&\le 2\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}U_i(f)\right|\right]\\
&= 2\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}|f(Z_i)-P(f)\right|\right]\\
&= 2n\cdot\mathbf{E}\|P_n-P\|_{\mathcal{F}}.
\end{aligned}$$

To prove the first inequality in (12.8), we write

$$\begin{aligned}
\mathbf{E}R_n(\mathcal{F}(Z^n)) &= \frac{1}{n}\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i\left[f(Z_i)-P(f)+P(f)\right]\right|\right]\\
&\le \frac{1}{n}\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i[f(Z_i)-P(f)]\right|\right] + \frac{1}{n}\mathbf{E}\left[\sup_{f\in\mathcal{F}}P(f)\cdot\left|\sum_{i=1}^{n}\sigma_i\right|\right]\\
&= \frac{1}{n}\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i[f(Z_i)-P(f)]\right|\right] + \frac{1}{n}\mathbf{E}\left|\sum_{i=1}^{n}\sigma_i\right|\\
&\le \frac{1}{n}\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i[f(Z_i)-P(f)]\right|\right] + \frac{1}{\sqrt{n}}.
\end{aligned}$$

Rearranging, we get the desired inequality. □

In this section, we will see that we can get a lot of mileage out of the *stochastic version* of the Rademacher average. To that end, let us define

$$r_n(\mathcal{F}(Z^n)) := \frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(Z_i)\right|. \tag{12.9}$$

The key difference between (12.4) and (12.9) is that, in the latter, we do *not* take the expectation over the Rademacher sequence σ^n. In other words, both $R_n(\mathcal{F}(Z^n))$ and $r_n(\mathcal{F}(Z^n))$ are random variables, but the former depends only on the training data Z^n, while the latter also depends on the n Rademacher random variables $\sigma_1, \ldots, \sigma_n$. We see immediately that $R_n(\mathcal{F}(Z^n)) = \mathbf{E}[r_n(\mathcal{F}(Z^n))|Z^n]$ and $\mathbf{E}R_n(\mathcal{F}(Z^n)) = \mathbf{E}r_n(\mathcal{F}(Z^n))$, where the expectation on the right-hand side is over both Z^n and σ^n. The following result will be useful:

Lemma 12.2 (Concentration inequalities for Rademacher averages). *For any $\varepsilon > 0$,*

$$\mathbf{P}\left\{r_n(\mathcal{F}(Z^n)) \geq \mathbf{E}R_n(\mathcal{F}(Z^n)) + \varepsilon\right\} \leq e^{-n\varepsilon^2/2} \tag{12.10}$$

and

$$\mathbf{P}\left\{r_n(\mathcal{F}(Z^n)) \leq \mathbf{E}R_n(\mathcal{F}(Z^n)) - \varepsilon\right\} \leq e^{-n\varepsilon^2/2}. \tag{12.11}$$

Proof. For each $1 \leq i \leq n$, let $U_i := (Z_i, \sigma_i)$. Then $r_n(\mathcal{F}(Z^n))$ can be represented as a real-valued function $g(U^n)$. Moreover, it is easy to see that this function has bounded differences with $c_1 = \ldots = c_n = 2/n$. Hence, McDiarmid's inequality tells us that for any $\varepsilon > 0$

$$\mathbf{P}\left\{g(U^n) \geq \mathbf{E}g(U^n) + \varepsilon\right\} \leq e^{-n\varepsilon^2/2},$$

and the same holds for the probability that $g(U^n) \leq \mathbf{E}g(U^n) - \varepsilon$. This completes the proof. □

12.3. Sequential learning algorithms

In a sequential learning algorithm, the sample complexity is a *random variable*. It is not known in advance, but rather is computed from data in the process of learning. In other words, instead of using a training sequence of fixed length, we keep drawing independent samples until we decide that we have acquired enough of them, and then compute an empirical risk minimizer.

To formalize this idea, we need the notion of a *stopping time*. Let $\{U_n\}_{n=1}^{\infty}$ be a random process. A random variable τ taking values in $\mathbb{N}$ is called a *stopping time* if and only if, for each $n \geq 1$, the occurrence of the event $\{\tau = n\}$ is determined by $U^n = (U_1, \ldots, U_n)$. More precisely:

Definition 12.2. *For each n, let Σ_n denote the σ-algebra generated by U^n (in other words, Σ_n consists of all events that occur by time n). Then a random variable τ taking values in $\mathbb{N}$ is a stopping time if and only if, for each $n \geq 1$, the event $\{\tau = n\} \in \Sigma_n$.*

In other words, denoting by U^∞ the entire sample path $(U_1, U_2, \ldots)$ of our random process, we can view τ as a function that maps U^∞ into $\mathbb{N}$. For each n, the indicator function of the event $\{\tau = n\}$ is a function of U^∞:

$$\mathbf{1}_{\{\tau=n\}} \equiv \mathbf{1}_{\{\tau(U^\infty)=n\}}.$$

Then τ is a stopping time if and only if, for each n and for all U^∞, V^∞ with $U^n = V^n$ we have

$$\mathbf{1}_{\{\tau(U^\infty)=n\}} = \mathbf{1}_{\{\tau(V^\infty)=n\}}.$$

Our sequential learning algorithms will work as follows. Given a desired accuracy parameter $\varepsilon > 0$ and a confidence parameter $\delta > 0$, let $\overline{n}(\varepsilon, \delta)$ be the initial sample size; we will assume that $\overline{n}(\varepsilon, \delta)$ is a nonincreasing function of both ε and δ. Let $\mathcal{T}(\varepsilon, \delta)$ denote the set of all stopping times τ such that

$$\sup_{P \in \mathcal{P}} \mathbf{P} \left\{ \|P_\tau - P\|_{\mathcal{F}} \le \varepsilon \right\} \ge \delta.$$

Now if $\tau \in \mathcal{T}(\varepsilon, \delta)$ and we let

$$\widehat{f}_\tau := \arg\min_{f \in \mathcal{F}} P_\tau(f) \equiv \arg\min_{f \in \mathcal{F}} \frac{1}{\tau} \sum_{i=1}^{\tau} f(Z_i),$$

then we immediately see that

$$\sup_{P \in \mathcal{P}} \left\{ P(\widehat{f}_\tau) \ge \inf_{f \in \mathcal{F}} P(f) + 2\varepsilon \right\} \le \delta.$$

Of course, the whole question is how to construct an appropriate stopping time *without knowing* P.

Definition 12.3. *A parametric family of stopping times* $\{\nu(\varepsilon, \delta) : \varepsilon > 0, \delta \in (0,1)\}$ *is called* strongly efficient (SE) *(w.r.t.* $\mathcal{F}$ *and* $\mathcal{P}$*) if there exist constants* $K_1, K_2, K_3 \ge 1$*, such that for all* $\varepsilon > 0, \delta \in (0,1)$

$$\nu(\varepsilon, \delta) \in \mathcal{T}(K_1 \varepsilon, \delta) \tag{12.12}$$

and for all $\tau \in \mathcal{T}(\varepsilon, \delta)$

$$\sup_{P \in \mathcal{P}} \mathbf{P} \left\{ \nu(K_2 \varepsilon, \delta) > \tau \right\} \le K_3 \delta. \tag{12.13}$$

In other words, Eq. (12.12) says that any SE stopping time $\{\nu(\varepsilon, \delta)\}$ guarantees that we can approximate statistical expectations by empirical expectations with accuracy $K_1 \varepsilon$ and confidence $1 - \delta$; similarly, Eq. (12.13) says that, with probability at least $1 - K_3 \delta$, we will require at most as many samples as would be needed by *any* sequential algorithm for empirical approximation with accuracy ε / K_2 and confidence $1 - \delta$.

Definition 12.4. *A family of stopping times* $\{\nu(\varepsilon, \delta) : \varepsilon > 0, \delta \in (0,1)\}$ *is* weakly efficient (WE) *for* $(\mathcal{F}, \mathcal{P})$ *if there exist constants* $K_1, K_2, K_3 \ge 1$*, such that for all* $\varepsilon > 0, \delta \in (0,1)$

$$\nu(\varepsilon, \delta) \in \mathcal{T}(K_1 \varepsilon, \delta) \tag{12.14}$$

and

$$\sup_{P \in \mathcal{P}} \mathbf{P} \left\{ \nu(K_2 \varepsilon, \delta) > N(\varepsilon; \delta) \right\} \le K_3 \delta. \tag{12.15}$$

If $\nu(\varepsilon, \delta)$ is a WE stopping time, then Eq. (12.14) says that we can solve the empirical approximation problem with accuracy $K_1 \varepsilon$ and confidence $1 - \delta$; Eq. (12.15) says that, with probability at most $1 - \delta$, the sample complexity will be less than the sample complexity of empirical approximation with accuracy ε / K_2 and confidence $1 - \delta$.

If $N(\varepsilon; \delta) \ge \overline{n}(\varepsilon, \delta)$, then $N(\varepsilon, \delta) \in \mathcal{T}(\varepsilon, \delta)$. Hence, any WE stopping time is also SE. The converse, however, is not true.

12.3.1. A strongly efficient sequential learning algorithm. Let $\{Z_n\}_{n=1}^\infty$ be an infinite sequence of i.i.d. draws from some $P \in \mathcal{P}$; let $\{\sigma_n\}_{n=1}^\infty$ be an i.i.d. Rademacher sequence independent of $\{Z_n\}$. Choose

$$\overline{n}(\varepsilon,\delta) \ge \left\lfloor \frac{2}{\varepsilon^2} \log \frac{2}{\delta(1-e^{-\varepsilon^2/2})} \right\rfloor + 1 \tag{12.16}$$

and let

$$\nu(\varepsilon,\delta) := \min\left\{ n \ge \overline{n}(\varepsilon,\delta) : r_n(\mathcal{F}(Z^n)) \le \varepsilon \right\}. \tag{12.17}$$

This is clearly a stopping time for each $\varepsilon > 0$ and each $\delta \in (0,1)$.

Theorem 12.1. *The family $\{\nu(\varepsilon,\delta) : \varepsilon > 0, \delta \in (0,1)\}$ defined in (12.17) with $\overline{n}(\varepsilon,\delta)$ set according to (12.16) is SE for any class $\mathcal{F}$ of measurable functions $f : \mathsf{Z} \to [0,1]$ and $\mathcal{P} = \mathcal{P}(\mathsf{Z})$ with $K_1 = 5, K_2 = 6, K_3 = 1$.*

Proof. Let $\overline{n} = \overline{n}(\varepsilon,\delta)$. We will first show that, for any $P \in \mathcal{P}(\mathsf{Z})$,

$$\|P_n - P\|_{\mathcal{F}} \le 2r_n(\mathcal{F}(Z^n)) + 3\varepsilon, \qquad \forall n \ge \overline{n} \tag{12.18}$$

with probability at least $1-\delta$. Since for $n = \nu(\varepsilon,\delta) \ge \overline{n}$ we have $r_n(\mathcal{F}(Z^n)) \le \varepsilon$, we will immediately be able to conclude that

$$\mathbf{P}\left\{ \|P_{\nu(\varepsilon,\delta)} - P\|_{\mathcal{F}} \ge 5\varepsilon \right\} \le \delta,$$

which will imply that $\nu(\varepsilon,\delta) \in \mathcal{T}(5\varepsilon,\delta)$. Now we prove (12.18). First of all, applying Lemma 12.2 and the union bound, we can write

$$\begin{aligned}
\mathbf{P}\left\{ \bigcup_{n\ge\overline{n}} \{ r_n(\mathcal{F}(Z^n)) \ge \mathbf{E}R_n(\mathcal{F}(Z^n)) + \varepsilon \} \right\} &\le \sum_{n\ge\overline{n}} e^{-n\varepsilon^2/2} \\
&= e^{-\overline{n}\varepsilon^2/2} \sum_{n\ge 0} e^{-n\varepsilon^2/2} \\
&= \frac{e^{-\overline{n}\varepsilon^2/2}}{1-e^{-\varepsilon^2/2}} \\
&\le \delta/2.
\end{aligned}$$

From the symmetrization inequality (12.5), we know that $\mathbf{E}\|P_n - P\|_{\mathcal{F}} \le 2\mathbf{E}R_n(\mathcal{F}(Z^n))$. Moreover, using (12.6) and the union bound, we can write

$$\begin{aligned}
\mathbf{P}\left\{ \bigcup_{n\ge\overline{n}} \{ \|P_n - P\|_{\mathcal{F}} \ge \mathbf{E}\|P_n - P\|_{\mathcal{F}} + \varepsilon \} \right\} &\le \sum_{n\ge\overline{n}} e^{-2n\varepsilon^2} \\
&\le \sum_{n\ge\overline{n}} e^{-n\varepsilon^2/2} \\
&\le \delta/2.
\end{aligned}$$

Therefore, with probability at least $1-\delta$,

$$\|P_n - P\|_{\mathcal{F}} \le \mathbf{E}\|P_n - P\|_{\mathcal{F}} + \varepsilon \le 2\mathbf{E}R_n(\mathcal{F}(Z^n)) + \varepsilon \le 2r_n(\mathcal{F}(Z^n)) + 3\varepsilon, \qquad \forall n \ge \overline{n}$$

which is (12.18). This shows that (12.12) holds for $\nu(\varepsilon,\delta)$ with $K_1 = 5$.

Next, we will prove that, for any $P \in \mathcal{P}(\mathsf{Z})$,

$$\mathbf{P}\left\{\min_{\overline{n} \le n < \nu(6\varepsilon,\delta)} \|P_n - P\|_{\mathcal{F}} < \varepsilon\right\} \le \delta. \tag{12.19}$$

In other words, (12.19) says that, with probability at least $1-\delta$, $\|P_n - P\|_{\mathcal{F}} \ge \varepsilon$ for all $\overline{n} \le n < \nu(6\varepsilon,\delta)$. This means that, for any $\tau \in \mathcal{T}(\varepsilon,\delta)$, $\nu(6\varepsilon,\delta) \le \tau$ with probability at least $1-\delta$, which will give us (12.13) with $K_2 = 6$ and $K_3 = 1$.

To prove (12.19), we have by (12.7) and the union bound that

$$\mathbf{P}\left\{\bigcup_{n \ge \overline{n}} \left\{\|P_n - P\|_{\mathcal{F}} \le \mathbf{E}\|P_n - P\|_{\mathcal{F}} - \varepsilon\right\}\right\} \le \delta/2.$$

By the desymmetrization inequality (12.8), we have

$$\mathbf{E}\|P_n - P\|_{\mathcal{F}} \ge \frac{1}{2}\mathbf{E}R_n(\mathcal{F}(Z^n)) - \frac{1}{2\sqrt{n}}, \qquad \forall n.$$

Finally, by the concentration inequality (12.10) and the union bound,

$$\mathbf{P}\left\{\bigcup_{n \ge \overline{n}} \left\{r_n(\mathcal{F}(Z^n)) \ge \mathbf{E}R_n(\mathcal{F}(Z^n)) + \varepsilon\right\}\right\} \le \delta/2.$$

Therefore, with probability at least $1-\delta$,

$$\|P_n - P\|_{\mathcal{F}} \ge \frac{1}{2}r_n(\mathcal{F}(Z^n)) - \frac{1}{2\sqrt{n}} - \frac{3\varepsilon}{2}, \qquad \forall n \ge \overline{n}.$$

If $\overline{n} \le n < \nu(6\varepsilon,\delta)$, then $r_n(\mathcal{F}(Z^n)) > 6\varepsilon$. Therefore, using the fact that $n \ge \overline{n}$ and $\overline{n}(\varepsilon,\delta)^{-1/2} \le \varepsilon$, we see that, with probability at least $1-\delta$,

$$\|P_n - P\|_{\mathcal{F}} > \frac{3\varepsilon}{2} - \frac{1}{2\sqrt{n}} \ge \frac{3\varepsilon}{2} - \frac{1}{2\sqrt{\overline{n}}} \ge \varepsilon, \qquad \overline{n} \le n < \nu(6\varepsilon,\delta).$$

This proves (12.19), and we are done. $\square$

12.3.2. A weakly efficient sequential learning algorithm. Now choose

$$\overline{n}(\varepsilon,\delta) \ge \left\lfloor \frac{2}{\varepsilon^2}\log\frac{4}{\delta}\right\rfloor + 1, \tag{12.20}$$

for each $k = 0,1,2,\ldots$ let $n_k := 2^k\overline{n}(\varepsilon,\delta)$, and let

$$\nu(\varepsilon,\delta) := \min\left\{n_k : r_{n_k}(\mathcal{F}(Z^{n_k})) \le \varepsilon\right\}. \tag{12.21}$$

Theorem 12.2. *The family $\{\nu(\varepsilon,\delta) : \varepsilon > 0, \delta \in (0,1/2)\}$ defined in (12.21) with $\overline{n}(\varepsilon,\delta)$ set according to (12.20) is WE for any class $\mathcal{F}$ of measurable functions $f : \mathsf{Z} \to [0,1]$ and $\mathcal{P} = \mathcal{P}(\mathsf{Z})$ with $K_1 = 5$, $K_2 = 18$, $K_3 = 3$.*

PROOF. As before, let $\overline{n} = \overline{n}(\varepsilon, \delta)$. The proof of (12.14) is similar to what we have done in the proof of Theorem 12.1, except we use the bounds

$$\begin{aligned}
\mathbf{P}\left\{\bigcup_{k=0}^{\infty}\{r_{n_k}(\mathcal{F}(Z^{n_k})) \ge \mathbf{E}R_{n_k}(\mathcal{F}(Z^{n_k})) + \varepsilon\}\right\} &\le \sum_{k=0}^{\infty} e^{-2^k\overline{n}\varepsilon^2/2} \\
&= e^{-\overline{n}\varepsilon^2/2} + e^{-\overline{n}\varepsilon^2/2}\sum_{k=1}^{\infty} e^{-\frac{\overline{n}\varepsilon^2}{2}(2^k-1)} \\
&\le e^{-\overline{n}\varepsilon^2/2} + e^{-\overline{n}\varepsilon^2/2}\sum_{k=1}^{\infty} e^{-(2^k-1)} \\
&\le e^{-\overline{n}\varepsilon^2/2} + e^{-\overline{n}\varepsilon^2/2}\sum_{k=1}^{\infty} e^{-k} \\
&\le 2e^{-\overline{n}\varepsilon^2/2} \\
&\le \delta/2,
\end{aligned}$$

where in the third step we have used the fact that $\overline{n}\varepsilon^2/2 \ge 1$. Similarly,

$$\mathbf{P}\left\{\bigcup_{k=0}^{\infty}\{\|P_{n_k} - P\|_{\mathcal{F}} \le \mathbf{E}\|P_{n_k} - P\|_{\mathcal{F}} + \varepsilon\}\right\} \le \delta^2.$$

Therefore,

$$\|P_{n_k} - P\|_{\mathcal{F}} \le 2r_{n_k}(\mathcal{F}(Z^{n_k})) + 3\varepsilon, \qquad \forall k = 0, 1, 2, \ldots$$

and consequently

$$\mathbf{P}\left\{\|P_{\nu(\varepsilon,\delta)} - P\|_{\mathcal{F}} \ge 5\varepsilon\right\} \le \delta,$$

which proves (12.14).

Now we prove (12.15). Let $N = N(\varepsilon, \delta)$, the sample complexity of empirical approximation that we have defined in (12.3). Let us choose k so that $n_k \le N < n_{k+1}$, which is equivalent to $2^k\overline{n} \le N < 2^{k+1}\overline{n}$. Then

$$\mathbf{P}\{\nu(18\varepsilon, \delta) > N\} \le \mathbf{P}\{\nu(18\varepsilon, \delta) > n_k\}.$$

We will show that the probability on the right-hand side is less than 3δ. First of all, since $N \ge \overline{n}$ (by hypothesis), we have $n_k \ge \overline{n}/2 \ge 1/\varepsilon^2$. Therefore, with probability at least $1 - \delta$

$$\|P_{n_k} - P\|_{\mathcal{F}} \ge \frac{1}{2}r_{n_k}(\mathcal{F}(Z^{n_k})) - \frac{1}{2\sqrt{n_k}} - \frac{9\varepsilon}{2} \ge \frac{1}{2}r_{n_k}(\mathcal{F}(Z^{n_k})) - 5\varepsilon. \tag{12.22}$$

If $\nu(18\varepsilon, \delta) > n_k$, then by definition $r_{n_k}(\mathcal{F}(Z^{n_k})) > 18\varepsilon$. Writing $r_{n_k} = r_{n_k}(\mathcal{F}(Z^{n_k}))$ for brevity, we see get

$$\begin{aligned}
\mathbf{P}\{\nu(18\varepsilon, \delta) > n_k\} &\le \mathbf{P}\{r_{n_k} > 18\varepsilon\} \\
&= \mathbf{P}\{r_{n_k} > 18\varepsilon, \|P_{n_k} - P\|_{\mathcal{F}} \ge 18\varepsilon\} + \mathbf{P}\{r_{n_k} > 18\varepsilon, \|P_{n_k} - P\|_{\mathcal{F}} < 4\varepsilon\} \\
&\le \mathbf{P}\{\|P_{n_k} - P\|_{\mathcal{F}} \ge 4\varepsilon\} + \mathbf{P}\{r_{n_k} > 18\varepsilon, \|P_{n_k} - P\|_{\mathcal{F}} < 4\varepsilon\}.
\end{aligned}$$

If $r_{n_k} > 18\varepsilon$ but $\|P_{n_k} - P\|_{\mathcal{F}} < 4\varepsilon$, the event in (12.22) cannot occur. Indeed, suppose it does. Then it must be the case that $4\varepsilon > 9\varepsilon - 5\varepsilon = 4\varepsilon$, which is a contradiction. Therefore,

$$\mathbf{P}\left\{r_{n_k} > 18\varepsilon, \|P_{n_k} - P\|_{\mathcal{F}} < 4\varepsilon\right\} \le \delta,$$

and hence

$$\mathbf{P}\left\{\nu(18\varepsilon, \delta) > n_k\right\} \le \mathbf{P}\left\{\|P_{n_k} - P\|_{\mathcal{F}} \ge 4\varepsilon\right\} + \delta.$$

For each $f \in \mathcal{F}$ and each $n \in \mathbb{N}$ define

$$S_n(f) := \sum_{i=1}^{n}[f(Z_i) - P(f)]$$

and let $\|S_n\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |S_n(f)|$. Then

$$\mathbf{P}\left\{\|P_{n_k} - P\|_{\mathcal{F}} \ge 4\varepsilon\right\} = \mathbf{P}\left\{\|S_{n_k}\|_{\mathcal{F}} \ge 4\varepsilon n_k\right\} \le \mathbf{P}\left\{\|S_{n_k}\|_{\mathcal{F}} \ge 2\varepsilon N\right\}.$$

Since $n_k \le N$, the $\mathcal{F}$-indexed stochastic processes $S_{n_k}(f)$ and $S_N(f) - S_{n_k}(f)$ are independent. Therefore, we use a technical result stated as Lemma 12.4 in the appendix with $\xi_1 = S_{n_k}$ and $\xi_2 = S_N(f) - S_{n_k}(f)$ to write

$$\mathbf{P}\left\{\|S_{n_k}\|_{\mathcal{F}} \ge 2\varepsilon N\right\} \le \frac{\mathbf{P}\left\{\|S_N\|_{\mathcal{F}} \ge \varepsilon N\right\}}{\inf_{f\in\mathcal{F}} \mathbf{P}\left\{|S_N(f) - S_{n_k}(f)| \le \varepsilon N\right\}}.$$

By definition of $N = N(\varepsilon, \delta)$, the probability in the numerator is at most δ. To analyze the probability in the denominator, we use Hoeffding's inequality to get

$$\begin{aligned}
\inf_{f\in\mathcal{F}} \mathbf{P}\left\{|S_N(f) - S_{n_k}(f)| \le \varepsilon N\right\} &= 1 - \sup_{f\in\mathcal{F}} \mathbf{P}\left\{|S_N(f) - S_{n_k}(f)| > \varepsilon N\right\} \\
&\ge 1 - 2e^{-N\varepsilon^2/2} \\
&\ge 1 - \delta.
\end{aligned}$$

Therefore,

$$\mathbf{P}\left\{\nu(18\varepsilon, \delta) > n_k\right\} \le \frac{\delta}{1-\delta} + \delta \le 3\delta$$

for $\delta < 1/2$. Therefore, $\{\nu(\varepsilon, \delta) : \varepsilon \in (0, 1), \delta \in (0, 1/2)\}$ is WE with $K_1 = 5, K_2 = 18, K_3 = 3$. □

12.4. A sequential algorithm for stochastic simulation

Armed with these results on sequential learning algorithms, we can take up the question of constructing efficient simulation strategies. We fix an accuracy parameter $\varepsilon > 0$, a confidence parameter $\delta \in (0, 1)$, and a level parameter $\alpha \in (0, 1)$. Given two probability distributions, P on the input space Z and Q on the parameter space Θ, we draw a large i.i.d. sample $Z_1, \ldots, Z_n$ from P and a large i.i.d. sample $\theta_1, \ldots, \theta_m$ from Q. We then compute

$$\widehat{\theta} = \operatorname*{arg\,min}_{\theta\in\{\theta_1,\ldots,\theta_m\}} L_n(\theta),$$

where

$$L_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, \theta).$$

The goal is to pick n and m large enough so that, with probability at least $1-\delta$, $\widehat{\theta}$ is an ε-minimizer of L to level α, i.e., with probability at least $1-\delta$ there exists some set $\Lambda \subset \Theta$ with $Q(\Lambda) \le \alpha$, such that Eq. (12.2) holds with probability at least $1-\delta$.

To that end, consider the following algorithm based on Theorem 12.2, proposed by Koltchinskii et al. [**KAA+00a, KAA+00b**]:

Algorithm 1
choose positive integers m and n such that $\quad m \ge \frac{\log(2/\delta)}{\log[1/(1-\alpha)]}$ and $n \ge \lfloor \frac{50}{\varepsilon^2} \log \frac{8}{\delta} \rfloor + 1$ draw m independent samples $\theta_1, \ldots, \theta_m$ from Q draw n independent samples $Z_1, \ldots, Z_n$ from P_Z evaluate the stopping variable $\quad \gamma = \max_{1 \le j \le m} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(Z_i, \theta_j) \right|$ where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher r.v.'s independent of θ^m and Z^n if $\gamma > \varepsilon/5$, then $\quad$ add n more i.i.d. samples from P_Z and repeat else stop and output $\quad \widehat{\theta} = \arg\min_{\theta \in \{\theta_1, \ldots, \theta_n\}} L_n(\theta)$

Then we claim that, with probability at least $1-\delta$, $\widehat{\theta}$ is an ε-minimizer of L to level α. To see this, we need the following result [**Vid03**, Lemma 11.1]:

Lemma 12.3. *Let Q be a probability distribution on the parameter set Θ, and let $h : \Theta \to \mathbb{R}$ be a (measurable) real-valued function on Θ, bounded from above, i.e., $h(\theta) < +\infty$ for all $\theta \in \Theta$. Let $\theta_1, \ldots, \theta_m$ be m i.i.d. samples from Q, and let*

$$\bar{h}(\theta^m) := \max_{1 \le j \le m} h(\theta_m).$$

Then for any $\alpha \in (0,1)$

$$Q\left(\left\{\theta \in \Theta : h(\theta) > \bar{h}(\theta^m)\right\}\right) \le \alpha \tag{12.23}$$

with probability at least $1-(1-\alpha)^m$.

Proof. For each $c \in \mathbb{R}$, let

$$F(c) := \mathbf{P}\left(\{\theta \in \Theta : h(\theta) \le c\}\right).$$

Note that F is the CDF of the random variable $\xi = h(\theta)$ with $\theta \sim Q$. Therefore, it is right-continuous, i.e., $\lim_{c' \searrow c} F(c') = F(c)$. Now define

$$c_\alpha := \inf\{c : F(c) \ge 1-\alpha\}.$$

Since F is right-continuous, $F(c_\alpha) \ge 1-\alpha$. Moreover, if $c < c_\alpha$, then $F(c) < 1-\alpha$. Now let us suppose that $\bar{h}(\theta^m) \ge c_\alpha$. Then, since F is monotone nondecreasing,

$$\mathbf{P}\left(\left\{\theta \in \Theta : h(\theta) \le \bar{h}(\theta^m)\right\}\right) = F\left(\bar{h}(\theta^m)\right) \ge F(c_\alpha) \ge 1-\alpha,$$

or, equivalently, if $\bar{h}(\theta^m) \ge c_\alpha$, then

$$\mathbf{P}\left(\left\{\theta \in \Theta : h(\theta) > \bar{h}(\theta^m)\right\}\right) \le \alpha.$$

Therefore, if θ^m is such that

$$\mathbf{P}\left(\left\{\theta \in \Theta : h(\theta) > \bar{h}(\theta^m)\right\}\right) > \alpha,$$

then it must be the case that $\bar{h}(\theta^m) < c_\alpha$, which in turn implies that $F(\bar{h}(\theta^m)) < 1 - \alpha$, the complement of the event in (12.23). But $\bar{h}(\theta^m) < c_\alpha$ means that $h(\theta_j) < c_\alpha$ for every $1 \le j \le m$. Since the θ_j's are independent, the events $\{h(\theta_j) < c_\alpha\}$ are independent, and each occurs with probability at most $1 - \alpha$. Therefore,

$$\mathbf{P}\left(\left\{\theta^m \in \Theta^m : Q\left(\left\{\theta \in \Theta : h(\theta) > \bar{h}(\theta^m)\right\}\right)\right\}\right) \le (1-\alpha)^m,$$

which is what we intended to prove. $\square$

We apply this lemma to the function $h(\theta) = -L(\theta)$. Then, provided m is chosen as described in Algorithm 1, we will have

$$Q\left(\left\{\theta \in \Theta : L(\theta) < \min_{1 \le j \le j} L(\theta_m)\right\}\right) \le \delta/2.$$

Now consider the *finite* class of functions $\mathcal{F} = \{f_j(z) = \ell(z, \theta_j) : 1 \le j \le m\}$. By Theorem 12.2, the final output $\widehat{\theta} \in \{\theta_1, \ldots, \theta_m\}$ will satisfy

$$\left| L(\widehat{\theta}) - \min_{1 \le j \le m} L(\theta_j) \right| \le \varepsilon$$

with probability at least $1 - \delta/2$. Hence, with probability at least $1 - \delta$ there exists a set $\Lambda \subset \Theta$ with $Q(\Lambda) \le \alpha$, such that (12.2) holds. Moreover, the total number of samples used up by Algorithm 1 will be, with probability at least $1 - 3\delta/2$, no more than

$$N_{\mathcal{F}, P_Z}(\varepsilon/18, \delta/2) \equiv \min\left\{n \in \mathbb{N} : \mathbf{P}\left(\|P_n - P_Z\|_{\mathcal{F}} > \varepsilon/18\right) < \delta/2\right\}.$$

We can estimate $N_{\mathcal{F}, P_Z}(\varepsilon/18, \delta/2)$ as follows. First of all, the function

$$\Delta(Z^n) := \|P_n - P_Z\|_{\mathcal{F}} \equiv \max_{1 \le j \le m} |P_n(f_j) - P_Z(f_j)|$$

has bounded differences with $c_1 = \ldots = c_n = 1/n$. Therefore, by McDiarmid's inequality

$$\mathbf{P}\left(\Delta(Z^n) \ge \mathbf{E}\Delta(Z^n) + t\right) \le e^{-2nt^2}, \qquad \forall t > 0.$$

Secondly, since the class $\mathcal{F}$ is finite with $|\mathcal{F}| = m$, the symmetrization inequality (12.5) and the Finite Class Lemma give the bound

$$\mathbf{E}\|P_n - P_Z\|_{\mathcal{F}} \le 4\sqrt{\frac{\log m}{n}}.$$

Therefore, if we choose $t = \varepsilon/18 - 4\sqrt{n^{-1}\log m}$ and n is large enough so that $t > \varepsilon/20$ (say), then

$$\mathbf{P}\left(\|P_n - P\|_{\mathcal{F}} > \varepsilon/18\right) \le e^{-n\varepsilon^2/200}.$$

Hence, a fairly conservative estimate is

$$N_{\mathcal{F}, P_Z}(\varepsilon/18, \delta/2) \le \max\left\{\left\lfloor \frac{200}{\varepsilon^2} \log \frac{2}{\delta} \right\rfloor + 1, \left\lfloor \left(\frac{720}{\varepsilon}\right)^2 \log m \right\rfloor + 1\right\}$$

It is instructive to compare Algorithm 1 with a simple Monte Carlo strategy:

Algorithm 0
choose positive integers m and n such that
$m \geq \frac{\log(2/\delta)}{\log[1/(1-\alpha)]}$ and $n \geq \frac{1}{2\varepsilon^2}\log\frac{4m}{\delta}$
draw m independent samples $\theta_1, \ldots, \theta_m$ from Q
draw n independent samples $Z_1, \ldots, Z_n$ from P_Z
for $j = 1$ to m
compute $L_n(\theta_j) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i, \theta_j)$
end for
output $\widehat{\theta} = \arg\min_{\theta \in \{\theta_1,\ldots,\theta_m\}} L_n(\theta_j)$

The selection of m is guided by the same considerations as in Algorithm 1. Moreover, for each $1 \leq j \leq m$, $L_n(\theta_j)$ is an average of n independent random variables $\ell(Z_i, \theta_j) \in [0,1]$, and $L(\theta_j) = \mathbf{E}L_n(\theta_j)$. Hence, Hoeffding's inequality says that

$$\mathbf{P}\left(\{Z^n \in \mathsf{Z}^n : |L_n(\theta_j) - L(\theta_j)| > \varepsilon\}\right) \leq 2e^{-2n\varepsilon^2}.$$

If we choose n as described in Algorithm 0, then

$$\begin{aligned}
\mathbf{P}\left(\left|L_n(\widehat{\theta}) - \min_{1\leq j\leq m} L(\theta_j)\right| > \varepsilon\right) &\leq \mathbf{P}\left(\bigcup_{j=1}^m |L_n(\theta_j) - L(\theta_j)| > \varepsilon\right) \\
&\leq \sum_{j=1}^m \mathbf{P}\left(|L_n(\theta_j) - L(\theta_j)| > \varepsilon\right) \\
&\leq \delta/2.
\end{aligned}$$

Hence, with probability at least $1-\delta$ there exists a set $\Lambda \subset \Theta$ with $Q(\Lambda) \leq \alpha$, so that (12.2) holds. It may seem at first glance that Algorithm 0 is more efficient than Algorithm 1. However, this is not the case in high-dimensional situations. There, one can actually show that, with probability practically equal to one, the empirical minimum of L can be *much larger* than the true minimum (cf. [**KAA+00b**] for a very vivid numerical illustration). This is an instance of the so-called *Curse of Dimensionality*, which adaptive schemes like Algorithm 1 can often avoid.

12.5. Technical lemma

LEMMA 12.4. *Let $\{\xi_1(f) : f \in \mathcal{F}\}$ and $\{\xi_2(f) : f \in \mathcal{F}\}$ be two independent $\mathcal{F}$-indexed stochastic processes with*

$$\|\xi_j\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\xi_j(f)| < \infty, \qquad j = 1, 2.$$

Then for all $t > 0, c > 0$

$$\mathbf{P}\{\|\xi_1\|_{\mathcal{F}} \geq t + c\} \leq \frac{\mathbf{P}\{\|\xi_1 - \xi_2\|_{\mathcal{F}} \geq t\}}{\inf_{f\in\mathcal{F}} \mathbf{P}\{|\xi_2(f)| \leq c\}}. \tag{12.24}$$

PROOF. If $\|\xi_1\|_{\mathcal{F}} \geq t + c$, then there exists some $f \in \mathcal{F}$, such that $|\xi_1(f)| \geq t + c$. Then for this particular f by the triangle inequality we see that

$$|\xi_2(f)| \leq c \qquad \Rightarrow \qquad |\xi_1(f) - \xi_2(f)| \geq t$$

Therefore,

$$\inf_{f\in\mathcal{F}} \mathbf{P}_{\xi_2}\Big\{|\xi_2(f)| \le c\Big\} \le \mathbf{P}_{\xi_2}\Big\{|\xi_2(f)| \le c\Big\} \le \mathbf{P}_{\xi_2}\Big\{\, |\xi_1(f) - \xi_2(f)| \ge t\Big\} \le \mathbf{P}_{\xi_2}\Big\{\|\xi_1{-}\xi_2\|_{\mathcal{F}} \ge t\Big\}.$$

The leftmost and the rightmost terms in the above inequality do not depend on the particular f, and the inequality between them is valid on the event $\{\|\xi_1\|_{\mathcal{F}} \ge t + c\}$. Therefore, integrating the two sides w.r.t. ξ_1 on this event, we get

$$\inf_{f\in\mathcal{F}} \mathbf{P}_{\xi_2}\Big\{|\xi_2(f)| \le c\Big\} \cdot \mathbf{P}_{\xi_1}\Big\{\|\xi_1\|_{\mathcal{F}} \ge t + c\Big\} \le \mathbf{P}_{\xi_1,\xi_2}\Big\{\|\xi_1 - \xi_2\|_{\mathcal{F}} \ge t\Big\}.$$

Rearranging, we get (12.24). □

Part 4

Advanced Topics

CHAPTER 13

Stability of learning algorithms

Recall the abstract formulation of the learning problem in Section 6.1: we have a collection $Z_1, \ldots, Z_n$ of i.i.d. samples from some unknown distribution P on a set Z and a class $\mathcal{F}$ of functions $f : \mathsf{Z} \to [0,1]$. A learning algorithm is a sequence $A = \{A_n\}_{n=1}^{\infty}$ of mappings $A_n : \mathsf{Z}^n \to \mathcal{F}$ that take training data as input and generate functions in $\mathcal{F}$ as output. We say that A is consistent if

$$L(\widehat{f}_n) = \int_{\mathsf{Z}} \widehat{f}_n(z) P(\mathrm{d}z), \qquad \widehat{f}_n = A_n(Z^n)$$

converges in some sense to $L^* = \inf_{f \in \mathcal{F}} L(f)$, for any P. If a consistent algorithm exists, we say that the problem is learnable. Early on (see Theorem 5.3, which is essentially an application of the mismatched minimization lemma, Lemma 5.1), we have identified a sufficient condition for the existence of a consistent learning algorithm: uniform convergence of empirical means (UCEM). One way of stating the UCEM property is to require that

$$\sup_P \mathbf{E}_P \|P_n - P\|_{\mathcal{F}} \xrightarrow{n\to\infty} 0, \tag{13.1}$$

where the expectation is with respect to an i.i.d. process $Z_1, Z_2, \ldots$ with common marginal distribution P, P_n is the empirical distribution based on the first n samples of the process:

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i},$$

and $\|\cdot\|_{\mathcal{F}}$ is the seminorm defined by $\|P - P'\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |P(f) - P'(f)|$. By the method of Theorem 5.3 we know that, if $\mathcal{F}$ satisfies (13.1), then the ERM algorithm

$$\widehat{f}_n = \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(Z_i)$$

is consistent. In some cases, the UCEM property is both necessary and sufficient for learnability — for example, in the binary classification setting, where $Z = (X, Y)$ with arbitrary X, $Y \in \{0,1\}$, and $f(Z) = f(X,Y)$ taking values in $\{0,1\}$.

However, it is easy to see that, in general, one can have learnability without the UCEM property. For example, suppose that the function class $\mathcal{F}$ is such that one can find a function $\widetilde{f} \notin \mathcal{F}$ with the property that $\widetilde{f}(z) < \inf_{f\in\mathcal{F}} f(z)$ for every $z \in \mathsf{Z}$. Consider now a modified class $\widetilde{\mathcal{F}} = \mathcal{F} \cup \{\widetilde{f}\}$ obtained by adding $\widetilde{f}$ to $\mathcal{F}$. Then the ERM algorithm over $\widetilde{\mathcal{F}}$ will always return $\widetilde{f}$, and moreover $L(\widetilde{f}) \equiv L^*(\widetilde{\mathcal{F}})$. Thus, not only do we have consistency, but we also have perfect generalization, and the only condition the original class $\mathcal{F}$ has to satisfy is that we can find at least one $\widetilde{f}$ with the desired property. Of course, this imposes some minimal richness requirements on the ranges of all functions in $\mathcal{F}$ — for example, we could not pull this off when the functions in $\mathcal{F}$ are binary valued. And yet, the UCEM property is not required for perfect learnability!

So, what's going on here? It turns out that the main attraction of the UCEM property – namely, its algorithm-independence – is also its main disadvantage. Learnability is closely tied up with properties of learning algorithms: how well can they generalize? How good are they at rejecting obviously bad hypotheses and focusing on good ones? Thus, our goal is to connect learnability to certain properties of learning algorithms. This lecture is based primarily on a paper by Shalev-Shwartz et al. [**SSSS10**].

13.1. An in-depth view of learning algorithms

We adopt yet another abstract framework for machine learning in this section, first introduced by Vapnik (1995). A learning problem is denoted by $(\mathsf{Z}, \mathcal{P}, \mathcal{F}, \ell)$, where

- Z represents a set of possible values of data samples
- $\mathcal{P}$ is a set of probability distributions for Z-valued random variables
- $\mathcal{F}$ is a nonempty, closed, convex subset of a Hilbert space $\mathcal{H}$
- $\ell : \mathcal{F} \times \mathsf{Z} \to \mathbb{R}$ is a loss function; $\ell(f, z)$ is the loss for using f on sample z.

This notation is rather flexible:

- $\mathcal{F}$ could be a set of functions on Z. For example, we could have $\ell(f, z) = f(z)$, in which case, for a given $P \in \mathcal{P}$, we would be seeking to select f to minimize $\mathbf{E}_P[f]$ as in the abstract framework for ERM introduced in Section 6.1.
- It could be that $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$, and a sample $z = (x, y)$ represents a feature vector x and a label y. (This is the case for the next three examples.)
- We could consider $\ell(f, (x, y)) = \varphi(-yf(x))$, where φ is a penalty function, and then ℓ represents surrogate loss.
- Or $\ell(f, (x, y)) = (y - f(x))^2$, for the problem of regression with quadratic loss.
- The elements of $\mathcal{F}$ could also be considered as vectors of scalar parameters, with the number of such parameters equal to the dimension of the Hilbert space $\mathcal{H}$ containing $\mathcal{F}$. The loss function could correspond, for example, to a support vector machine type classifier, $\ell(f, (x, y)) = \varphi(-y\langle f, \psi(x)\rangle)$, where $\psi : \mathsf{X} \to \mathcal{H}$ maps an unlabeled data sample x to a feature vector $\psi(x)$.

For the examples above, if the penalty function φ is convex, then $\ell(f, z)$ is a convex function of f for each fixed z. We will still use the notation

$$L_P(f) = \mathbf{E}_P[\ell(f, Z)] \equiv \int_{\mathsf{Z}} \ell(f, z) P(\mathrm{d}z)$$

for the expected loss of f with respect to P, and will often omit the subscript P when it's clear from context. Given an n-tuple $Z^n = (Z_1, \dots, Z_n)$ of i.i.d. samples from P, we have the empirical loss

$$L_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i).$$

Finally, we define the minimum risk and empirical minimum risk by:

$$L^*(\mathcal{F}) := \inf_{f \in \mathcal{F}} L(f) \qquad \text{and} \qquad L_n^*(\mathcal{F}) := \inf_{f \in \mathcal{F}} L_n(f).$$

Here, $L^*(\mathcal{F})$ is a deterministic quantity that depends on ℓ, $\mathcal{F}$, and the underlying distribution P, whereas $L_n^*(\mathcal{F})$ is a random variable.

Also, let us define $\mathsf{Z}^* := \bigcup_{n=1}^{\infty} \mathsf{Z}^n$, i.e., Z^* is the collection of all tuples over Z. This definition allows us to treat a learning algorithm as a *single mapping* $A : \mathsf{Z}^* \to \mathcal{F}$ — the size of the training set is now clear from context. For example, $A(Z^n)$ is the output of A fed with an n-tuple $Z^n = (Z_1, \dots, Z_n)$, and so, in particular,

$$L(A(Z^n)) = \int_{\mathsf{Z}} \ell(A(Z^n), z) P(\mathrm{d}z)$$

is the expected loss of the function $A(Z^n) \in \mathcal{F}$ on a fresh sample $Z \sim P$, independent of Z^n. In contrast, $L_n(A(Z^n))$, given by

$$L_n(A(Z^n)) = \frac{1}{n} \sum_{i=1}^{n} \ell(A(Z^n), Z_i),$$

is the empirical loss of the algorithm output $A(Z^n)$ on the same sample Z^n that was supplied to A.

Our goal is to understand what makes a good learning algorithm. To keep things simple, we will focus on expected-value guarantees. We say that a learning algorithm A is *consistent* if

$$c_n(A) := \sup_P \mathbf{E}_P \left[L(A(Z^n)) - L^* \right] \xrightarrow{n\to\infty} 0. \tag{13.2}$$

We say that the learning problem specified by ℓ and $\mathcal{F}$ is learnable if there exists at least one consistent learning algorithm A. As we have already seen on multiple occasions, under certain conditions the ERM algorithm is consistent. A learning algorithm A is an *Asymptotic Empirical Risk Minimizer* (AERM) if

$$e_n(A) := \sup_P \mathbf{E}_P \left[L_n(A(Z^n)) - L_n^* \right] \xrightarrow{n\to\infty} 0. \tag{13.3}$$

Of course, if A is an exact ERM algorithm, then $e_n(A) = 0$ for all n, but there are many situations in which it is preferable to use AERM algorithms. Next, we say that A *generalizes* if

$$g_n(A) := \sup_P \mathbf{E}_P \left| L(A(Z^n)) - L_n(A(Z^n)) \right| \xrightarrow{n\to\infty} 0. \tag{13.4}$$

A weaker notion of generalization is as follows: A generalizes *on average* if

$$\bar{g}_n(A) := \sup_P \left| \mathbf{E}_P [L(A(Z^n)) - L_n(A(Z^n))] \right| \xrightarrow{n\to\infty} 0. \tag{13.5}$$

Our goal is to show, without requiring the UCEM property, that learnability is possible. Instead of relying on the UCEM property, we will investigate the relationship between the above properties of learning algorithms to *stability*, i.e., weak dependence of the algorithm output on any individual training sample. (We don't focus on examples such that the UCEM property provably does not hold – it is in fact not easy to come up with realistic examples for which UCEM can be disproved and yet consistency can be proved. The main point in this chapter is to not use the UCEM property in the proof of consistency.)

13.2. Learnability without uniform convergence

We now show that we can prove learnability without assuming uniform convergence:

THEOREM 13.1. *Suppose $\mathcal{F}$ is a convex subset of a Hilbert space $\mathcal{H}$, and suppose there are constants $L, m > 0$, such that, for every $z \in \mathsf{Z}$, the function $f \mapsto \ell(f, z)$ is m-strongly convex and L-Lipschitz. Then the ERM algorithm*

$$\widehat{f}_n = A(Z^n) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i)$$

is such that

$$L(\widehat{f}_n) - L^* \le \frac{2L^2}{\delta m n},$$

with probability at least $1 - \delta$.

If the mapping $f \mapsto \ell(f, z)$ is convex but not strongly convex, a variation of Theorem 13.1 can be obtained by using an additive regularization term, as described at the end of the section.

PROOF. The idea is to compare the output of A on the original training data Z^n to the output of A on the modified data, with one of the training samples replaced. Specifically, let $Z'_1, \ldots, Z'_n$ be n i.i.d. samples from P, independent of Z^n. For each i, define the modified training data

$$Z^n_{(i)} := (Z_1, \ldots, Z_{i-1}, Z'_i, Z_{i+1}, \ldots, Z_n).$$

For any $f \in \mathcal{F}$, consider the original and the perturbed empirical risks:

$$\begin{aligned} L_n(f) &:= \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i) \\ L^{(i)}_n(f) &:= \frac{1}{n} \ell(f, Z'_i) + \frac{1}{n} \sum_{j:\, j \neq i} \ell(f, Z_j) \\ &= L_n(f) + \frac{1}{n}\Big(\ell(f, Z'_i) - \ell(f, Z_i)\Big), \end{aligned}$$

Likewise, consider the ERM solutions for the original and the perturbed data:

$$\widehat{f}_n := \arg\min_{f \in \mathcal{F}} L_n(f), \qquad \widehat{f}^{(i)}_n := \arg\min_{f \in \mathcal{F}} L^{(i)}_n(f).$$

The function $f \mapsto \frac{1}{n}\left(\ell(f, Z'_i) - \ell(f, Z_i)\right)$ is $\frac{2L}{n}$-Lipschitz. Hence, by the stability of minimizers of strongly convex functions under Lipschitz perturbations (Lemma 3.2).

$$\|\widehat{f}_n - \widehat{f}^{(i)}_n\| \le \frac{2L}{mn}. \tag{13.6}$$

In other words, arbitrarily replacing any one sample in Z^n by some other Z'_i has only limited effect on the ERM solution, i.e., the algorithm output $A(Z^n)$ does not depend too much on any individual sample! But, because ℓ is Lipschitz, this implies that, for any $z \in \mathsf{Z}$,

$$\Big|\ell(\widehat{f}_n, z) - \ell(\widehat{f}^{(i)}_n, z)\Big| \le L\|\widehat{f}_n - \widehat{f}^{(i)}_n\| \le \frac{2L^2}{mn}. \tag{13.7}$$

We now claim that the *stability property* (13.7) implies

$$\mathbf{E}\left[L(\widehat{f}_n) - L_n(\widehat{f}_n)\right] \le \frac{2L^2}{mn}. \tag{13.8}$$

Indeed, since $\widehat{f}_n$ is a function of Z^n, and since $Z'_1, \ldots, Z'_n$ are independent of $Z_1, \ldots, Z_n$ and are draws from the same distribution, we can write

$$\mathbf{E}L(\widehat{f}_n) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[\ell(\widehat{f}_n, Z'_i)].$$

On the other hand, since, for every i, $\ell(\widehat{f}_n, Z_n)$ and $\ell(\widehat{f}_n^{(i)}, Z'_i)$ have the same distribution, we have

$$\mathbf{E}L_n(\widehat{f}_n) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[\ell(\widehat{f}_n, Z_i)] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[\ell(\widehat{f}_n^{(i)}, Z'_i)]$$

Therefore,

$$\begin{aligned} \mathbf{E}\left[L(\widehat{f}_n) - L_n(\widehat{f}_n)\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\ell(\widehat{f}_n, Z'_i) - \ell(\widehat{f}_n^{(i)}, Z'_i)\right] \\ &\le \frac{1}{n}\sum_{i=1}^n \sup_{z \in \mathsf{Z}} \left|\ell(\widehat{f}_n, z) - \ell(\widehat{f}_n^{(i)}, z)\right| \\ &\le \frac{2L^2}{mn}, \end{aligned} \tag{13.9}$$

as claimed. Eq. (13.8) shows that the ERM algorithm *generalizes well on average*, i.e., the empirical risk of $\widehat{f}_n = A(Z^n)$ on the data Z^n is a good estimate of $L(\widehat{f}_n) = L(A(Z^n))$ in expectation.

Now let $f^* \in \mathcal{F}$ achieve L^*. Since f^* doesn't depend on Z^n, we have $L^* = L(f^*) = \mathbf{E}[L_n(f^*)] \ge \mathbf{E}[L_n(\widehat{f}_n)]$, and therefore

$$\mathbf{E}\left[L(\widehat{f}_n) - L^*\right] \le \mathbf{E}\left[L(\widehat{f}_n) - L_n(\widehat{f}_n)\right] \le \frac{2L^2}{mn}.$$

From Markov's inequality, it then follows that

$$L(\widehat{f}_n) - L^* \le \frac{2L^2}{\delta mn}$$

with probability at least $1-\delta$, and we are done. □

The following variation of Theorem 13.1 relaxes the strong convexity assumption on the mapping $f \mapsto \ell(f,z)$ by switching to complexity-regularized ERM based on a penalty term.

THEOREM 13.2. *Let $\mathcal{F}$ be a convex and norm-bounded subset of a Hilbert space $\mathcal{H}$, i.e., there exists some $B < \infty$, such that $\|f\| \le B$ for all $f \in \mathcal{F}$. Suppose also that, for each $z \in \mathsf{Z}$, the function $f \mapsto \ell(f,z)$ is convex and L-Lipschitz (note: we are not assuming strong convexity). For each $\lambda > 0$, consider the complexity-regularized ERM algorithm*

$$\widehat{f}_{n,\lambda} = A_\lambda(Z^n) := \operatorname*{arg\,min}_{f \in \mathcal{F}} \left\{\frac{1}{n}\sum_{i=1}^n \ell(f, Z_i) + \frac{\lambda}{2}\|f\|^2\right\}.$$

Then $\widehat{f}_n = \widehat{f}_{n,\lambda}$ *with* $\lambda = \frac{2L}{B\sqrt{n}}$ *satisfies*

$$L(\widehat{f}_n) \le L^* + \frac{LB}{\sqrt{n}} + \frac{LB}{\delta\sqrt{n}} + \frac{8LB}{\delta n} \tag{13.10}$$

with probability at least $1-\delta$.

Proof. The idea is to apply Theorem 13.1 for the loss function $\ell_\lambda(f,z) \triangleq \ell(f,z) + \frac{\lambda}{2}\|f\|^2$. This function is λ-strongly convex. The function $\frac{\lambda}{2}\|f\|^2$ has gradient λf, and $\|\lambda f\| \le \lambda B$ for $f \in \mathcal{F}$, so $\frac{\lambda}{2}\|f\|^2$ is λB Lipschitz over $\mathcal{F}$, so for each z, $\ell_\lambda(f,z)$ is $L + \lambda B$ Lipschitz over $\mathcal{F}$. Therefore, Theorem 13.1 yields that with probability at least $1-\delta$,

$$L_\lambda(\widehat{f}_{n,\lambda}) \le L_\lambda^* + \frac{2(L+\lambda B)^2}{\delta \lambda n},$$

where L_λ^* is the minimum generalization regularized risk. Since $\frac{\lambda}{2}\|f\|^2 \le \frac{\lambda B^2}{2}$ for $f \in \mathcal{F}$ it follows that $L_\lambda^* \le L^* + \frac{\lambda B^2}{2}$. Also, $L(\widehat{f}_{n,\lambda}) \le L_\lambda(\widehat{f}_{n,\lambda})$ with probability one, and $\left(1 + \frac{2}{\sqrt{n}}\right)^2 \le 1 + \frac{8}{\sqrt{n}}$. Combining these observations and the choice of λ yields (13.10). □

13.3. Learnability and stability

The ERM algorithm considered in the previous system is stable with respect to replace one perturbations, in the sense of (13.7), due to the assumed strong convexity of the loss functions. It was then shown that the stability property implies that the algorithm generalizes well, in the sense of (13.8). The idea of this section is to focus on the fact that stability of a learning algorithm with respect to replace one perturbations implies that the algorithm generalizes well. That is, it does not overfit the data. We begin by revisting (13.9), with the notation changed to indicate the role of the algorithm A.

$$\mathbf{E}\left[L(A(Z^n))\right] - \mathbf{E}\left[L_n(A(Z^n)\right] = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n \ell(A(Z^n), Z_i') - \ell(A(Z_{(i)}^n), Z_i')\right] \tag{13.11}$$

To recall why (13.11) is true, observe that on the righthand side, for each i, the first term in the sum, $\ell(A(Z^n), Z_i')$, is the loss for A on a fresh data sample and the term $\ell(A(Z_{(i)}^n), Z_i')$ is the loss for A (trained on $Z_{(i)}^n$) on the i^{th} training sample used in the training data $Z_{(i)}^n$.

Next we state the definition of stability on average, and for convenience we restate the definition of generalization on average:

Definition 13.1. *An algorithm A is stable on average (with respect to replace-one operation) if*

$$\overline{s}_n(A) \triangleq \sup_P \frac{1}{n}\left|\sum_{i=1}^n \mathbf{E}\left[\ell(A(Z^n), Z_i') - \ell(A(Z_{(i)}^n), Z_i')\right]\right| \stackrel{n\to\infty}{\longrightarrow} 0. \tag{13.12}$$

An algorithm A generalizes on average if

$$\overline{g}_n(A) \triangleq \sup_P \left|\mathbf{E}\left[L(A(Z^n))\right] - \mathbf{E}\left[L_n(A(Z^n)\right]\right| \stackrel{n\to\infty}{\longrightarrow} 0.$$

Lemma 13.1. *For any learning algorithm, $\overline{s}_n(A) = \overline{g}_n(A)$. In particular, A is stable on average if and only if it generalizes on average.*

PROOF. Take the absolute value and supremum over P on each side of (13.11) to get $\overline{s}_n(A) = \overline{g}_n(A)$. $\square$

DEFINITION 13.2. *An algorithm A is consistent if*

$$c_n(A) \triangleq \sup_P \mathbf{E}\left[L(A(Z^n)) - L^*\right] \overset{n\to\infty}{\longrightarrow} 0. \tag{13.13}$$

In some applications it is too demanding to find an ERM algorithm, but algorithms satisfying the weaker property in the following definition may be available.

DEFINITION 13.3. *An algorithm A is an asymptotic empirical risk minimizer (AERM) if*

$$e_n(A) \triangleq \sup_P \mathbf{E}\left[L_n(A(Z^n)) - L_n^*\right] \overset{n\to\infty}{\longrightarrow} 0, \tag{13.14}$$

where L_n^* is the (random) minimum empirical risk: $L_n^* = L_n^*(\mathcal{F}) = \inf_{f\in\mathcal{F}} L_n(f)$.

REMARK 13.1. *Note that in (13.13), the notation L is short for L_P and L^* is short for L_P^*. Since $L(A(Z^n)) - L^* \geq 0$ for any Z^n, any P, and any algorithm A, $c_n(A) = \sup_P \mathbf{E}\left[\left|L(A(Z^n)) - L^*\right|\right]$. Similarly, $L_n(A(Z^n)) - L_n^* \geq 0$ for any Z^n and any algorithm A, so $e_n(A) = \sup_P \mathbf{E}\left[\left|L_n(A(Z^n)) - L_n^*\right|\right]$.*

The following is perhaps the most useful result about the virtues of stability:

THEOREM 13.3. *For any algorithm A, $c_n(A) \leq \overline{s}_n(A) + e_n(A)$. Therefore, an AERM algorithm that is stable on average is consistent.*

PROOF. For any algorithm A and any probability distribution P, by the definitions of $\overline{g}_n(A)$ and $e_n(A)$,

$$\begin{aligned}
\mathbf{E}\left[L(A(Z^n))\right] &\leq \mathbf{E}\left[L_n(A(Z^n))\right] + \overline{g}_n(A) \\
&\leq \mathbf{E}\left[L_n^*\right] + e_n(A) + \overline{g}_n(A) \\
&\leq L^* + e_n(A) + \overline{g}_n(A).
\end{aligned}$$

Together with the fact $\overline{g}_n(A) = \overline{s}_n(A)$, this implies that $c_n(A) \leq \overline{s}_n(A) + e_n(A)$ as claimed. $\square$

Theorem 13.3 implies that stability on average, or equivalently, generalization on average, is sufficient for an ERM algorithm to be consistent. It turns out that an ERM algorithm generalizes on average (see (13.4) for definition) if and only if it generalizes, as shown next.

PROPOSITION 13.1. *If A is an AERM algorithm that generalizes on average, then it generalizes, and moreover*

$$g_n(A) \leq \bar{g}_n(A) + 2e_n(A) + \frac{2}{\sqrt{n}}. \tag{13.15}$$

PROOF. We begin by decomposing the difference $L_n(A(Z^n)) - L(A(Z^n))$:

$$\begin{aligned}
&L_n(A(Z^n)) - L(A(Z^n)) \\
&= L_n(A(Z^n)) - L_n^* + \underbrace{L_n^* - L_n(f^*)}_{\leq 0} + L_n(f^*) - L(f^*) + \underbrace{L(f^*) - L(A(Z^n))}_{\leq 0} \\
&\leq L_n(A(Z^n)) - L_n^* + L_n(f^*) - L(f^*).
\end{aligned}$$

Applying Lemma 13.5 in the Appendix to $U := L_n(A(Z^n)) - L(A(Z^n))$ and $V := L_n(A(Z^n)) - L_n^* + L_n(f^*) - L(f^*)$, we get

$$\begin{aligned}
&\mathbf{E}\,|L_n(A(Z^n)) - L(A(Z^n))| \\
&\le |\mathbf{E}[L_n(A(Z^n)) - L(A(Z^n))]| + 2\mathbf{E}\,|L(A(Z^n)) - L_n^* + L_n(f^*) - L(f^*)| \\
&\le |\mathbf{E}[L_n(A(Z^n)) - L(A(Z^n))]| + 2\mathbf{E}\,|L(A(Z^n)) - L_n^*| + 2\mathbf{E}|L_n(f^*) - L(f^*)| \\
&\le \bar{g}_n(A) + 2e_n(A) + \frac{2}{\sqrt{n}},
\end{aligned}$$

where in the last line we have used the assumed properties of A, together with the fact that, for any f, $\mathbf{E}|L_n(f) - L(f)| = \mathbf{E}|L_n(f) - \mathbf{E}L_n(f)| \le \sqrt{\mathbf{E}(L_n(f) - \mathbf{E}L_n(f)^2)} \le \frac{1}{\sqrt{n}}$, since ℓ is bounded between 0 and 1. This completes the proof. □

13.4. Stability of stochastic gradient descent

One of the most popular algorithms for learning over complicated hypothesis classes (such as deep neural networks) is the Stochastic Gradient Descent (SGD) algorithm. The basic idea behind SGD is as follows. For a fixed training set $Z^n = (Z_1, \ldots, Z_n)$, the usual ERM approach requires minimizing the function

$$L_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i) \tag{13.16}$$

over the hypothesis class $\mathcal{F}$. One way to go about this is to use gradient descent: assuming that the function $f \mapsto \ell(f, z)$ is differentiable for each $z \in \mathsf{Z}$, we can set the initial condition $f_0 \in \mathcal{F}$ and iteratively compute

$$f_t = \Pi\left(f_{t-1} - \alpha_t \nabla L_n(f_{t-1})\right), \qquad t = 1, 2, \ldots \tag{13.17}$$

where $\Pi : \mathcal{H} \to \mathcal{F}$ is the projection operator onto $\mathcal{F}$,

$$\nabla L_n(f_{t-1}) = \frac{1}{n}\sum_{i=1}^n \nabla\ell(f_{t-1}, Z_i)$$

is the gradient of L_n at f_{t-1}, and $\{\alpha_t\}_{t=1}^\infty$ is a monotonically decreasing sequence of positive reals typically referred to as step sizes. Under certain mild conditions on ℓ and $\mathcal{F}$, and with appropriately tuned step sizes, one can guarantee that

$$L_n(f_t) \to \inf_{f\in\mathcal{F}} L_n(f) \equiv L_n^* \qquad \text{as } t \to \infty.$$

In other words, for each n, we can find a large enough T_n, such that $A(Z^n) = f_{T_n}$ is an AERM algorithm.

However, one disadvantage of gradient descent is that each update (13.17) requires a sweep through the entire sample Z^n in order to compute the gradient ∇L_n. Thus, the complexity of each step of the gradient descent method scales as $O(n)$. SGD offers a way around this limitation and allows to reduce the complexity of each iteration to $O(1)$. If we look at (13.16), we see that the empirical loss $L_n(f)$ can be written as an average of n functions of f:

$$L_n(f) = \frac{1}{n}\sum_{i=1}^n \ell_i(f), \qquad \text{where } \ell_i(f) := \ell(f, Z_i).$$

In each iteration of SGD, we pick a random index $I_t \in \{1, \dots, n\}$ and update

$$f_t = \Pi\left(f_{t-1} - \alpha_t \nabla \ell_{I_t}(f_{t-1})\right) \equiv \Pi\left(f_{t-1} - \alpha_t \nabla \ell(f_{t-1}, Z_{I_t})\right). \tag{13.18}$$

Thus, SGD is a *randomized* algorithm. We will assume from now on that $(I_t, t = 1, 2, \dots)$ is a random process with values in $[n]$ that is statistically invariant with respect to permutations of the values. That is, for any permutation $\pi : [n] \to [n]$, $(I_t : t = 1, 2, \dots)$ has the same distribution as $(\pi \circ I_t : t = 1, 2, \dots)$. For example, the selections I_t could be independent and uniformly distributed over $[n]$, or the sequence could be periodic with period n such that $I_1, \dots, I_n$ is a random permutation of $[n]$, or the blocks of the form $I_{kn+1} \dots, I_{kn+n}$ could be independent permutations of $[n]$ for all $k \geq 0$.

In a recent paper, Hardt, Recht, and Singer [**HRS16**] have shown that SGD with suitably tuned step sizes and number of updates gives a stable learning algorithm. Under different assumptions on the loss function ℓ, we end up with different conditions for stability. In order to proceed, let us first examine the evolution of SGD updates for a fixed training set Z^n. Fix a differentiable function $\varphi : \mathcal{F} \to \mathbb{R}$ and a step size $\alpha \geq 0$, and define an operator $G_{\varphi,\alpha} : \mathcal{F} \to \mathcal{F}$ by

$$G_{\varphi,\alpha}(f) := \Pi\left(f - \alpha \nabla \varphi(f)\right). \tag{13.19}$$

Then we can write the tth update of SGD as

$$f_t = G_t(f_{t-1}), \qquad \text{where } G_t := G_{\ell(\cdot, Z_{I_t}), \alpha_t}. \tag{13.20}$$

Now let us fix some $i^* \in \{1, \dots, n\}$ and consider running SGD with the same realization of the random indices $\{I_t\}$ on another training set $Z'^n = (Z'_1, \dots, Z'_n)$, where $Z'_{i^*} \neq Z_{i^*}$ and $Z'_j = Z_j$ for all $j \neq i^*$. Denoting by $\{f'_t\}$ the corresponding updates with $f'_0 = f_0$, we can write

$$f'_t = G'_t(f'_{t-1}), \qquad \text{where } G'_t := G_{\ell(\cdot, Z'_{I_t}), \alpha_t}. \tag{13.21}$$

For each $t = 0, 1, \dots$, let $\delta_t := \|f_t - f'_t\|$, with the initial condition $\delta_0 = 0$. Define the following quantities:

$$\eta_t := \sup_{f, f' \in \mathcal{F}} \frac{\|G_t(f) - G_t(f')\|}{\|f - f'\|} \tag{13.22}$$

and

$$c_t := \sup_{f \in \mathcal{F}} \left(\|G_t(f) - f\| \vee \|G'_t(f) - f\|\right). \tag{13.23}$$

We can now track the evolution of δ_t as follows:

- If $I_t \neq i^*$, then $\ell(\cdot, Z_{I_t}) = \ell(\cdot, Z'_{I_t})$, and therefore

$$\begin{aligned}
\delta_t &= \|f_t - f'_t\| \\
&= \|G_t(f_{t-1}) - G'_t(f'_{t-1})\| \\
&= \|G_t(f_{t-1}) - G_t(f'_{t-1})\| \\
&\leq \eta_t \|f_{t-1} - f'_{t-1}\|.
\end{aligned}$$

- If $I_t = i^*$, then

$$\begin{aligned}\delta_t &= \|G_t(f_{t-1}) - G'_t(f'_{t-1})\| \\ &\le \|G_t(f_{t-1}) - G_t(f'_{t-1})\| + \|G_t(f'_{t-1}) - G'_t(f'_{t-1})\| \\ &\le \eta_t\|f_{t-1} - f'_{t-1}\| + \|G_t(f'_{t-1}) - f'_{t-1}\| + \|G'_t(f'_{t-1}) - f'_{t-1}\| \\ &\le \eta_t\|f_{t-1} - f'_{t-1}\| + 2c_t.\end{aligned}$$

We can combine these two cases into a single inequality:

$$\delta_t \le \eta_t\delta_{t-1} + 2c_t\mathbf{1}_{\{I_t=i^*\}}. \tag{13.24}$$

This will be our main tool for analyzing the stability of SGD.

We also need to use the contraction property of the gradient update map. As shown in Proposition 4.2, projection onto a closed convex subset of a Hilbert space is nonexpansive. Thus, if the gradient descent map of Lemma 3.4 is followed by projection onto a convex subset of the Hilbert space as in this section, the contraction properties of Lemma 3.4 hold with the projection included. Therefore, if $\alpha_t \le \frac{2}{M}$, then $\eta_t \le 1$ for all t, and, if in addition, φ is m-strongly convex and $\alpha_t \le \frac{1}{M}$, then $\eta_t \le 1 - \frac{\alpha_t m}{2}$.

Now we can analyze the stability of SGD under various assumptions on the loss ℓ.

THEOREM 13.4. *Suppose that, for all $z \in \mathsf{Z}$, the function $f \mapsto \ell(f,z)$ is convex, M-smooth, and L-Lipschitz. If SGD is run with $\alpha_t \le \frac{2}{M}$ for T time steps, then, for any two datasets Z^n and Z'^n that differ only in one sample,*

$$\sup_{z\in\mathsf{Z}} \mathbf{E}\,|\ell(f_T,z) - \ell(f'_T,z)| \le \frac{2L^2}{n}\sum_{t=1}^T \alpha_t, \tag{13.25}$$

where the expectation is only with respect to the internal randomness of SGD, i.e., the selection process (I_t).

PROOF. Let $i^* \in [n]$ be the coordinate where Z^n and Z'^n differ. By the contraction property of the gradient descent map discussed above (i.e. $\eta_t \le 1$) and the assumption that φ is L-Lipschitz,

$$\begin{aligned}\|G_{\varphi,\alpha}(f) - f\| &= \|\Pi(f - \alpha\nabla\varphi(f)) - \Pi(f)\| \\ &\le \alpha\|\nabla\varphi(f)\| \\ &\le \alpha L,\end{aligned}$$

so $c_t \le \alpha_t L$ for all t. Using these two estimates in (13.24) gives

$$\delta_t \le \delta_{t-1} + 2\alpha_t L\mathbf{1}_{\{I_t=i^*\}}. \tag{13.26}$$

By the symmetry assumption on (I_t), $\mathbf{P}[I_t = i^*] = \frac{1}{n}$ for all i^* and all t. Therefore, taking expectations of both sides of (13.26) with respect to (I_t), we get

$$\mathbf{E}[\delta_t] \le \mathbf{E}[\delta_{t-1}] + \frac{2\alpha_t L}{n}.$$

Since $\delta_0 = 0$, we obtain

$$\mathbf{E}[\delta_T] \le \frac{2L}{n}\sum_{t=1}^T \alpha_t. \tag{13.27}$$

Finally, using the Lipschitz continuity of ℓ, we get for any $z \in \mathsf{Z}$

$$\mathbf{E}|\ell(f_T, z) - \ell(f'_T, z)| \le L \cdot \mathbf{E}[\delta_T] \tag{13.28}$$

$$\le \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t. \tag{13.29}$$

Since z was arbitrary, we have (13.25). □

For example, if we set $T = n$ and $\alpha_t = \frac{2}{M\sqrt{n}}$ for all t, then

$$\sum_{t=1}^{n} \alpha_t = \frac{2\sqrt{n}}{M}, \tag{13.30}$$

and then the algorithm $A(Z^n)$ obtained by running SGD for n steps with constant step size $\alpha = 2/M\sqrt{n}$ is stable with $\bar{s}_n(A) \le \frac{4L^2}{M\sqrt{n}}$.

If we now assume that ℓ is also strongly convex, we get a bound that does not depend on the number of iterations T:

Theorem 13.5. *Suppose that ℓ satisfies the conditions of Theorem 13.4, and also that the function $f \mapsto \ell(f, z)$ is m-strongly convex for each $z \in \mathsf{Z}$. Suppose that we run SGD with a constant step size $\alpha \le \frac{1}{M}$ for T time steps. Then, for any two datasets Z^n and Z'^n that differ in only one sample,*

$$\sup_{z \in \mathsf{Z}} \mathbf{E}|\ell(f_T, z) - \ell(f'_T, z)| \le \frac{4L^2}{mn}, \tag{13.31}$$

where the expectation is only with respect to the internal randomness of SGD (i.e., index selection).

Proof. The proof is similar to the proof of Theorem 13.4. First of all, under our assumptions on ℓ and on α, by the contraction property of the gradient descent map and the Lipschitz assumptions,

$$\eta_t \le 1 - \frac{\alpha m}{2} \qquad \text{and} \qquad c_t \le \alpha L. \tag{13.32}$$

Using this in (13.24), we have

$$\delta_t \le \left(1 - \frac{\alpha m}{2}\right) \delta_{t-1} + 2\alpha L \cdot \mathbf{1}_{\{I_t = i^*\}}. \tag{13.33}$$

Taking expectations of both sides with respect to (I_t) yields

$$\mathbf{E}[\delta_t] \le \left(1 - \frac{\alpha m}{2}\right) \mathbf{E}[\delta_{t-1}] + \frac{2\alpha L}{n}, \tag{13.34}$$

with the initial condition $\delta_0 = 0$. Unwinding the recursion, we get

$$\mathbf{E}[\delta_T] \le \frac{2\alpha L}{n} \sum_{t=1}^{T} \left(1 - \frac{\alpha m}{2}\right)^{t-1} \le \frac{2\alpha L}{n} \cdot \frac{2}{\alpha m} = \frac{4L}{mn}. \tag{13.35}$$

The result follows. □

Finally, we derive a stability estimate for SGD without requiring convexity, but still assuming Lipschitz-continuity and smoothness:

THEOREM 13.6. *Suppose that, for each $z \in \mathsf{Z}$, the loss function $f \mapsto \ell(f, z)$ is M-smooth, L-Lipschitz, and bounded between 0 and 1. Suppose that we run SGD with $I_t \overset{\text{i.i.d.}}{\sim} \text{Uniform}([n])$ and step sizes $\alpha_t \le c/t$ for T time steps, where $c > 0$ is some constant. Then, for any two datasets Z^n and Z'^n that differ in only one sample,*

$$\sup_{z \in \mathsf{Z}} \mathbf{E}|\ell(f_T, z) - \ell(f'_T, z)| \le \frac{1 + 1/Mc}{n}(2cL^2)^{\frac{1}{Mc+1}} T^{\frac{Mc}{Mc+1}}, \tag{13.36}$$

where the expectation is only with respect to the internal randomness of SGD (i.e., index selection).

REMARK 13.2. *The assumption that I_t are i.i.d. is necessary here.*

PROOF. If φ is L-Lipschitz continuous and M-smooth, then for any f, f' and $\alpha > 0$,

$$\begin{aligned}\|G_{\varphi,\alpha}(f) - G_{\varphi,\alpha}(f')\| &= \|f - f' - \alpha(\nabla\varphi(f) - \nabla\varphi(f'))\| \\ &\le \|f - f'\| + \alpha\|\nabla\varphi(f) - \nabla\varphi(f')\| \le (1 + \alpha M)\|f - f'\|.\end{aligned}$$

Therefore, $\eta_t \le 1 + \alpha_t M \le 1 + \frac{Mc}{t}$ for all t. The Lipschitz assumption and choice of α_t gives $c_t \le \frac{cL}{t}$. Taking expectations on each side of (13.24) yields

$$\mathbf{E}[\delta_t] \le \left(1 + \frac{cM}{t}\right)\mathbf{E}[\delta_{t-1}] + \frac{2Lc}{nt}. \tag{13.37}$$

Using only (13.37) and the initial condition $\delta_0 = 0$ does not give a good bound on $\mathbf{E}[\delta_T]$ because of the large step sizes for small t. To proceed, let $t_0 \in \{0, \dots, n\}$. We know $\delta_{t_0} \neq 0$ only if sample i^* is used by time t_0 (i.e. only if $I_t = i^*$ for some t with $1 \le t \le t_0$). So, since independent uniform sampling is assumed, $\mathbf{P}[\delta_{t_0} = 0] \ge \left(1 - \frac{1}{n}\right)^{t_0} \ge 1 - \frac{t_0}{n}$. Using the the fact $|\ell_T(f_T) - \ell_T(f'_T)| \le L\delta_T$ on the event $\delta_{t_0} = 0$ and the fact $|\ell_T(f_T) - \ell_T(f'_T)| \le 1$ in general,

$$\mathbf{E}|\ell_T(f_T) - \ell_T(f'_T)| \le L\Delta_T + \frac{t_0}{n}, \tag{13.38}$$

where $\Delta_t := \mathbf{E}[\delta_t | \delta_{t_0} = 0]$ for $t \ge t_0$. The same reasoning that led to (13.37) yields

$$\begin{aligned}\Delta_t &\le \left(1 + \frac{cM}{t}\right)\Delta_{t-1} + \frac{2Lc}{nt} \\ &\le \exp\left(\frac{Mc}{t}\right)\Delta_{t-1} + \frac{2Lc}{nt} \qquad \text{for } t \ge t_0 \end{aligned} \tag{13.39}$$

where in the last line we have used the inequality $1 + u \le \exp(u)$. Unwinding the recursion down to $t = t_0 + 1$ and using the initial condition $\Delta_{t_0} = 0$, we have

$$\begin{aligned}
\Delta_T &\le \sum_{t=t_0+1}^{T} \prod_{k=t+1}^{T} \exp(Mc/k) \frac{2cL}{tn} \\
&= \sum_{t=t_0+1}^{T} \exp\left(Mc \sum_{k=t+1}^{T} \frac{1}{k} \right) \frac{2cL}{tn} \\
&\le \sum_{t=t_0+1}^{T} \exp\left(Mc \log \frac{T}{t} \right) \frac{2cL}{tn} \\
&\le \frac{2cL}{n} T^{Mc} \sum_{t=t_0+1}^{T} t^{-(1+Mc)} \\
&\le \frac{2cL}{n} T^{Mc} \frac{1}{Mc} \left(t_0^{-Mc} - T^{-Mc} \right) \\
&\le \frac{2L}{nM} \left(\frac{T}{t_0} \right)^{Mc}.
\end{aligned}$$

Plugging this estimate into (13.38), we get

$$\mathbf{E}\left[|\ell(f_T, z) - \ell(f'_T, z)| \right] \le \frac{t_0}{n} + \frac{2L^2}{nM} \left(\frac{T}{t_0} \right)^{Mc}.$$

The right-hand side is (approximately) minimized by setting

$$t_0 = \left(2cL^2 \right)^{\frac{1}{q+1}} T^{\frac{q}{q+1}}, \qquad q = Mc$$

which gives

$$\mathbf{E}\left[|\ell(f_T, z) - \ell(f'_T, z)| \right] \le \frac{1 + 1/Mc}{n} (2cL^2)^{\frac{1}{Mc+1}} T^{\frac{Mc}{Mc+1}}.$$

□

In this case, we can set $T = n^{(1-\varepsilon)(Mc+1)/Mc}$ for any $\varepsilon \in (0, 1)$, and obtain the stability bound

$$s_n(A) \le (1 + 1/Mc)(2cL^2)^{\frac{1}{Mc+1}} n^{-\varepsilon} \tag{13.40}$$

for $A(Z^n) = f_T$.

13.5. Analysis of Stochastic Gradient Descent

This section summarizes an analysis of stochastic gradient descent following [BCN16]. Convergence results are presented first under smoothness and strong convexity assumptions, and under mild assumptions on the stochastic gradient. Then the convexity assumption is dropped, allowing for the possibility of convergence to a local minimum, but it is shown that the mean square gradient still converges to zero in an average sense.

The setting is the following.

- $\mathcal{F}$ is a Hilbert space, possibly a finite dimensional Euclidian space $\mathbb{R}^d$

- $\Gamma : \mathcal{F} \to \mathbb{R}$ is continuously differentiable, with a finite infimum Γ^*
- $f_1 \in \mathcal{F}$ (Initial state)
- $(\xi_t : t = 1, 2, \ldots)$ is a sequence of mutually independent random variables
- $g(f_t, \xi_t)$ is a random element of the Hilbert space for each t. It is a stochastic gradient of Γ evaluated at f_t.

The *stochastic gradient descent* (SGD) algorithm is given by the update

$$f_{t+1} = f_t - \alpha_t g(f_t, \xi_t)$$

such that $(\alpha_t : t = 1, 2, \ldots)$ is a sequence of nonnegative stepsizes.

Example 13.1. *The function Γ could have the form of an empirical risk: $\Gamma(f) = \frac{1}{n}\sum_{i=1}^n \ell(f, z_i)$ for some deterministic data sequence z^n (perhaps coming by sampling n times from some distribution) and ℓ is a loss function. The random variables (ξ_t) could be independent, and each uniformly distributed over $[n]$ and $g(\cdot, \xi_t) = \nabla \ell(\cdot, z_{\xi_t})$. More generally, the random variables could be independent and each uniformly distributed over $\binom{[n]}{k}$, the set of subsets of $[n]$ of cardinality k (k is the batch size), and*

$$g(\cdot, \xi_t) = \frac{1}{|\xi_t|}\sum_{j \in \xi_t} \nabla \ell(\cdot, z_j).$$

Assumption 13.1. *(i) There exists $\mu > 0$ such that for all $t \geq 1$,*

$$\langle \nabla \Gamma(f_t), \mathbf{E}_{\xi_t}[g(f_t, \xi_t)]\rangle \geq \mu \|\nabla \Gamma(f_t)\|^2. \tag{13.41}$$

(ii) There exist $B \geq 0$ and $B_G \geq 0$ such that

$$\mathbf{E}_{\xi_t}[\|g(f_t, \xi_t)\|^2] \leq B + B_G \|\nabla \Gamma(f_t)\|^2. \tag{13.42}$$

The assumption (13.41) would be satisfied with $\mu = 1$ if, given f_t, the stochastic gradient is an unbiased estimator of the actual gradient of Γ: $\mathbf{E}_{\xi_t}[g(f_t, \xi_t)] = \nabla \Gamma(f_t)$. If $B > 0$, the stochastic gradient can continue to exhibit variance bounded away from zero, even as the actual gradient becomes arbitrarily small. It can be shown that if $\|\nabla \Gamma(f_t)\|$ can be arbitrarily large, then (13.41) and (13.42) imply that $B_G \geq \mu^2$. Also, if $B = 0$ and $\mu = B_G = 1$, the SGD algorithm must be the deterministic gradient descent algorithm with stepsizes (α_t).

Theorem 13.7. *(SGD with fixed stepsize for smooth, strongly convex objective function) Suppose Γ is M-smooth and m-strongly convex for some $M, m > 0$, and suppose Assumption 13.1 holds. Consider SGD run with fixed stepsize α such that $0 < \alpha \leq \frac{\mu}{MB_G}$. Then the optimality gap satisfies*

$$\begin{aligned}\mathbf{E}\Gamma(f_t) - \Gamma^* &\leq \frac{\alpha M B}{2m\mu} + (1 - \alpha m \mu)^{t-1}\left(\Gamma(f_1) - \Gamma_* - \frac{\alpha M B}{2m\mu}\right) \\ &\xrightarrow{t \to \infty} \frac{\alpha M B}{2m\mu}\end{aligned} \tag{13.43}$$

Proof. By the M smoothness assumption,

$$\begin{aligned}\Gamma(f_{t+1}) - \Gamma(f_t) &\leq \langle \nabla \Gamma(f_t), f_{t+1} - f_t\rangle + \frac{M}{2}\|f_{t+1} - f_t\|^2 \\ &= -\alpha \langle \nabla \Gamma(f_t), g(f_t, \xi_t)\rangle + \frac{\alpha^2 M}{2}\|g(f_t, \xi_t)\|^2\end{aligned}$$

Taking expectation with respect to the randomness of ξ_t, invoking Assumption 13.1, and using the assumption $0 < \alpha \leq \frac{\mu}{MB_G}$ yields

$$\begin{aligned}\mathbf{E}_{\xi_t}[\Gamma(f_{t+1})] - \Gamma(f_t) &\leq -\mu\alpha\|\nabla\Gamma(f_t)\|^2 + \frac{\alpha^2 M}{2}\left(B + B_G\|\nabla\Gamma(f_t)\|^2\right) \\ &= -\alpha\left(\mu - \frac{\alpha M B_G}{2}\right)\|\nabla\Gamma(f_t)\|^2 + \frac{\alpha^2 MB}{2} \qquad (13.44)\\ &\leq -\frac{\alpha\mu}{2}\|\nabla\Gamma(f_t)\|^2 + \frac{\alpha^2 MB}{2}. \qquad (13.45)\end{aligned}$$

The bound (13.45) shows how smoothness can be translated into a bound on descent to be expected in terms of the norm of the gradient at f_t. The next step is to use strong convexity, which implies that the norm of the gradient is lower bounded by $\Gamma(f_t)-\Gamma(f^*)$.) The m-strong convexity of Γ implies (see Lemma 3.2(5)) :

$$\Gamma(f_t) - \Gamma^* \leq \frac{\|\nabla\Gamma(f_t)\|^2}{2m}$$

which, when substituted into (13.45), yields

$$\mathbf{E}_{\xi_t}[\Gamma(f_{t+1})] - \Gamma(f_t) \leq -\alpha m\mu(\Gamma(f_t) - \Gamma^*) + \frac{\alpha^2 MB}{2}$$

Subtracting Γ^* from both sides, taking total expectations, and rearranging, yields

$$\mathbf{E}\Gamma(f_{t+1}) - \Gamma^* \leq (1 - \alpha m\mu)\left(\mathbf{E}\Gamma(f_t) - \Gamma^*\right) + \frac{\alpha^2 MB}{2},$$

from which (13.43) follows by induction on t. □

The bound (13.43) on the expected optimality gap does not converge to zero as $t \to \infty$ because even near a minimizer the gradient is permitted to be noisy if $B > 0$. If diminishing stepsizes are used then essentially the law of large numbers helps out to make the expected optimality gap converge to zero. The classical schedule of Robbins and Monroe is to select (α_t) so that

$$\sum_{t=1}^{\infty} \alpha_t = \infty \text{ and } \sum_{t=1}^{\infty} \alpha_t^2 \leq \infty. \qquad (13.46)$$

The first condition in (13.46) is needed to ensure that the sequence f_t can move arbitrarily far, within regions where the gradient is bounded. The second condition is equivalent to $\sum_{t=T}^{\infty} \alpha_t^2 \to 0$ as $T \to \infty$, so that for very large T, the total variance of noise in the future of the stochastic gradient is converging to zero.

Theorem 13.8. *(SGD with diminishing stepsize for smooth, strongly convex objective function) Suppose Γ is M-smooth and m-strongly convex for some $M, m > 0$. Suppose Assumption 13.1 holds. Consider SGD run with stepsizes $\alpha_t = \frac{c}{\gamma+t}$ such that $c > \frac{1}{m\mu}$ and γ is large enough that $\alpha_1 \leq \frac{\mu}{MB_G}$. Then for all $t \geq 1$ the optimality gap satisfies*

$$\mathbf{E}\Gamma(f_t) - \Gamma^* \leq \frac{\nu}{\gamma + t} \qquad (13.47)$$

where

$$\nu \triangleq \max\left\{\frac{c^2MB}{2(cm\mu-1)}, (\gamma+1)(\Gamma(f_1)-\Gamma^*).\right\} \tag{13.48}$$

PROOF. Note that $\alpha_t \le \frac{\mu}{MB_G}$ for all $t \ge 1$, so the proof of Theorem 13.7 goes through with α replaced by α_t to yield

$$\begin{aligned}
\mathbf{E}\Gamma(f_{t+1}) - \Gamma^* &\le (1-\alpha_t m\mu)\left(\mathbf{E}\Gamma(f_t)-\Gamma^*\right) + \frac{\alpha_t^2 MB}{2},\\
&= \left(1-\frac{cm\mu}{\gamma+t}\right)\left(\mathbf{E}\Gamma(f_t)-\Gamma^*\right) + \frac{c^2MB}{2(\gamma+t)^2}\\
&\le \left(1-\frac{cm\mu}{\gamma+t}\right)\left(\mathbf{E}\Gamma(f_t)-\Gamma^*\right) + \frac{\nu(cm\mu-1)}{(\gamma+t)^2}, \qquad (13.49)
\end{aligned}$$

where the last step uses the fact $\frac{c^2MB}{2} \le \nu(cm\mu-1)$. Finally, (13.47) holds for $t=1$ by the choice of ν, and it follows for all $t \ge 1$ from (13.49) by induction on t. □

SGD without convexity. The following drops the assumption that the objective function is convex. The objective function is still assumed to be bounded below (i.e. have a finite infimum, Γ^*.) The proof follows the first half of the proof of Theorem 13.7, which uses the smoothness assumption. With the convexity assumption dropped, the algorithm could become stuck near a local minimum. If it is near a local minimum for a long time, its step sizes will get small because the gradient will be small. After that it might escape to a new local minimum and for a time have larger gradients. Since Γ is assumed to be bounded below, however, the algorithm can't forever experience large gradients on average.

THEOREM 13.9. *(SGD with fixed stepsize for smooth objective function) Suppose Γ is M-smooth for some $M>0$, and suppose Assumption 13.1 holds. Consider SGD run with fixed stepsize α such that $0<\alpha\le\frac{\mu}{MB_G}$. Then the expected average-squared-gradients of Γ corresponding to the SG iterates satisfies the following inequality for all $T\ge 1$.*

$$\mathbf{E}\frac{1}{T}\sum_{t=1}^{T}\|\nabla\Gamma(f_t)\|^2 \le \frac{\alpha MB}{\mu} + \frac{2(\Gamma(f_1)-\Gamma(f^*))}{T\mu\alpha} \tag{13.50}$$

$$\xrightarrow{T\to\infty} \frac{\alpha MB}{\mu}. \tag{13.51}$$

PROOF. Taking the total expectation of each side of (13.45) yields

$$\mathbf{E}\Gamma(f_{t+1}) - \mathbf{E}\Gamma(f_t) \le -\frac{\alpha\mu}{2}\mathbf{E}\|\nabla\Gamma(f_t)\|^2 + \frac{\alpha^2MB}{2}. \tag{13.52}$$

Summing each side of (13.52) over $1\le t\le T$, using the fact $\Gamma(f_{T+1})\ge\Gamma^*$, and dividing by T yields (13.50). □

Here is a variation with decreasing step sizes.

THEOREM 13.10. *(SGD with diminishing stepsize for smooth objective function) Suppose Γ is M-smooth for some $M>0$, and suppose Assumption 13.1 holds. Consider SGD run* *with stepsizes* $\alpha_t \geq 0$ *such that* $A_T \triangleq \sum_{t=1}^T \alpha_t \stackrel{T\to\infty}{\longrightarrow} \infty$, $\sum_{t=1}^\infty \alpha_t^2 < \infty$, *and* $\alpha_t \leq \frac{\mu}{MB_G}$ *for all* t. *Then*

$$\mathbf{E}\sum_{t=1}^{\infty} \alpha_t \|\nabla\Gamma(f_t)\|^2 < \infty$$

and therefore

$$\frac{1}{A_T}\mathbf{E}\sum_{t=1}^{T} \alpha_t \|\nabla\Gamma(f_t)\|^2 \stackrel{T\to\infty}{\longrightarrow} 0. \tag{13.53}$$

Proof. By the assumptions, (13.52) holds for each t with α replaced by α_t. The sum of the left side of (13.52) over $1 \leq t \leq T-1$ is $\mathbf{E}\Gamma(f_T) - \mathbf{E}\Gamma(f_t) \geq \Gamma^* - \Gamma(f_1)$. Thus,

$$\mathbf{E}\sum_{t=1}^{\infty} \alpha_t \|\nabla\Gamma(f_t)\|^2 \leq \frac{2(\Gamma(f_1) - \Gamma^*)}{\mu} + \frac{MB}{\mu}\sum_{t=1}^{\infty}\alpha_t^2 < \infty.$$

□

The condition (13.53) shows that the expected squared gradients converge to zero in a certain average sense.

13.6. Differentially private algorithms and generalization

Recall that we have defined a randomized learning algorithm A to be stable if the outputs $A(Z^n)$ and $A(Z'^n)$ of A on two training sets Z^n and Z'^n that differ in only one example are close in terms of their losses: for example, A is ε-uniformly stable if

$$\sup_{z\in\mathsf{Z}} [\mathbf{E}\ell(A(Z^n), z) - \mathbf{E}\ell(A(Z'^n), z)] \leq \varepsilon \tag{13.54}$$

for all Z^n and Z'^n that differ in only one example.

In this section, we will examine a much stronger stability property that pertains to the sensitivity of the conditional distribution of the output of A given $Z^n = z^n$ to individual training examples comprising Z^n. For this purpose, it is convenient to think of $F = A(Z^n)$ as a random object taking values in the hypothesis class $\mathcal{F}$. Then the operation of A is fully described by the conditional distribution $P_{F|Z^n}$. Moreover, we can rewrite the stability condition (13.54) in the following equivalent form:

$$\sup_{z\in\mathsf{Z}} [\mathbf{E}[\ell(F,z)|Z^n = z^n] - \mathbf{E}[\ell(F,z)|Z^n = z'^n]] \leq \varepsilon \tag{13.55}$$

for any two training sets z^n, z'^n that differ in at most one example. Let us now consider a stronger property that compares the conditional *distribution* of F given $Z^n = z^n$ against the one given $Z^n = z'^n$:

Definition 13.4. *A randomized algorithm* A *specified by the conditional distribution* $P_{F|Z^n}$ *is* (ε, δ)-differentially private *if, for any measurable subset* B *of* $\mathcal{F}$ *and for any two trainig sets* z^n, z'^n *that differ in at most one example, we have*

$$P\left[F \in B \middle| Z^n = z^n\right] \leq e^{\varepsilon} P\left[F \in B \middle| Z^n = z'n\right] + \delta. \tag{13.56}$$

Equivalently, $P_{F|Z^n}$ *is* (ε,δ)*-differentially private if, for any function* $g : \mathcal{F} \to [0,1]$,

$$\mathbf{E}[g(F)|Z^n = z^n] \leq e^{\varepsilon}\mathbf{E}[g(F)|Z^n = z'^n] + \delta. \tag{13.57}$$

This definition was proposed by Cynthia Dwork in the context of protecting individual information in statistical databases [**Dwo06**]. Of course, it is useful only for $\delta \in [0,1)$ and for suitably small values of ε.

We start with the following simple observation:

LEMMA 13.2. *If a learning algorithm $P_{F|Z^n}$ is (ε,δ)-differentially private, then it is $(e^\varepsilon - 1 + \delta)$-uniformly stable in the sense of (13.54). If $\varepsilon \in [0,1]$, then the algorithm is $(2\varepsilon+\delta)$-uniformly stable.*

PROOF. A direct consequence of the definition: let z^n, z'^n be two training sets differing in only one example. Then, for any $z \in \mathsf{Z}$,

$$\begin{aligned}\mathbf{E}[\ell(F,z)|Z^n = z^n] - \mathbf{E}[\ell(F,z)|Z^n = z'^n] &\le (e^\varepsilon - 1)\mathbf{E}[\ell(F,z)|Z^n = z'^n] + \delta \\ &\le e^\varepsilon - 1 + \delta.\end{aligned}$$

Since $e^u - 1 \le 2u$ for $u \in [0,1]$, we also obtain the second part of the lemma. □

This stability estimate immediately implies that a differentially private algorithm should generalize. However, the resulting bounds are rather loose. We will now present a tighter bound, due to Nissim and Stemmer [**NS15**].

First, we need to collect some preliminaries on the properties of differentially private algorithms. Fix a randomized algorithm $A = P_{F|Z^n}$ and consider a new algorithm obtained by running M copies of A in parallel on m training sets $(Z_{j,1},\dots,Z_{j,n})$, $1 \le j \le m$. In other words, we form the matrix

$$Z^{m\times n} = \begin{pmatrix} Z_{1,1} & Z_{1,2} & \dots & Z_{1,n} \\ Z_{2,1} & Z_{2,2} & \dots & Z_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{m,1} & Z_{m,2} & \dots & Z_{m,n} \end{pmatrix}$$

and let F_j be the output of A on the jth row of $Z^{m\times n}$. This defines a new algorithm, which we denote by A^m and which is described by the following conditional distribution $P_{F^m|Z^{m\times n}}$: For any m measurable sets $B_1,\dots,B_m \subset \mathcal{F}$,

$$P\left[F_1 \in B_1,\dots,F_m \in B_n \middle| Z^{m\times n} = z^{m\times n}\right] = \prod_{j=1}^m P\left[F \in B_m \middle| Z^n = (z_{j,1},\dots,z_{j,n})\right].$$

If A is (ε,δ)-differentially private, then the algorithm A^m constructed in this way is also (ε,δ)-differentially private. This follows almost immediately from the fact that the jth component of the output of the new algorithm depends only on the jth row of the matrix $Z^{m\times n}$.

Another way of combining algorithms is by *adaptive composition*. Consider two randomized algorithms, $A_1 = P_{F_1|Z^n}$ and $A_2 = P_{F_2|Z^n,F_1}$. Here, the first algorithm takes a dataset Z^n and produces an output $F_1 \in \mathcal{F}_1$; the second algorithm takes a dataset Z^n and an additional $\mathcal{F}_1$-valued input F_1 and produces an output $F_2 \in \mathcal{F}_2$. The *adaptive composition* of A_1 and A_2 takes Z^n as input and produces an output $F_2 \in \mathcal{F}_2$ using a two-stage procedure:

- Generate F_1 by running A_1 on Z^n.
- Generate F_2 by running A_2 on Z^n and on F_1 generated by A_1.

Suppose that A_1 is (ε_1,δ_1)-differentially private, and that, for each $f_1 \in \mathcal{F}_1$, $P_{F_2|Z^n,F_1=f_1}$ is (ε_2,δ_2)-differentially private. Then their adaptive composition is $(\varepsilon_1+\varepsilon_2,\delta_1+\delta_2)$-differentially private. We will now prove this: Fix an arbitrary function $g : \mathcal{F}_2 \to [0,1]$ and two datasets z^n, z'^n that differ in only one sample, and write

$$
\begin{aligned}
\int_{\mathcal{F}_2} g(f_2) P_{F_1,F_2|Z^n=z^n}(\mathrm{d}f_1,\mathrm{d}f_2) &= \int_{\mathcal{F}_1} \left(\int_{\mathcal{F}_2} g(f_2) P_{F_2|Z^n=z^n,F_1=f_1}(\mathrm{d}f_2) \right) P_{F_1|Z^n=z^n}(\mathrm{d}f_1) \\
&\le \int_{\mathcal{F}_1} \min\left(1, e^{\varepsilon_2} \int_{\mathcal{F}_2} g(f_2) P_{F_2|Z^n=z'^n,F_1=f_1}(\mathrm{d}f_2) + \delta_2 \right) P_{F_1|Z^n=z^n}(\mathrm{d}f_1) \\
&= \int_{\mathcal{F}_1} \min\left(1, e^{\varepsilon_1} \int_{\mathcal{F}_2} g(f_2) P_{F_2|Z^n=z'^n,F_1=f_1}(\mathrm{d}f_2) \right) P_{F_1|Z^n=z^n}(\mathrm{d}f_1) + \delta_2, \qquad (13.58)
\end{aligned}
$$

where we have used the differential privacy assumption on A_2. Now, for a fixed realization z'^n, we can define the function

$$
g'(f_1) := \min\left(1, e^{\varepsilon_2} \int_{\mathcal{F}_2} g(f_2) P_{F_2|Z^n=z'^n,F_1=f_1}(\mathrm{d}f_2)\right)
$$

that takes values in $[0,1]$. Therefore, by the differential privacy assumption on A_1,

$$
\begin{aligned}
&\int_{\mathcal{F}_1} \min\left(1, e^{\varepsilon_2} \int_{\mathcal{F}_2} g(f_2) P_{F_2|Z^n=z'^n,F_1=f_1}(\mathrm{d}f_2)\right) P_{F_1|Z^n=z^n}(\mathrm{d}f_1) \\
&\qquad = \int_{\mathcal{F}_1} g'(f_1) P_{F_1|Z^n=z^n}(\mathrm{d}f_1) \\
&\qquad \le e^{\varepsilon_1} \int_{\mathcal{F}_1} g'(f_1) P_{F_1|Z^n=z'^n}(\mathrm{d}f_1) + \delta_1 \\
&\qquad \le e^{\varepsilon_1+\varepsilon_2} \int_{\mathcal{F}_1} \int_{\mathcal{F}_2} g(f_2) P_{F_1,F_2|Z^n=z'^n}(\mathrm{d}f_1,\mathrm{d}f_2) + \delta_1. \qquad (13.59)
\end{aligned}
$$

Using the bound (13.59) in (13.58), we obtain

$$
\mathbf{E}[g(F_2)|Z^n = z^n] \le e^{\varepsilon_1+\varepsilon_2} \mathbf{E}[g(F_2)|Z^n = z'^n] + (\delta_1 + \delta_2).
$$

Since g was arbitrary, we have established the desired differential privacy property.

Finally, we will need a particular differentially private algorithm, the so-called *exponential mechanism* of McSherry and Talwar [**MT07**]. Suppose that we are given a function $U : \mathcal{S} \times \mathsf{Z}^n \to \mathbb{R}$, where $\mathcal{S}$ is a finite set, such that

$$
\max_{s \in \mathcal{S}} |U(s, z^n) - U(s, z'^n)| \le 1
$$

for all z^n, z'^n that differ in only one sample. Consider a randomized algorithm that takes input Z^n and generates an output S taking values in $\mathcal{S}$ according to the following distribution:

$$
P_{S|Z^n=z^n}(s) = \frac{e^{\varepsilon U(s,z^n)/2}}{\sum_{s' \in \mathcal{S}} e^{\varepsilon U(s',z^n)/2}}. \qquad (13.60)
$$

We have the following:

Lemma 13.3. *The exponential algorithm (13.60) has the following properties:*

(1) It is ε-differentially private.

(2) Let $U^*(z^n) := \max_{s \in \mathcal{S}} U(s, z^n)$. Then, for any t,

$$
P\left[U(S, Z^n) < U^*(Z^n) - t \middle| Z^n = z^n\right] \le |\mathcal{S}| e^{-\varepsilon t/4}. \qquad (13.61)
$$

PROOF. For part 1, fix z^n, z'^n differing in only one sample. Then we have

$$\begin{aligned}
\frac{P[S=s|Z^n=z^n]}{P[S=s|Z^n=z'^n]} &= \frac{e^{\varepsilon U(s,z^n)/2}}{e^{\varepsilon U(s,z'^n)/2}} \cdot \frac{\sum_{s'\in\mathcal{S}} e^{\varepsilon U(s',z'^n)/2}}{\sum_{s'\in\mathcal{S}} e^{\varepsilon U(s',z^n)/2}} \\
&= \exp\left(\frac{\varepsilon(U(s,z^n)-U(s,z'^n))}{2}\right) \cdot \frac{\sum_{s'\in\mathcal{S}} e^{\varepsilon U(s',z'^n)/2}}{\sum_{s'\in\mathcal{S}} e^{\varepsilon U(s',z^n)/2}} \\
&\le e^{\varepsilon/2} \cdot \frac{|\mathcal{S}| e^{(\varepsilon/2)\max_{s\in\mathcal{S}} U(s,z^n)}}{|\mathcal{S}| e^{(\varepsilon/2)\min_{c\in\mathcal{S}} U(s,z'^n)}} \\
&\le e^{\varepsilon/2} \cdot \exp\left(\frac{\varepsilon}{2} \cdot \max_{s\in\mathcal{S}} |U(s,z^n)-U(s,z'^n)|\right) \\
&\le e^{\varepsilon}.
\end{aligned}$$

For part 2: for each t, define the set

$$\mathcal{S}_t := \{s \in \mathcal{S} : U(s,z^n) \ge U^*(z^n) - t\}.$$

Then

$$\begin{aligned}
P\left[S \in \mathcal{S}_t \middle| Z^n = z^n\right] &= \frac{\sum_{s\in\mathcal{S}_t} e^{\varepsilon U(s,z^n)/2}}{\sum_{s\in\mathcal{S}} e^{\varepsilon U(s,z^n)/2}} \\
&= \frac{\sum_{s\in\mathcal{S}_t} e^{\varepsilon U(s,z^n)/2}}{\sum_{s\in\mathcal{S}_{t/2}} e^{\varepsilon U(s,z^n)/2} + \sum_{s\in\mathcal{S}^c_{t/2}} e^{\varepsilon U(s,z^n)/2}} \\
&\le \frac{\sum_{s\in\mathcal{S}_t} e^{\varepsilon U(s,z^n)/2}}{\sum_{s\in\mathcal{S}^c_{t/2}} e^{\varepsilon U(s,z^n)/2}} \\
&\le e^{(\varepsilon/2)(U^*(z^n)-t)} e^{-(\varepsilon/2)(U^*(z^n)-t/2)} |\mathcal{S}_t| \\
&\le |\mathcal{S}| e^{-\varepsilon t/4}.
\end{aligned}$$

□

Finally, we need the following result, due to Nissim and Stemmer [**NS15**]:

LEMMA 13.4. *Let the parameters ε, δ be such that $0 < \delta \le \varepsilon \le \frac{1}{5}$ and $m = \frac{\varepsilon}{\delta}$ is an integer. Consider an algorithm B that takes an input $Z^{m\times n}$ and outputs a pair $(F_J, J) \in \mathcal{F} \times \{1,\dots,m\}$. If B is (ε,δ)-differentially private, then*

$$\mathbf{P}\left[L_n^{(J)}(F_J) \le L(F_J) + 5\varepsilon\right] \ge \varepsilon. \tag{13.62}$$

Here, for each $j \in \{1,\dots,m\}$, $L_n^{(j)}(f)$ denotes the empirical loss of $f \in \mathcal{F}$ on the jth row of the matrix $Z^{m\times n}$.

PROOF. We first derive a version of this result that holds in expectation. Let $Z'^{m\times n}$ be an independent copy of $Z^{m\times n}$, and let $Z^{m\times n}_{(ji)}$ be obtained from $Z^{m\times n}$ by replacing the sample Z_{ji} in the jth row and the ith column with Z'_{ji}.

Now, we write

$$
\begin{aligned}
\mathbf{E}\left[L_n^{(J)}(F_J)\right] &= \sum_{j=1}^{m} \mathbf{E}\left[L_n^{(J)}(F_J)\mathbf{1}\{J=j\}\right] \\
&= \frac{1}{n}\sum_{j=1}\sum_{i=1}^{n} \mathbf{E}[\ell(F_J, Z_{ji})\mathbf{1}\{J=j\}] \\
&= \frac{1}{n}\sum_{j=1}\sum_{i=1}^{n} \int P_{Z^{m\times n}}(\mathrm{d}z^{m\times n}) \int P_{Z'^{m\times n}}(\mathrm{d}z'^{m\times n}) \int P_{(F_J,J)|Z^{m\times n}=z^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z_{ji}) \\
&= \frac{1}{n}\sum_{j=1}\sum_{i=1}^{n} \int P_{Z^{m\times n}}(\mathrm{d}z^{m\times n}) \int P_{Z'^{m\times n}}(\mathrm{d}z'^{m\times n}) \int P_{(F_J,J)|Z^{m\times n}=z_{ji}^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z'_{ji}),
\end{aligned} \tag{13.63}
$$

where in the last line we have used the assumption that the entries of $Z^{m\times n}$ and $Z'^{m\times n}$ are i.i.d. draws from the same distribution. Since $P_{(F_J,J)|Z^{m\times n}}$ is (ε,δ)-differentially private, we have

$$
\int P_{(F_J,J)|Z^{m\times n}=z_{ji}^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z'_{ji}) \le e^{\varepsilon} \int P_{(F_J,J)|Z^{m\times n}=z^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z'_{ji}) + \delta.
$$

Averaging with respect to $Z^{m\times n}$ and $Z'^{m\times n}$ and exploiting independence, we obtain

$$
\begin{aligned}
&\int P_{Z^{m\times n}}(\mathrm{d}z^{m\times n}) \int P_{Z'^{m\times n}}(\mathrm{d}z'^{m\times n}) \int P_{(F_J,J)|Z^{m\times n}=z_{ji}^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z'_{ji}) \\
&\le e^{\varepsilon} \int P_{Z^{m\times n}}(\mathrm{d}z^{m\times n}) \int P_{Z'^{m\times n}}(\mathrm{d}z'^{m\times n}) \int P_{(F_J,J)|Z^{m\times n}=z_{ji}^{m\times n}}(\mathrm{d}f_j, j)\ell(f_j, z'_{ji}) + \delta \\
&= e^{\varepsilon}\mathbf{E}[\ell(F_J, Z'_{ji})\mathbf{1}\{J=j\}] + \delta \\
&= e^{\varepsilon}\mathbf{E}[L(F_J)\mathbf{1}\{J=j\}] + \delta.
\end{aligned}
$$

Substituting this into (13.63), we obtain

$$
\mathbf{E}\left[L_n^{(J)}(F_J)\right] \le e^{\varepsilon}\mathbf{E}[L(F_J)] + m\delta \le \mathbf{E}[L(F_J)] + 2\varepsilon + m\delta \le \mathbf{E}[L(F_J)] + 3\varepsilon, \tag{13.64}
$$

where the second step follows from the inequality $ae^x \le 2x + a$ for $a, x \in [0,1]$. The probability bound (13.62) follows from Lemma 13.6 in the Appendix. □

Now we are ready to state and prove the main result of this section:

THEOREM 13.11 (Nissim–Stemmer). *Let the parameters ε, δ be such that $0 < \delta \le \varepsilon \le \frac{1}{10}$ and $m = \frac{\varepsilon}{\delta}$ is an integer. Let $A = P_{F|Z^n}$ be an (ε,δ)-differentially private randomized learning algorithm operaitng on a sample of size $n \ge \frac{4}{\varepsilon^2}\log\frac{8}{\delta}$. Assume that $\varepsilon \ge \delta$. Then, for any loss function $\ell : \mathcal{F} \times \mathsf{Z} \to [0,1]$,*

$$
\mathbf{P}\left[|L_n(F) - L(F)| > 13\varepsilon\right] \le \frac{2\delta}{\varepsilon}\log\frac{2}{\varepsilon}. \tag{13.65}
$$

13.6.1. The proof of Theorem 13.11. Suppose, to the contrary, that A does not generalize, i.e., that

$$\mathbf{P}\left[L_n(F) - L(F) > 13\varepsilon\right] > \frac{\delta}{\varepsilon}\log\frac{2}{\varepsilon}. \tag{13.66}$$

Draw $m+1$ independent datasets $(Z_{j,1},\dots,Z_{j,n})$, $j \in \{1,\dots,m+1\}$, from P_Z, and form the $(m+1)\times n$ matrix

$$Z^{(m+1)\times n} == \begin{pmatrix} Z_{1,1} & Z_{1,2} & \dots & Z_{1,n} \\ Z_{2,1} & Z_{2,2} & \dots & Z_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{m,1} & Z_{m,2} & \dots & Z_{m,n} \\ Z_{m+1,1} & Z_{m+1,2} & \dots & Z_{m+1,n} \end{pmatrix}.$$

Think of the first m rows of this matrix as m independent training sets, and of the last row as a separate validation set. For $j \in \{1,\dots,m\}$, let $F_j \in \mathcal{F}$ be the output of an independent copy of A on the jth row of this matrix. Next, for each $f \in \mathcal{F}$, define the empirical losses

$$L^{(j)}_n(f) := \frac{1}{n}\sum_{i=1}^n \ell(f, Z_{j,i}), \qquad j \in \{1,\dots,m+1\}$$

of f on the jth training set and on the validation set. For the random set $\mathcal{S} = \{(F_j, j)\}_{j=1}^m$, define a function $U : \mathcal{S}\times \mathsf{Z}^{(m+1)\times n} \to \mathbb{R}$ by

$$U((F_j, j), z^{(m+1)\times n}) := n\left(L^{(j)}_n(F_j) - L^{(m+1)}_n(F_j)\right) \tag{13.67}$$

and generate a random pair $(F_I, I) \in \mathcal{S}$ by running the McSherry–Talwar exponential algorithm (13.60) with this function U on $Z^{(m+1)\times n}$.

For $j \in \{1,\dots,m\}$, denote by E_j the event $\{L^{(j)}_n(F_j) - L(F_j) > 13\varepsilon\}$, and let $E = \bigcup_{j=1}^m E_j$. By hypothesis, cf. Eq. (13.66), $\mathbf{P}[E_j] > \frac{\delta}{\varepsilon}\log\frac{2}{\varepsilon}$ for each j. Since the events $E_1,\dots,E_m$ are independent,

$$\mathbf{P}[E^c] = \mathbf{P}\left[\bigcap_{j=1}^m E^c_j\right] = \prod_{j=1}^m \mathbf{P}[E^c_j] \le \left(1 - \frac{\delta}{\varepsilon}\log\frac{2}{\varepsilon}\right)^m = \left(1 - \frac{\delta}{\varepsilon}\log\frac{2}{\varepsilon}\right)^{\varepsilon/\delta} \le \frac{\varepsilon}{2}.$$

Next, for each $j \in \{1,\dots,m\}$, let G_j denote the event that $|L^{(m+1)}_n(F_j) - L(F_j)| \le \varepsilon$, and let $G = \bigcap_{j=1}^m G_j$. On the other hand, since $F_1,\dots,F_m$ are independent of the last row of the matrix $Z^{(m+1)\times n}$, Hoeffding's lemma and the union bound guarantee that

$$\mathbf{P}\left[G\right] \ge 1 - 2me^{-2n\varepsilon^2} = 1 - \frac{2\varepsilon}{\delta}e^{-2n\varepsilon^2}.$$

By the union bound,

$$\begin{aligned}\mathbf{P}[E\cap G] &= 1 - \mathbf{P}[E^c \cup G^c] \ge 1 - (\mathbf{P}[E^c] + \mathbf{P}[G^c]) \\ &\ge 1 - \frac{2\varepsilon}{\delta}e^{-2n\varepsilon^2} - \frac{\varepsilon}{2}.\end{aligned}$$

Consequently, if we choose $n \ge \frac{1}{2\varepsilon^2}\log\frac{8}{\delta}$, then we will have $\mathbf{P}[E\cap G] \ge 1 - \frac{3\varepsilon}{4}$.

Now, on the event $E \cap G$, the function U defined in (13.67) will satisfy

$$\begin{aligned}
U^*(Z^{(m+1)\times n}) &= \max_{j\in\{1,\ldots,m\}} U(j, Z^{(m+1)\times n}) \\
&= n \max_{j\in\{1,\ldots,m\}} \left[L_n^{(j)}(F_j) - L_n^{(m+1)}(F_j)\right] \\
&= n \max_{j\in\{1,\ldots,m\}} \left[\left(L_n^{(j)}(F_j) - L(F_j)\right) + \left(L(F_j) - L_n^{(m+1)}(F_j)\right)\right] \\
&\geq 12n\varepsilon.
\end{aligned}$$

Therefore, on the event $E \cap G$, with probability at least $1 - me^{-n\varepsilon^2/4}$, the output (F_I, I) the exponential mechanism with (13.67) will be such that

$$\begin{aligned}
L_n^{(I)}(F_I) - L(F_I) &= L_n^{(I)}(F_I) - L_n^{(m+1)}(F_I) + L_n^{(m+1)}(F_I) - L(F_I) \\
&= \frac{U(I, Z^{(m+1)\times n})}{n} + L_n^{(m+1)}(F_I) - L(F_I) \\
&\geq \frac{U^*(Z^{(m+1)\times n})}{n} - 2\varepsilon > 10\varepsilon.
\end{aligned}$$

Thus, if n is also chosen to be larger then $\frac{4}{\varepsilon^2}\log\frac{8}{\delta}$, then the output (F_I, I) will satisfy

$$\mathbf{P}\left[L_n^{(I)}(F_I) \leq L(F_I) + 10\varepsilon\right] \leq \varepsilon. \tag{13.68}$$

By Lemma 13.4, this is impossible if we can show that the algorithm $B = P_{(F_J,J)|Z^{(m+1)\times n}}$ is $(2\varepsilon, \delta)$-differentially private.

To see this, we observe that the algorithm $B = P_{(F_J,J)|Z^{(m+1)\times n}}$ is an adaptive composition of A^m and the McSherry–Talwar algorithm. Since A is (ε, δ)-differentially private, so is A^m, and the McSherry–Talwar algorithm is $(\varepsilon, 0)$-differentially private. Therefore, B is $(2\varepsilon, \delta)$-differentially private. Therefore, by Lemma 13.4, its output must satisfy

$$\mathbf{P}\left[L_n^{(J)}(F_J) \leq L(F_J) + 10\varepsilon\right] \geq 2\varepsilon.$$

which contradicts (13.68).

13.7. Technical lemmas

LEMMA 13.5. *Let U and V be two random variables, such that $U \leq V$ almost surely. Then*

$$\mathbf{E}|U| \leq |\mathbf{E}U| + 2\mathbf{E}|V|. \tag{13.69}$$

PROOF. We have

$$\mathbf{E}|U| = \mathbf{E}|(V-U) - V| \leq \mathbf{E}|V-U| + \mathbf{E}|V| = \mathbf{E}[V-U] + \mathbf{E}|V| \leq |\mathbf{E}U| + 2\mathbf{E}|V|.$$

□

LEMMA 13.6. *Let U and V be two random variables, such that $0 \leq U, V \leq 1$ almost surely, and*

$$\mathbf{E}[U] \leq \mathbf{E}[V] + 3\varepsilon$$

for some $0 \leq \varepsilon \leq \frac{1}{5}$. Then

$$\mathbf{P}[U \leq V + 5\varepsilon] \geq \varepsilon. \tag{13.70}$$

Proof. Suppose, by way of contradiction, that $\mathbf{P}[U \le V + 5\varepsilon] < \varepsilon$. Then

$$\begin{aligned}\mathbf{E}[U] &= \mathbf{E}[U\mathbf{1}\{U - V \le 5\varepsilon\}] + \mathbf{E}[U\mathbf{1}\{U - V > 5\varepsilon\}] \\ &> \mathbf{E}[(5\varepsilon + V)\mathbf{1}\{U - V > 5\varepsilon\}] \\ &= (5\varepsilon)\,\mathbf{P}[U - V > 5\varepsilon] + \mathbf{E}[V\mathbf{1}\{U - V > 5\varepsilon\}].\end{aligned}$$

On the other hand, since $0 \le V \le 1$

$$\mathbf{E}[V\mathbf{1}\{U - V \le 5\varepsilon\}] \le \mathbf{E}[\mathbf{1}\{U - V \le 5\varepsilon\}] = \mathbf{P}[U - V \le 5\varepsilon] < \varepsilon.$$

Therefore,

$$\begin{aligned}\mathbf{E}[U] &> (5\varepsilon)\,\mathbf{P}[U - V > 5\varepsilon] + \mathbf{E}[V] - \varepsilon \\ &> (5\varepsilon)(1 - \varepsilon) + \mathbf{E}[V] - \varepsilon \\ &= \mathbf{E}[V] + 4\varepsilon - 5\varepsilon^2 \\ &\ge \mathbf{E}[V] + 3\varepsilon,\end{aligned}$$

which contradicts the assumption that $\mathbf{E}[U] \le \mathbf{E}[V] + 3\varepsilon$. □

CHAPTER 14

Online optimization algorithms

The main topic for this chapter is online algorithms for convex optimization. In the models for statistical learning problems discussed earlier, it is assumed the data Z^n are generated by independent draws from a probability distribution P on Z. The probability distribution P is unknown, and for a problem to be PAC learnable, there should be an algorithm that is probably almost correct, where the probability of almost correctness, $1-\delta$, should converge to one uniformly over all P in some class $\mathcal{P}$. Thus, the definition of PAC has some min-max aspect.

We can buy into the minimax modeling philosophy more fully by dropping the assumption that the samples are drawn at random from some distribution P. Rather, we could consider the samples $z_1, z_2, \ldots$ to be arbitrary. The learner can be viewed as a *player*, and the variables $z_1, z_2, \ldots$ can be viewed as chosen by an *adversary*. Usually in this context we won't be modeling the adversary, but just assume the adversary could come up with an arbitrary sequence $z_1, z_2, \ldots$. While the problem formulation is somewhat different from the statistical learning framework we have been focusing on all this time, much of the same tools we have seen can be applied to the learning problems.

The performance analysis of stochastic gradient descent we looked at in Chapter 13 is about the performance of the stochastic gradient descent algorithm, such that the data points $z_1, \ldots, z_n$ were fixed. All the stochastic effects were due to internal randomization in the algorithm, and the guarantees were uniform in $z_1, \ldots, z_n$.

We shall focus on the online, or adversarial, function minimization problem.[1] An influential paper in this line of work is the one of [**Zin03**], although there is a long line of literature on learning in an adversarial context.

14.1. Online convex programming and a regret bound

The paper of Zinkevich [**Zin03**] sparked much interest in the adversarial framework for modeling online function minimization. The paper shows that a projected gradient descent algorithm achieves zero asymptotic average regret rate for minimizing an arbitrary sequence of uniformly Lipschitz convex functions over a closed bounded convex set in $\mathbb{R}^d$. The framework involves objects familiar to us, although the terminology is a bit closer to game theory.

[1]There is work on a related *online learning* problem for which the player can select a label $\widehat{y}_t$ at each time after observing x_t. Then the actual label y_t is revealed and the player incurs the loss $\ell((x_t, y_t), \widehat{y}_t)$. A simple case is the 0-1 loss. An algorithm A for this problem produces a label $\widehat{y}_t = A(x_1, \ldots, x_t, y_1, \ldots, y_{t-1})$ for each t. See [**SSBD14**] for an introduction.

- Let $\mathcal{F}$ be a nonempty, closed, convex subset of a Hilbert space $\mathcal{H}$. It is assumed $\mathcal{F}$ is bounded, so $D := \max\{\|f - f'\| : f, f' \in \mathcal{F}\|\} < \infty$. The player selects actions from $\mathcal{F}$.
- Let Z be a set, denoting the possible actions of the adversary.
- Let $\ell : \mathcal{F} \times \mathsf{Z} \to \mathbb{R}_+$. The interpretation is that $\ell(f_t, z_t)$ is the loss to the player for step t. We sometimes use the notation $\ell_t : \mathsf{Z} \to \mathbb{R}_+$, defined by $\ell_t(f) = \ell(f, z_t)$.
- Suppose the player has access to an algorithm that can compute $\ell_t(f)$ and $\nabla \ell_t(f)$ for a given f.
- Suppose the player has access to an algorithm that can calculate $\Pi(f)$ for any $f \in \mathbb{R}^d$, where $\Pi : \mathcal{H} \to \mathcal{F}$ is the projection mapping: $\Pi(f) = \arg\min\{\|f - f'\|^2 : f' \in \mathcal{F}\}$, that maps any $f \in \mathcal{H}$ to a nearest point in $\mathcal{F}$.
- $T \geq 1$ represents a *time horizon* of interest.

The online convex optimization game proceeds as follows.

- At each time step t from 1 to T, the player chooses $f_t \in \mathcal{F}$
- The adversary chooses $z_t \in \mathsf{Z}$
- The player observes z_t and incurs the loss $\ell(f_t, z_t)$.

Roughly speaking, the player would like to select the sequence of actions (f_t) to minimize the total loss for some time-horizon T, or equivalently, minimize the corresponding average loss per time step:

$$J_T((f_t)) := \sum_{t=1}^{T} \ell(f_t, z_t) \qquad L_T((f_t)) := \frac{1}{T} J_T((f_t)).$$

If we wanted to emphasize the dependence on z^T we could have writen $J_T((f_t), z^T)$ and $L_T((f_t), z^T)$ instead. A possible strategy of the player is to use a fixed $f^* \in \mathcal{F}$ for all time, in which case we write the total loss as $J_T(f^*) := \sum_{t=1}^{T} \ell(f^*, z_t)$ and the loss per time step as $L_T(f^*) = \frac{1}{T} J_T(f^*)$. Note that $L_T(f^*)$ is the empirical loss for f^* for T samples. If the player is extremely lucky, or if for each t a genie knowing z_t in advance reveals an optimal choice to the player, the player could use $f_t^{\text{genie}} := \arg\min_{f \in \mathsf{F}} \ell(f, z_t)$. Typically it is unreasonable to expect a player, without knowing z_t before selecting f_t, to achieve, or even nearly achieve, the genie-assisted minimum loss.

It turns out that a realistic goal is for the player to make selections that perform nearly as well as *any fixed strategy* f^* that could possibly be selected after the sequence z^T is revealed. Specifically, if the player uses (f_t) then the *regret* (for not using an optimal fixed strategy) is defined by:

$$R_T((f_t)) = J_T((f_t)) - \inf_{f^* \in \mathcal{F}} J_T(f^*),$$

where for a particular f^*, $J_T((f_t)) - J_T(f^*)$ is the regret for using (f_t) instead of f^*. We shall be interested in strategies the player can use to (approximately) minimize the regret. Even this goal seems ambitious, but one important thing the player can exploit is that the player can let f_t depend on t, whereas the performance the player aspires to match is that of the best policy that is constant over all steps t.

Zinkevich [**Zin03**] showed that the projected gradient descent algorithm, defined by

$$f_{t+1} = \Pi(f_t - \alpha_t \nabla \ell_t(f_t)), \tag{14.1}$$

meets some performance guarantees for the regret minimization problem. Specifically, under convexity and the assumption that the functions ℓ_t are all L-Lipschtiz continuous, Zinkevich showed that regret $O(LD\sqrt{T})$ is achievable by gradient descent. Under such assumptions the $\sqrt{T}$ scaling is, in fact, the best possible. The paper of Hazan, Agarwal, and Kale [**HAK07**] shows that if, in addition, the functions ℓ_t are all m-strongly convex for some $m > 0$, then gradient descent can achieve $O\left(\frac{L^2}{m}\log T\right)$ regret. The paper [**HAK07**] ties together several different previous approaches including follow-the-leader, exponential weighting, Cover's algorithm, and gradient descent. The following theorem combines the analysis of [**Zin03**] for the case of Lipschitz continuous objective functions and the analysis of [**HAK07**] for strongly convex functions. The algorithms used for the two cases differ only in the stepsize selections. Recall that D is the diameter of $\mathcal{F}$.

THEOREM 14.1. *Suppose $\ell(\cdot, z)$ is convex and L-Lipschtiz continuous for each z, and suppose the projected gradient algorithm (14.1) is run with stepsizes $(\alpha_t)_{t\geq 1}$.*
(a) If $\alpha_t = \frac{c}{\sqrt{t}}$ for $t \geq 1$, then the regret is bounded as follows:

$$R_T((f_t)) \leq \frac{D^2\sqrt{T}}{2c} + \left(\sqrt{T} - \frac{1}{2}\right) L^2 c,$$

which for $c = \frac{D}{L\sqrt{2}}$ gives:

$$R_T((f_t)) \leq DL\sqrt{2T}.$$

(b) If, in addition, $\ell(\cdot, z)$ is m-strongly convex for some $m > 0$ and $\alpha_t = \frac{1}{mt}$ for $t \geq 1$, then the regret is bounded as follows:

$$R_T((f_t)) \leq \frac{L^2(1 + \log T)}{2m}.$$

PROOF. Most of the proof is the same for parts (a) and (b), where for part (a) we simply take $m = 0$. Let $f_{t+1}^\flat = f_t - \alpha_t \nabla \ell_t(f_t)$, so that $f_{t+1} = \Pi(f_{t+1}^\flat)$. Let $f^* \in \mathcal{F}$ be any fixed policy. Note that

$$f_{t+1}^\flat - f^* = f_t - f^* - \alpha_t \nabla \ell_t(f_t)$$
$$\|f_{t+1}^\flat - f^*\|^2 = \|f_t - f^*\|^2 - 2\alpha_t\langle f^t - f^*, \nabla \ell_t(f_t)\rangle + \alpha_t^2 \|\nabla \ell_t(f_t)\|^2.$$

By the fact $f^* = \Pi(f^*)$ and the contraction property of Π (see Proposition 4.2), $\|f_{t+1} - f^*\| \leq \|f_{t+1}^\flat - f^*\|$. Also, by the Lipschitz assumption, $\|\nabla \ell_t(f_t)\| \leq L$. Therefore,

$$\|f_{t+1} - f^*\|^2 \leq \|f_t - f^*\|^2 - 2\alpha_t\langle f_t - f^*, \nabla \ell_t(f_t)\rangle + \alpha_t^2 L^2$$

or, equivalently,

$$2\big\langle f_t - f^*, \nabla \ell_t(f_t)\big\rangle \leq \frac{\|f_t - f^*\|^2 - \|f_{t+1} - f^*\|^2}{\alpha_t} + \alpha_t L^2. \tag{14.2}$$

(Equation (14.2) captures well the fact that this proof is based on the use of $\|f_t - f^*\|$ as a potential function. The only property of the gradient vectors $\nabla \ell_t(f_t)$ used so far is $\|\nabla \ell_t(f_t)\| \leq L$. The specific choice of using gradient vectors is exploited next, to bound differences in the loss function.) The strong convexity of ℓ_t implies $\ell_t(f^*) - \ell_t(f_t) \geq \langle f^* -$

$f_t, \nabla \ell_t(f_t)\rangle + \frac{m}{2}\|f^* - f_t\|^2$, or equivalently,
$2(\ell_t(f_t) - \ell_t(f^*)) \leq 2\langle f_t - f^*, \nabla \ell_t(f_t)\rangle - m\|f_t - f^*\|^2$, which combined with (14.2) gives:

$$2(\ell_t(f_t) - \ell_t(f^*)) \leq \frac{\|f_t - f^*\|^2 - \|f_{t+1} - f^*\|^2}{\alpha_t} + \alpha_t L^2 - m\|f_t - f^*\|^2 \tag{14.3}$$

We shall use the following for $1 \leq t \leq T-1$:

$$\frac{\|f_t - f^*\|^2 - \|f_{t+1} - f^*\|^2}{\alpha_t} = \frac{\|f_t - f^*\|^2}{\alpha_t} - \frac{\|f_{t+1} - f^*\|^2}{\alpha_{t+1}} + \left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right)\|f_{t+1} - f^*\|^2.$$

Summing each side of (14.3) from $t = 1$ to T yields:

$$\begin{aligned}
2(J_T((f_t)) - J_T(f^*)) &\leq \left(\frac{1}{\alpha_1} - m\right)\|f_1 - f^*\|^2 - \frac{1}{\alpha_T}\|f_{T+1} - f^*\|^2 \\
&\quad + \sum_{t=1}^{T-1}\left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t} - m\right)\|f_{t+1} - f^*\|^2 + L^2\sum_{t=1}^{T}\alpha_t \\
&\leq D^2\left(\frac{1}{\alpha_1} - m + \sum_{t=1}^{T-1}\left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t} - m\right)\right) + L^2\sum_{t=1}^{T}\alpha_t \\
&\leq D^2\left(\frac{1}{\alpha_T} - mT\right) + L^2\sum_{t=1}^{T}\alpha_t \qquad (14.4)
\end{aligned}$$

(Part (a)) If $m = 0$ the bound (14.4) becomes

$$2(J_T((f_t)) - J_T(f^*)) \leq \frac{D^2}{\alpha_T} + L^2\sum_{t=1}^{T}\alpha_t \tag{14.5}$$

Now if $\alpha_t = \frac{c}{\sqrt{t}}$, then

$$\sum_{t=1}^{T}\alpha_t = c + \sum_{t=2}^{T}\frac{c}{\sqrt{t}} \leq c + c\int_{t=1}^{T}\frac{c\mathrm{d}t}{\sqrt{t}} = (2\sqrt{T} - 1)c$$

and we get

$$J_T((f_t)) - J_T(f^*) \leq \frac{D^2\sqrt{T}}{2c} + \left(\sqrt{T} - \frac{1}{2}\right)L^2 c$$

If $c = \frac{D}{L\sqrt{2}}$ then $J_T((f_t)) - J_T(f^*) \leq DL\sqrt{2T}$. Since $f^* \in \mathfrak{F}$ is arbitrary it follows that $R_T((f_t)) \leq DL\sqrt{2T}$.

(Part (b)) For the case $m > 0$ and $\alpha_t = \frac{1}{mt}$ for all $t \geq 1$, the first term on the right-hand side of (14.4) is zero, and

$$\sum_{t=1}^{T}\alpha_t = \frac{1}{m}\left(1 + \sum_{t=2}^{T-1}\frac{1}{t}\right) \leq \frac{1 + \log T}{m},$$

so part (b) of the theorem follows from (14.4) and the fact $f^* \in \mathfrak{F}$ is arbitrary. $\square$

14.2. Online perceptron algorithm

The perceptron algorithm of Rossenblatt (1958) is an iterative algorithm for training binary classifiers of the form $y = \mathrm{sgn}(\langle f, x\rangle)$. In this section we show that the perceptron algorithm can be viewed as an instance of the gradient descent algorithm for a certain online convex optimization problem, and adapt the proof of Theorem 14.1 to bound the total number of iterations in the realizable case. Let the original data set be denoted by $\widetilde{z}^n$. At each time step t, the gradient descent algorithm will be applied to a sample $z_t = \widetilde{z}_{I_t}$ for a choice of index $I_t \in [n]$ to be described in what follows. We use z and (x, y) interchangeably, and, similarly, z_t and (x_t, y_t) interchangeably. Consider the surrogate loss function $\widetilde{\ell}(f, x) = (1 - y\langle f, x\rangle)_+$, which is the penalized version of 0-1 loss arising from use of the hinge penalty function $\varphi(u) = (1 + u)_+$. At step t the learner selects f_t, and then the adversary selects a convex loss function. Usually we would think to use $\widetilde{\ell}_t(\cdot) = \widetilde{\ell}(\cdot, z_t)$. However, the rules of online convex function minimization allow the adversary to present a different sequence of convex functions $\ell_t : t \geq 1$, determined as follows:

- If $y_t\langle f, x_t\rangle \geq 0$ (i.e. if f_t correctly classifies z_t), then $\ell_t \equiv 0$.
- If $y_t\langle f, x_t\rangle < 0$ then $\ell_t \equiv \widetilde{\ell}_t$.

Note that ℓ_t is convex for each t and

$$\nabla \ell_t(f_t) = \begin{cases} 0 & \text{if } y_t = \mathrm{sgn}(\langle f_t, x_t\rangle) \\ -y_t x_t & \text{else.} \end{cases}$$

The gradient descent algorithm, with no projection and constant stepsize α, becomes::

$$f_{t+1} = \begin{cases} f_t & \text{if } y_t = \mathrm{sgn}(\langle f_t, x_t\rangle) \\ f_t + \alpha y_t x_t & \text{else,} \end{cases} \tag{14.6}$$

where we use the initial state $f_1 = 0$. Since f_t is proportional to α for all t, the classifiers f_t are all proportional to α, so the 0-1 loss performance of the algorithm does not depend on α. We included the stepsize $\alpha > 0$ only for the proof. We now specify how the index I_t is chosen. If there is some sample that f_t does not correctly label, then I_t is the index of such a sample. Otherwise, $I_t \in [n]$ is arbitrary. The classical perceptron algorithm corresponds to this choice of I_t, the update rule (14.6), and stopping at the first time t it is found that f_t separates the original data.

Proposition 14.1. *(Perceptron classification, realizable case) Let*

$$L \geq \max_{1 \leq i \leq n} \|x_i\|$$

and

$$B \geq \min\{\|f^*\| : y_i\langle f^*, x_i\rangle \geq 1 \text{ for } i \in [n]\}.$$

At most B^2L^2 updates are needed for the perceptron algorithm to find a separating classifier.

Proof. (Variation of the proof of Theorem 14.1.) By the assumptions, ℓ_t is L-Lipschitz continuous for all t. Let f^* be a vector so that $\|f^*\| \leq B$ and $y_i\langle f^*, x_i\rangle \geq 1$ for $i \in [n]$. By (14.3) with $m = 0$ and $\alpha_t = \alpha$ for all t,

$$2(\ell_t(f_t) - \ell_t(f^*)) \leq \frac{\|f_t - f^*\|^2 - \|f_{t+1} - f^*\|^2}{\alpha} + \alpha L^2$$

Summing over t from 1 to T yields

$$\begin{aligned}2(J_T((f_t)) - J_T(f^*)) &\leq \frac{1}{\alpha}\|f_1 - f^*\|^2 - \frac{1}{\alpha}\|f_{T+1} - f^*\|^2 + \alpha L^2 T \\ &\leq \frac{B^2}{\alpha} + \alpha L^2 T\end{aligned}$$

Now $\ell_t(f_t) \geq 1$ if f_t does not separate the data, whereas $\ell_t(f^*) = 0$ for all t. Thus, if none of $f_1, \ldots f_T$ separate the data, then $2T \leq 2(J_T((f_t)) - J_T(f^*)) \leq \frac{B^2}{\alpha} + \alpha L^2 T$ for any $\alpha > 0$. Taking $\alpha = \frac{B}{L\sqrt{T}}$ yields $T \leq BL\sqrt{T}$, or $T \leq (BL)^2$. Thus, at most $(BL)^2$ updates are needed for the algorithm to find a separating classifier. □

14.3. On the generalization ability of online learning algorithms

This section is based on [**CBCG04**]. An instance of the on-line convex function minimization framework of Section 14.1 is represented by a tuple $(\mathcal{F}, \mathsf{Z}, \ell)$. An online algorithm A prescribes an action f_1 and then, for each $t \geq 1$, an action $f_{t+1} = A(f_1, \ldots, f_t, z_1, \ldots, z_t)$. Section 14.1 considered regret for arbitrary sequences of samples. Consider instead, the statistical learning framework, such that the samples $Z_1, Z_2, \ldots, Z_n$ are independent and identically distributed random variables with values in Z and some probability distribution P. Consider using an algorithm A for online convex function minimization with $T = n$, so the algorithm makes one pass through the data using the samples $Z_1, \ldots, Z_n$ in order.

The sequence $f_1, f_2, \ldots, f_n$ produced by A is random, due to the randomness of the Z's. However, for each $t \in [n]$, f_t is determined by A and $Z_1, \ldots, Z_{t-1}$, so f_t is independent of Z_t. Therefore, given f_t, $\ell(f_t, Z_t)$ is an unbiased estimator of the generalization performance of f_t. In essence, each subsequent sample Z_t is a test sample for f_t. It is therefore to be expected that online algorithms in the statistical framework have good generalization ability. An application of the Azuma-Hoeffding inequality (Theorem 2.1) makes this precise:

Theorem 14.2. *(Generalization ability of on-line algorithms) Suppose ℓ is bounded, with values in $[0, 1]$. Suppose $Z_1, Z_2, \ldots$ are independent and identically distributed, and let (f_t) be produced by an online learning algorithm A. Then for any $T \geq 1$, with probability at least $1 - \delta$,*

$$\frac{1}{T}\sum_{i=1}^{T} L(f_t) \leq \frac{1}{T}\sum_{t=1}^{T} \ell(f_t, Z_t) + \sqrt{\frac{2\log\frac{1}{\delta}}{T}}, \tag{14.7}$$

where $L(f)$ denotes the generalization loss of a hypothesis f for the probability distribution P used to generate the samples, and $\frac{1}{T}\sum_{t=1}^{T} \ell(f_t, Z_t) = L_T((f_t))$ is the average loss per sample suffered by the online learner. Furthermore, if $\ell(\cdot, z)$ is convex for each z fixed, and $\overline{f}_T = \frac{1}{T}\sum_{t=1}^{T} f_t$, then with probability at least $1 - \delta$,

$$L(\overline{f}_T) \leq \frac{1}{T}\sum_{t=1}^{T} \ell(f_t, Z_t) + \sqrt{\frac{2\log\frac{1}{\delta}}{T}}. \tag{14.8}$$

Proof. Let $Y_0 = 0$ and $Y_t = \sum_{s=1}^{t}(L(f_s) - \ell(f_s, Z_s))$. Since $\mathbf{E}\left[L(f_s) - \ell(f_s, Z_s) | Z_1, \ldots, Z_{s-1}\right] = 0$ for all s, (Y_t) is a martingale. Also,

$|Y_t - Y_{t-1}| = |L(f_t) - \ell(f_t, Z_t)| \leq 1$ for all t. Thus, by the Azuma-Hoeffding inequality with $B_t = Y_{t-1}$ and $c = 2$, for any $\epsilon > 0$,

$$\mathbf{P}[Y_T \geq \epsilon T] \leq \exp\left\{-\frac{2\epsilon^2 T^2}{4T}\right\} = \exp\left\{-\frac{\epsilon^2 T}{2}\right\}.$$

Setting $\epsilon = \sqrt{\frac{2\log\frac{1}{\delta}}{T}}$ and rearranging yields (14.7). For the second part, the assumed convexity of $\ell(\cdot, z)$ implies convexity of L, so by Jensen's inequality $L(\overline{f}_T) \leq \frac{1}{T}\sum_{i=1}^{T} L(f_t)$. Therefore, (14.8) follows from (14.7). □

An interpretation of (14.7) is that the average (over $t \in [T]$) generalization loss of the hypotheses generated by the online learning algorithm A is not much larger than $L_T((f_t))$, with high probability. Recall that $L_T((f_t))$ is the loss incurred by the learner, which is version of empirical loss, although for a sequence of hypotheses (f_t) rather than a single hypothesis. The bound (14.8) shows that, in the case $\ell(f, z)$ is convex in f, the hypotheses of the algorithm can be combined to yield a hypothesis $\overline{f}_T$ that has generalization loss, in the standard sense, not much larger than $L_T((f_t))$ incurred by the online learning algorithm.

In the statistical learning framework, if $\ell(f, z)$ is convex in f, Theorem 14.1 implies that the average $\overline{f}_T$ of the hypotheses (f_t) provided by the projected gradient descent algorithm represents an asymptotic ERM (AERM) algorithm. As we've seen earlier, generalization and AERM together provide consistency; the following corollary illustrates that point.[2]

Corollary 14.1. *Suppose $\mathcal{F}$ is a closed convex subset of a Hilbert space with finite diameter D. Suppose $\ell : \mathcal{F} \times \mathsf{Z} \to [0, 1]$ is such that $\ell(\cdot, z)$ is L-Lipschitz continuous for each fixed $z \in \mathsf{Z}$. Suppose $Z_1, Z_2, \ldots, Z_n$ are independent and identically distributed. Let $\overline{f}_n = \frac{1}{n}\sum_{t=1}^{n} f_t$, where $f_1, \ldots, f_n$ is produced by the projected gradient descent algorithm with step size $\alpha_t = \frac{D}{L\sqrt{2t}}$ for $t \geq 1$ (as in Theorem 14.1(a)). Then with probability at least $1 - 2\delta$,*

$$L(\overline{f_n}) \leq L^* + DL\sqrt{\frac{2}{n}} + \sqrt{\frac{8\log\frac{1}{\delta}}{n}},$$

Proof. The last statement of Theorem 14.2 with $T = n$ implies that, with probability at least $1 - \delta$,

$$L(\overline{f}_n) \leq \frac{1}{n}\sum_{t=1}^{n} \ell(f_t, Z_t) + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}.$$

Let f^* minimize the generalization loss: $L^* = L(f^*)$. Theorem 14.1(a) implies that, with probability one,

$$\frac{1}{n}\sum_{t=1}^{n} \ell(f_t, Z_t) \leq \frac{1}{n}\sum_{t=1}^{n} \ell(f^*, Z_t) + DL\sqrt{\frac{2}{n}}.$$

[2]Earlier we also found that, in a certain sense, generalization is equivalent to stability with respect to replace one sample perturbations. In the context of this section the Azuma-Hoeffding inequality is used to show generalization; stability is not used in this section.

The Azuma-Hoeffding inequality, as used in the proof of Theorem 14.2, and the choice $L(f^*) = L^*$, implies that with probability at least $1 - \delta$,

$$\frac{1}{n}\sum_{t=1}^{n} \ell(f^*, Z_t) \leq L^* + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}.$$

By the union bound, with probability at least $1-2\delta$, the previous three centered inequalities all hold, implying the corollary. □

If $\ell(f, z)$ is not convex in f, the bound (14.7) still implies that at least one of the hypotheses (f_t) generated by the algorithm has generalization loss $L(f_t)$ not much larger than the loss of the on-line learning algorithm. Moreover, for each t, the generalization loss, $L(f_t)$, can be estimated by applying f_t not only to Z_t, but to $Z_t, Z_{t+1}, \ldots, Z_T$, to help identify a value t^* so $L(f_{t^*}) \approx \min_t L(f_t)$. See [**CBCG04**] for details. This provides a single output hypothesis $f_{\hat{t}}$ with good generalization performance, even for non-convex loss. If somehow an AERM property is also true, a consistency result along the lines of Corollary 14.1 could be achieved for nonconvex loss.

CHAPTER 15

Minimax lower bounds

Now that we have a good handle on the performance of ERM and its variants, it is time to ask whether we can do better. For example, consider binary classification: we observe n i.i.d. training samples from an unknown joint distribution P on $\mathsf{X} \times \{0,1\}$, where X is some feature space, and for a fixed class $\mathcal{F}$ of candidate classifiers $f : \mathsf{X} \to \{0,1\}$ we let $\widehat{f}_n$ be the ERM solution

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} 1_{\{f(X_i) \neq Y_i\}}. \tag{15.1}$$

If $\mathcal{F}$ is a VC class with VC dimension V, then the excess risk of $\widehat{f}_n$ over the best-in-class performance $L^*(\mathcal{F}) \equiv \inf_{f \in \mathcal{F}} L(f)$ satisfies

$$L(\widehat{f}_n) \le L^*(\mathcal{F}) + C\left(\sqrt{\frac{V}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

with probability at least $1-\delta$, where $C > 0$ is some absolute constant. Integrating, we also get the following bound on the expected excess risk:

$$\mathbf{E}\left[L(\widehat{f}_n) - L^*(\mathcal{F})\right] \le C\sqrt{\frac{V}{n}}, \tag{15.2}$$

for some constant $C > 0$. Crucially, the bound (15.2) holds for *all* possible joint distributions P on $\mathsf{X} \times \{0,1\}$, and the right-hand side is *independent* of P — it depends only on the properties of the class $\mathcal{F}$! Thus, we deduce the following remarkable *distribution-free guarantee* for ERM: for any VC class $\mathcal{F}$, the ERM algorithm (15.1) satisfies

$$\sup_{P \in \mathcal{P}(\mathsf{X} \times \{0,1\})} \mathbf{E}_P\left[L_P(\widehat{f}_n) - L_P^*(\mathcal{F})\right] \le C\sqrt{\frac{V}{n}}. \tag{15.3}$$

(We have used subscript P to explicitly indicate the fact that the quantity under the supremum depends on the underlying distribution P. In the sequel, we will often drop the subscript to keep the notation uncluttered.) Let's take a moment to reflect on the significance of the bound (15.3). What it says is that, regardless of how "weird" the stochastic relationship between the feature $X \in \mathsf{X}$ and the label $Y \in \{0,1\}$ might be, as long as we scale our ambition back and aim at approaching the performance of the best classifier in some VC class $\mathcal{F}$, the ERM algorithm will produce a good classifier with a uniform $O(\sqrt{V/n})$ guarantee on its excess risk.

At this point, we stop and ask ourselves: could this bound be too pessimistic, even when we are so lucky that the optimal (Bayes) classifier happens to be in $\mathcal{F}$? (Recall that the

Bayes classifier for a given P has the form

$$f^*(x) = \begin{cases} 1, & \text{if } \eta(x) \ge 1/2 \\ 0, & \text{otherwise} \end{cases},$$

where $\eta(x) = \mathbf{E}[Y|X = x] = \mathbf{P}(Y = 1|X = x)$ is the *regression function*.) Let $\mathcal{P}(\mathcal{F})$ denote the subset of $\mathcal{P}(\mathsf{X} \times \{0, 1\})$ consisting of all joint distributions of $X \in \mathsf{X}$ and $Y \in \{0, 1\}$, such that $f^* \in \mathcal{F}$. Then from (15.2) we have

$$\sup_{P \in \mathcal{P}(\mathcal{F})} \mathbf{E}\left[L(\widehat{f}_n) - L(f^*)\right] \le C\sqrt{\frac{V}{n}}, \tag{15.4}$$

where $\widehat{f}_n$ is the ERM solution (15.1). However, we know that if the relationship between X and Y is deterministic, i.e., if $Y = f^*(X)$, then ERM performs much better. More precisely, let $\mathcal{P}_0(\mathcal{F})$ be the *zero-error class*:

$$\mathcal{P}_0(\mathcal{F}) := \{P \in \mathcal{P}(\mathcal{F}) : Y = f^*(X) \text{ a.s.}\}.$$

Then one can show that

$$\sup_{P \in \mathcal{P}_0(\mathcal{F})} \mathbf{E}\left[L(\widehat{f}_n) - L(f^*)\right] \le C\frac{V}{n}, \tag{15.5}$$

a much better bound than the "global" bound (15.4) (see, e.g., the book by Vapnik [**Vap98**]). This suggests that the performance of ERM is somehow tied to how "sharp" the behavior of η is around the decision boundary that separates the sets $\{x \in \mathsf{X} : \eta(x) \ge 1/2\}$ and $\{x \in \mathsf{X} : \eta(x) < 1/2\}$. To see whether this is the case, let us define, for any $h \in [0, 1]$, the class of distributions

$$\mathcal{P}(h, \mathcal{F}) := \{P \in \mathcal{P}(\mathcal{F}) : |2\eta(X) - 1| \ge h \text{ a.s.}\}$$

(in that case, the distributions in $\mathcal{P}(h, \mathcal{F})$ are said to satisfy the *Massart noise condition* with margin h.) We have already seen the two extreme cases:

- $h = 0$ — this gives $\mathcal{P}(0, \mathcal{F}) = \mathcal{P}(\mathcal{F})$ (the bound $|2\eta - 1| \ge 0$ holds trivially for any P).
- $h = 1$ — this gives the zero-error regime $\mathcal{P}(1, \mathcal{F}) = \mathcal{P}_0(\mathcal{F})$ (if $|2\eta - 1| \ge 1$ a.s., then η can take only values 0 and 1 a.s.).

However, intermediate values of h are also of interest: if a distribution P belongs to $\mathcal{P}(h, \mathcal{F})$ for some $0 < h < 1$, then its regression function η makes a jump of size h as we cross the decision boundary. With this in mind, for any $n \in \mathbb{N}$ and $h \in [0, 1]$ let us define the *minimax (excess) risk*

$$R_n(h, \mathcal{F}) := \inf_{\widetilde{f}_n} \sup_{P \in \mathcal{P}(h, \mathcal{F})} \mathbf{E}\left[L(\widetilde{f}_n) - L(f^*)\right], \tag{15.6}$$

where the infimum is over *all* learning algorithms $\widetilde{f}_n$ based on n i.i.d. training samples. The term "minimax" indicates that we are *minimizing* over all admissible learning algorithms, while *maximizing* over all distributions in a given class. The following result was proved by Pascal Massart and Élodie Nédélec [**MN06**]:

THEOREM 15.1. *Let $\mathcal{F}$ be a VC class of binary-valued functions on X with VC dimension $V \geq 2$. Then for any $n \geq V$ and any $h \in [0,1]$ we have the lower bound*

$$R_n(h, \mathcal{F}) \geq c \min\left(\sqrt{\frac{V}{n}}, \frac{V}{nh}\right), \tag{15.7}$$

where $c > 0$ is some absolute constant ($c = 1/32$ is sufficient).

Let us examine some implications:

- When $h = 0$, the right-hand side of (15.7) is equal to $c\sqrt{V/n}$. Thus, without any further assumptions, ERM is as good as it gets (it is minimax-optimal), up to multiplicative constants.
- When $h = 1$ (the zero-error case), the right-hand side of (15.7) is equal to cV/n, which matches the upper bound (15.5) up to constants. Thus, if we happen to know that we are in a zero-error situation, ERM is minimax-optimal as well.
- For intermediate values of h, the lower bound depends on the relative sizes of h, V, and n. In particular, if $h \geq \sqrt{V/n}$, we have the minimax lower bound $R_n(h, \mathcal{F}) \geq cV/nh$. Alternatively, for a fixed $h \in (0,1)$, we may think of $n^* = \lceil V/h^2 \rceil$ as the *cutoff sample size*, beyond which the effect of the margin condition on η can be "spotted" and exploited by a learning algorithm.
- In the same paper, Massart and Nédélec obtain the following upper bound on ERM:

$$\sup_{P \in \mathcal{P}(h,\mathcal{F})} \mathbf{E}\left[L(\widehat{f}_n) - L(f^*)\right] \leq \begin{cases} C\sqrt{\dfrac{V}{n}}, & \text{if } h \leq \sqrt{V/n} \\ C\dfrac{V}{nh}\left(1 + \log\dfrac{nh^2}{V}\right), & \text{if } h > \sqrt{V/n} \end{cases}. \tag{15.8}$$

  Thus, ERM is nearly minimax-optimal (we say "nearly" because of the extra log factor in the above bound; in fact, as Massart and Nédélec show, the log factor is unavoidable when the function class $\mathcal{F}$ is "rich" in a certain sense). The proof of the above upper bound is rather involved and technical, and we will not get into it here.

The appearance of the logarithmic term in (15.8) is rather curious. Given the lower bound of Theorem 15.1, one may be tempted to dismiss it as an artifact of the analysis used by Massart and Nédélec. However, as we will now see, in certain situations this logarithmic term is unavoidable. To that end, we first need a definition: We say that a class $\mathcal{F}$ binary-valued functions $f : \mathsf{X} \to \{0,1\}$ is (N,D)-*rich*, for some $N, D \in \mathbb{N}$, if there exist N distinct points $x_1, \ldots, x_N \in \mathsf{X}$, such that the projection

$$\mathcal{F}(x^N) = \Big\{ (f(x_1), \ldots, f(x_N)) : f \in \mathcal{F} \Big\}$$

of $\mathcal{F}$ onto x^N contains all binary strings of Hamming weight[1] D. Some examples:

- If $\mathcal{F}$ is a VC-class with VC dimension V, then it is (V,D)-rich for all $1 \leq D \leq V$. This follows directly from definitions.

[1] The Hamming weight of a binary string is the number of nonzero bits it has.

- A nontrivial example, and one that is relevant to statistical learning, is as follows. Let $\mathcal{F}$ be the collection of indicators of all halfspaces in $\mathbb{R}^d$, for some $d \ge 2$. There is a result in computational geometry which says that, for any $N \ge d+1$, one can find N distinct points $x_1, \dots, x_N$, such that $\mathcal{F}(x^n)$ contains all strings in $\{0,1\}^N$ with Hamming weight up to, and including, $\lfloor d/2 \rfloor$. Consequently, $\mathcal{F}$ is $(N, \lfloor d/2 \rfloor)$-rich for all $N \ge d+1$.

We can now state the following result [**MN06**]:

THEOREM 15.2. *Given some $D \ge 1$, suppose that $\mathcal{F}$ is (N,D)-rich for all $N \ge 4D$. Then*

$$R_n(h, \mathcal{F}) \ge c(1-h)\frac{D}{nh}\left[1 + \log\frac{nh^2}{D}\right] \tag{15.9}$$

for any $\sqrt{D/n} \le h < 1$, where $c > 0$ is some absolute constant ($c = 1/72$ is sufficient).

We will present the proofs of Theorems 15.1 and 15.2 in Sections 15.2 and 15.4, after giving some necessary background on minimax lower bounds.

15.1. The Bhattacharyya coefficient and bounds on average error for binary hypothesis testing

This section presents useful bounds from the theory of detection that will be used in the next section to provide a lower bound on the probability of error for concept learning. Consider a binary hypothesis testing problem with equal prior probabilities $\pi_0 = \pi_1 = \frac{1}{2}$. Suppose that the observation Y has probability mass function p_i under hypothesis H_i for $i = 0$ or $i = 1$. For any decision rule $f : \mathsf{Y} \to \{0,1\}$, the average probability of error is $\frac{1}{2}\sum_y p_0(y)f(y) + p_1(y)(1-f(y))$. A decision rule is Bayes optimal (i.e. minimizes the average error probability), if and only if it minimizes $p_0(y)f(y) + p_1(y)(1-f(y))$ for each y. That is, the Bayes optimal decision rules are given by

$$f^*(y) = \begin{cases} 1 & \text{if } p_1(y) > p_0(y) \\ 0 & \text{if } p_1(y) < p_0(y) \\ 0 \text{ or } 1 & \text{if } p_0(y) = p_1(y). \end{cases}$$

The corresponding minimum probability of error is given by

$$p_e^* = \frac{1}{2}\sum_y p_0(y) \wedge p_1(y).$$

Let the Bhattacharyya coefficient for a pair of discrete probability distributions p_0 and p_1 be defined by:

$$\rho = \sum_y \sqrt{p_0(y)p_1(y)}$$

Given probability mass functions $p_1, \dots, p_n$, let $\otimes_{j=1}^n p_j$ denote the product form joint probability mass function, $(y_1, \dots, y_n) \mapsto p_1(y_1)\cdots p_n(y_n)$. In other words, the joint distribution $\otimes_{j=1}^n p_j$ is the *tensor product* of the marginal distributions. The following lemma goes back at least as far as a report of Kraft in the 1950's (see [**Kai67**]).

Lemma 15.1. *(a) For any two distributions p_0 and p_1, the minimum average probability of error for the binary hypothesis testing problem with equal priors satisfies*[2]

$$\frac{\rho^2}{4} \le p_e^* \le \frac{\rho}{2}. \tag{15.10}$$

(b) The Bhattacharyya coefficient tensorizes. That is, the Bhattacharyya coefficient for tensor products is the product of the Bhattacharyya coefficients:

$$\rho(\otimes_{j=1}^n p_{1,j}, \otimes_{j=1}^n p_{0,j}) = \prod_{j=1}^n \rho(p_{1,j}, p_{0,j})$$

Proof. Summing the identity $p_0(y) + p_1(y) = 2(p_0(y) \wedge p_1(y)) + |p_0(y) - p_1(y)|$ over y yields

$$2 = 4p_e^* + \sum_y |p_0(y) - p_1(y)|. \tag{15.11}$$

The Cauchy-Schwarz inequality yields

$$\begin{aligned}
\sum_y |p_0(y) - p_1(y)| &= \sum_y \left|\sqrt{p_0(y)} + \sqrt{p_1(y)}\right|\left|\sqrt{p_0(y)} - \sqrt{p_1(y)}\right| \\
&\le \sqrt{\left(\sum_y \left(\sqrt{p_0(y)} + \sqrt{p_1(y)}\right)^2\right)\left(\sum_y \left(\sqrt{p_0(y)} - \sqrt{p_1(y)}\right)^2\right)} \\
&= \sqrt{2(1+\rho)2(1-\rho)} = 2\sqrt{1-\rho^2}. \qquad (15.12)
\end{aligned}$$

Combining (15.11)and (15.12), and using the fact $\sqrt{1-u} \le 1 - \frac{u}{2}$ for $0 \le u \le 1$ (square both sides to check) yields

$$p_e^* \ge \frac{1}{2}\left[1 - \sqrt{1-\rho^2}\right] \ge \frac{\rho^2}{4}.$$

For the other direction, note that $p_0(y) \wedge p_1(y) \le \sqrt{p_0(y)p_1(y)}$. Summing over y and dividing through by 2 yields $p_e \le \rho/2$. The proof of part (b) is left to the reader. □

Example 15.1. *Let $0 \le h \le 1$. Suppose under p_1, Y has the Bernoulli$\left(\frac{1+h}{2}\right)$ distribution and under p_0, Y has the Bernoulli$\left(\frac{1-h}{2}\right)$ distribution. Then $\rho = \sqrt{1-h^2}$ so that $\frac{1-h^2}{4} \le p_e^* \le \frac{\sqrt{1-h^2}}{2}$. More generally, if there are n observations, such that under H_i they are independent and identically distributed with the Bernoulli distribution p_i just mentioned, then the Bhattacharyya coefficient is $\rho^n = \sqrt{(1-h^2)^n}$ and the minimum average Bayesian error satisfies*

$$\frac{(1-h^2)^n}{4} \le p_{e,n,h}^* \le \frac{(1-h^2)^{n/2}}{2}. \tag{15.13}$$

[2] The proof shows the slightly stronger inequality, $\frac{1}{2}\left[1 - \sqrt{1-\rho^2}\right] \le p_e^* \le \frac{\rho}{2}$, which is equivalent to $\frac{1}{2}H^2 \le d_{TV} \le H\sqrt{1 - H^2/4}$, where $d_{TV}(p_0, p_1) = \frac{1}{2}\sum_y |p_0(y) - p_1(y)|$ and $H^2(p_0, p_1) \triangleq \sum_y(\sqrt{p_0(y)} - \sqrt{p_1(y)})^2$. H^2 is the version of Hellinger distance without factor $\frac{1}{2}$, and ρ is also known as the Hellinger affinity.

15.2. Proof of Theorem 15.1

Theorem 15.1 is proved in this section. For convenience, we repeat the assumptions:

- X is feature space
- $\mathcal{F}$ is a family of classifiers f such that $f : \mathsf{X} \to \{0, 1\}$.
- $\mathcal{P} = \mathcal{P}(\mathsf{X} \times [0, 1])$ is the set of all probability measures on $\mathsf{X} \times \{0, 1\}$
- For a given P, $L(f) = P\{Y \neq f(X)\}$ for $f \in \mathcal{F}$, and $L^* = \inf_{f \in \mathcal{P}} L(f)$. (As usual, for brevity, the subscript "P" is omitted from L in this instance.)
- For a given $P \in \mathcal{P}$, $\eta(X) \triangleq P(Y = 1|X)$. That is, η is the Bayes optimal predictor of Y given X for mean square error loss (not 0-1 loss!) under probability distribution P.
- For $h \in [0, 1]$, $\mathcal{P}(h)$ is the set of $P \in \mathcal{P}$ such that the corresponding estimator η can be taken to satisfy $\eta(x) \in \left[0, \frac{1-h}{2}\right] \cup \left[\frac{1+h}{2}, 1\right]$ for all $x \in \mathsf{X}$.
- A learning algorithm A maps Z^n to $\mathcal{F}$ for any given number of samples n. The min-max excess risk for $\mathcal{P}(h)$ and $\mathcal{F}$ is given by: $R_n(h, \mathcal{F}) \triangleq \inf_A \sup_{P \in \mathcal{P}(h)} E_n \left[L(\widehat{f}_n) - L^*\right]$, where $\widehat{f}_n = A(Z_1, \ldots, Z_n)$ and $Z_1, \ldots, Z_n$ are independent, each with distribution P. Note that $R_n(h, \mathcal{F})$ is nonincreasing in h because if $h' \geq h$, then $\mathcal{P}(h') \subset \mathcal{P}(h)$.

Suppose the VC dimension V of $\mathcal{F}$ is finite and $V \geq 2$. We shall show that for any $h \in [0, 1]$ and $n \geq V - 1$,

$$R_n(h, \mathcal{F}) \geq \frac{1}{16} \min\left(\sqrt{\frac{V-1}{n}}, \frac{V-1}{nh}\right). \tag{15.14}$$

This will imply Theorem 15.1.

By the definition of VC dimension, there exists a collection of V points, $\overline{x}_1, \ldots, \overline{x}_V$ in X such that $\mathcal{F}$ shatters $\overline{x}_1, \ldots, \overline{x}_V$. Thus, for every $b \in \{0, 1\}^{V-1}$, there is a function $f_b \in \mathcal{F}$ such that $(f_b(\overline{x}_1), \ldots, f_b(\overline{x}_{V-1})) = b$, and $f_b(\overline{x}_V) = 0$. Let $0 \leq p \leq 1$, to be determined later. Let P_X denote the discrete probability distribution on X that assigns probability $\frac{p}{V-1}$ to each of $\overline{x}_1, \ldots, \overline{x}_{V-1}$ and the remaining probability, $1 - p$, to the point x_V. For each $b \in \{0, 1\}^{V-1}$, let P_b denote the joint probability distribution on $\mathsf{X} \times \{0, 1\}$ that has marginal distribution of X equal to P_{X}, and conditional distribution of Y given X based on the classifier f_b with a noisy label, where the probability of label flipping is $\frac{1-h}{2}$. That is, P_b is a discrete probability distribution on $\mathsf{X} \times \{0, 1\}$ with marginal distribution P_{X} and $\eta_b(x) = \frac{1-h}{2} + h f_b(x)$ for $x \in \mathsf{X}$. Note the following facts about P_b and f_b :

- For $b \in \{0, 1\}^{V-1}$ and $f \in \mathcal{F}$, the generalization loss under P_b is given by $L_b(f) = \frac{1-h}{2} + \frac{ph}{V-1} \sum_{v=1}^{V-1} \mathbf{1}_{\{f(\overline{x}_v) \neq b_v\}} + (1-p)h\mathbf{1}_{\{f(\overline{x}_V) \neq 0\}}$.
- $P_b \in \mathcal{P}(h)$ for any $b \in \{0, 1\}^{V-1}$.
- f_b is a Bayes optimal classifier for probability distribution P_b and $L_b^* = L_b(f_b) = \frac{1-h}{2}$.

Let $\overline{P}^n$ denote the joint probability distribution on $\{0, 1\}^{V-1} \times (\mathsf{X} \times \{0, 1\})^n$ such that the joint distribution of $B, (X_1, Y_1), \ldots, (X_n, Y_n)$ under $\overline{P}^n$ is such that B is uniformly distributed over $\{0, 1\}^{V-1}$ and given $B = b$, the conditional distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$ is P_b^n.

Let $\widehat{f}_n$ denote a classifier determined by an arbitrary learning algorithm A applied to n random labeled samples, so $\widehat{f}_n = A((X_1, Y_1), \ldots, (X_n, Y_n))$. Note that for any $b \in \{0, 1\}^V$,

the joint distribution of
$(X_1,Y_1),\dots,(X_n,Y_n),\widehat{f}_n,b$ under P_b^n is the same as the conditional joint distribution of $(X_1,Y_1),\dots,(X_n,Y_n),\widehat{f}_n,B$ under $\overline{P}^n$ given $B=b$. For simplicity and without loss of generality, we assume $\widehat{f}_n$ always assigns the optimal label for feature $\overline{x}_V$, namely, $\widehat{f}_n(\overline{x}_V)=0$.

The following facts should now be apparent:

$$L_B(\widehat{f}_n)=\frac{1-h}{2}+\frac{ph}{V-1}\sum_{v=1}^{V-1}\mathbf{1}_{\{\widehat{f}_n(\overline{x}_v)\neq B_v\}}$$

$$\overline{E}^n[L_B(\widehat{f}_n)]=\frac{1-h}{2}+\frac{ph}{V-1}\sum_{v=1}^{V-1}\overline{P}^n\{\widehat{f}_n(\overline{x}_v)\neq B_v\}$$

$$\frac{1}{2^{V-1}}\sum_{b\in\{0,1\}^{V-1}}E_b^n[L_b(\widehat{f}_n)]=\frac{1-h}{2}+\frac{ph}{V-1}\sum_{v=1}^{V-1}\overline{P}^n\{\widehat{f}_n(\overline{x}_v)\neq B_v\}. \tag{15.15}$$

Recalling that $L_b^*=\frac{1-h}{2}$ for all $b\in\{0,1\}^{V-1}$ and using the fact that the maximum is greater than or equal to the average, (15.15) yields

$$\begin{aligned}\sup_{P\in\mathcal{P}(h)}(E^n[L(\widehat{f}_n)]-L^*)&\geq\max_{b\in\{0,1\}^{V-1}}(E_b^n[L_b(\widehat{f}_n)]-L_b^*)\\&\geq\frac{ph}{V-1}\sum_{v=1}^{V-1}\overline{P}^n\{\widehat{f}_n(\overline{x}_v)\neq B_v\}.\end{aligned} \tag{15.16}$$

We thus focus on finding a lower bound for the right-hand side of (15.16). For any $v\in[V-1]$ let $N_v=\big|\{i:X_i=\overline{x}_v\}\big|$, so that N_v is the number of samples with feature vector equal to $\overline{x}_v$. Given $X^n=(X_1,\dots,X_n)$, the $V-1$ vectors of labels $(Y_i:X_i=\overline{x}_v)_{1\leq v\leq V-1}$ are independent. So for each v, for the purposes of estimating B_v, it is optimal to base the decision on the N_v observations $(Y_i:X_i=\overline{x}_v)$. Thus, for each v, the learner faces a binary hypothesis testing problem of the form of Example 15.1. Applying the lower bound in (15.13) to each term on the right-hand side of (15.16), and using the fact that for each v, under $\overline{P}^n$, N_v has the binomial distribution with parameters n and $\frac{p}{V-1}$, thus yields

$$\begin{aligned}\overline{P}^n\{\widehat{f}_n(\overline{x}_v)\neq B_v|N_v\}&\geq\frac{1}{4}(1-h^2)^{N_v}\\\overline{P}^n\{\widehat{f}_n(\overline{x}_v)\neq B_v\}&\geq\frac{1}{4}\overline{E}^n[(1-h^2)^{N_v}]\\&=\frac{1}{4}\left(\frac{p}{V-1}(1-h^2)+1-\frac{p}{V-1}\right)^n\\&=\frac{1}{4}\left(1-\frac{ph^2}{V-1}\right)^n\end{aligned}$$

Then (15.16) yields:

$$\sup_{P\in\mathcal{P}(h)}(E^n[L(\widehat{f}_n)]-L^*)\geq\frac{ph}{4}\left(1-\frac{ph^2}{V-1}\right)^n \tag{15.17}$$

Since (15.17) holds for any learning algorithm A, it follows that

$$R_n(h, \mathcal{F}) \geq \frac{ph}{4}\left(1 - \frac{ph^2}{V-1}\right)^n \tag{15.18}$$

for $h, p \in [0,1]$, $V \geq 2$, and $n \geq 1$.

Case 1: Suppose $\sqrt{\frac{V-1}{n}} \leq h \leq 1$. Let $p = \frac{V-1}{(n+1)h^2}$. Then (15.18) yields

$$\begin{aligned} R_n(h, \mathcal{F}) &\geq \frac{V-1}{4(n+1)h}\left(1 - \frac{1}{n+1}\right)^n \\ &= \frac{V-1}{4nh}\left(1 - \frac{1}{n+1}\right)^{n+1} \geq \frac{V-1}{16hn} \end{aligned}$$

Case 2: Suppose $0 \leq h \leq \sqrt{\frac{V-1}{n}}$. Then the monotonicity of $R_n(h, \mathcal{F})$ and Case 1 for $h = \sqrt{\frac{V-1}{n}}$ (this value of h is in $[0,1]$ by the assumption $n \geq V-1$) yield

$$R_n(h, \mathcal{F}) \geq R_n\left(\sqrt{\frac{V-1}{n}}, \mathcal{F}\right) \geq \frac{1}{16}\sqrt{\frac{V-1}{n}}.$$

Thus, in either case, (15.14) holds, and Theorem 15.1 is proved.

15.3. A bit of information theory

The entropy of a random variable X with pmf p_X is defined by $H(X) = \sum_i p_X(i) \log \frac{1}{p_X(i)}$. The entropy is a measure of randomness or spread of a probability distribution. If $\log \frac{1}{p_X(i)}$ is a measure of how surprising it is to observe the value $X = i$ then entropy is the mean surprise of an observation. The maximum entropy distribution on n points is the uniform distribution, with entropy $\log n$. In this section the logarithms can be taken to be natural logarithms; $\log \mathrm{e} = 1$. If X and Y are jointly distributed, the conditional entropy of X given Y satisfies (applying Jensen's inequality to the concave function $-u \log u$):

$$\begin{aligned} H(Y|X) &\triangleq \sum_i H(Y|X=i) p_X(i) \\ &= \sum_i \sum_j -\left(p_{Y|X}(j|i) \log p_{Y|X}(j|i)\right) p_X(i) \\ &\leq \sum_j -\left(\sum_i p_{Y|X}(j|i) p_X(i)\right) \log \left(\sum_i p_{Y|X}(j|i) p_X(i)\right) \\ &= \sum_j -p_Y(j) \log p_Y(j) = H(Y). \end{aligned}$$

It is easy to verify the decomposition rule $H(X,Y) = H(X) + H(X|Y)$. Let $X - Y - Z$ denote the condition that X and Z are conditionally independent given Y. Then $X - Y - Z$ implies $H(X|Y,Z) = H(X|Y)$.

The Shannon mutual information between X and Y is defined by $I(X;Y) = H(X) - H(X|Y)$. Thus, $I(X;Y)$ is the reduction in the uncertainty of X due to learning Y.

The data processing inequality: If $X - Y - Z$ then $I(X;Y) \geq I(X;Z)$. The data processing inequality follows from the fact $H(X|Y) = H(X|Y,Z) \geq H(X|Z)$.

Recall that the KL divergence between two probability distributions on the same set is defined by

$$D(p\|q) = \sum_i p(i) \log \frac{p(i)}{q(i)},$$

with the understanding that $0 \log 0 = 0$. The mutual information can be expressed in terms of the KL divergence in the following way.[3]

$$I(X;Y) = \sum_i p_X(i) D(\, p_{Y|X}(\cdot|i) \parallel p_Y) \tag{15.19}$$

$$\triangleq D(p_{Y|X}\|p_Y|p_X) \tag{15.20}$$

$$= \sum_i \left\{ \sum_j p_{Y|X}(j|i) \log \frac{p_{Y|X}(j|i)}{p_Y(j)} \right\} p_X(i) \tag{15.21}$$

The representation (15.19), denoted by the notation (15.20), has a very nice interpretation. It shows that $I(X;Y)$ is the weighted average of the divergences of the conditional distributions, $p_{Y|X}(\cdot|i)$, of Y from their weighted average distribution, where the weights are $p_X(i)$. Replacing p_Y by another distribution Q in (15.21) yields

$$\begin{aligned} I(X;Y) &= \sum_i \left\{ \sum_j p_{Y|X}(j|i) \log \frac{p_{Y|X}(j|i)}{Q(j)} \right\} p_X(i) + \sum_i \sum_j \left\{ p_{Y|X}(j|i) \log \frac{Q(j)}{p_Y(j)} \right\} p_X(i) \\ &= D(p_{Y|X}\|Q|p_X) + \sum_j p_Y(j) \log \frac{Q(j)}{p_Y(j)} \\ &= D(p_{Y|X}\|Q|p_X) - D(p_Y\|Q) \end{aligned} \tag{15.22}$$

It is not hard to show using Jensen's inequality that $D(p_Y\|Q) \geq 0$ with equality if and only if $p_Y = Q$. Thus, (15.22) shows that for any distribution Q on the space of Y:

$$I(X;Y) \leq D(p_{Y|X}\|Q|p_X), \tag{15.23}$$

with equality if and only if $Q = p_Y$. The advantage of the bound (15.23) is that often Q can be a simpler distribution than p_Y, but for the bound to be effective, Q should be a reasonably good approximation to p_Y so the difference $D(p_Y\|Q)$ is small.

In the context of the geometric interpretation (15.19) of mutual information, (15.23) implies p_Y is the most central distribution on the space of Y values, in the sense that for any other distribution Q, the weighted average of the divergences of the conditional distributions from Q is minimized by $Q = p_Y$. (This reflects the fact that the KL divergence $D(\cdot\|\cdot)$ is a Bregman divergence.)

We close with two lemmas used in the next section. Let $\mathsf{h}(p) = -p \log p - (1-p) \log(1-p)$ for $0 < p < 1$ and values $\mathsf{h}(0) = \mathsf{h}(1) = 0$. That is, $\mathsf{h}(p)$ is the entropy of a Bernoulli(p) random variable.[4]

[3]Another expression for mutual information in terms of divergence is: $I(X;Y) = D(P_{X,Y}\|P_X \otimes P_Y)$, where $P_X \otimes P_Y$ is the product form distributions with the same marginal distribution as $P_{X,Y}$.

[4]Using sans serif here because the letter h is used for something else in the next section.

Lemma 15.2. *Let B be a random vector with values in $\{0,1\}^N$ such that $\mathbf{E}\left[\sum_i B_i\right] \le pN$ for some p with $0 \le p \le 0.5$. Then $H(B) \le N\mathsf{h}(p)$.*

Proof. Let $p_i = \mathbf{P}\{B_i = 1\}$. The assumption implies $\frac{1}{N}\sum_i p_i \le p$.

$$\begin{aligned} H(B) &= H(B_1) + H(B_2|B_1) + \cdots + H(B_N|B_1, \cdots, B_{N-1}) \\ &\le H(B_1) + H(B_2) + \cdots + H(B_N) \\ &= \mathsf{h}(p_1) + \cdots + \mathsf{h}(p_N) \\ &\le N\mathsf{h}\left(\frac{1}{N}\sum_{i=1}^N p_i\right) \le N\mathsf{h}(p), \end{aligned}$$

where the last two inequalities follow from the concavity of h and Jensen's inequality, and the fact h is increasing over the interval $[0, 0.5]$. □

Lemma 15.3. *For any $c_1 \in (0,1)$ there exists $c_2 \in (0,1)$ such that whenever $0 < a < b \le 0.25$ and $\mathsf{h}(a) \ge c_1 b \log\frac{1}{b}$ (e.g. $a\log\frac{1}{a} \ge c_1 b\log\frac{1}{b}$) then $a \ge c_2 b$. In particular, if $c_1 = \frac{1}{2}$, then $c_2 = \frac{1}{8}$ is sufficient.*

Proof. Given $c_1 \in (0,1)$, let $c_2 > 0$ be so small that

$$c_1 \ge 2c_2\left[1 + \frac{\log(1/c_2)}{\log(1/0.25)}\right].$$

It is easy to check that if $c_1 = \frac{1}{2}$ then $c_2 = \frac{1}{8}$ is sufficient. We will use the fact $a\log\frac{1}{a} \ge (1-a)\log\frac{1}{1-a}$ for $0 \le a \le 0.25$. (The inequality is readily checked for small a and can be checked numerically for larger a.) Suppose $0 < a < b \le 0.25$ and $\mathsf{h}(a) \ge c_1 b\log\frac{1}{b}$. Then,

$$\begin{aligned} a\log\frac{1}{a} &\ge \frac{1}{2}\mathsf{h}(a) \ge \frac{c_1 b}{2}\log\frac{1}{b} \\ &\ge c_2\left[1 + \frac{\log(1/c_2)}{\log(1/0.25)}\right] b\log\frac{1}{b} \\ &\ge c_2\left[b\log\frac{1}{b} + b\log\frac{1}{c_2}\right] = (bc_2)\log\frac{1}{bc_2} \end{aligned}$$

The function $u \mapsto u\log\frac{1}{u}$ is monotone increasing over $0 \le u \le 0.25$, so it follows that $a \ge bc_2$. □

15.4. Proof of Theorem 15.2

Theorem 15.2 is proved in this section. The first part of the proof closely follows the proof of Theorem 15.1, although it is simpler in one respect; we don't need to reserve one point (such as $\overline{x}_V$) and insist all classifiers of the form f_b assign label 0 to that point. Therefore we don't need the parameter p. Given $N \ge 4D$, let $\overline{x}_1, \ldots, \overline{x}_N$ denote any set of points such that the projection $\mathcal{F}(\overline{x}^N)$ contains $\{0,1\}^N_D$, the set of all binary strings of length N and Hamming weight D. For $b \in \{0,1\}^N_D$, let $f_b \in \mathcal{F}$ such that $(f(x_1), \ldots, f(x_N)) = b$. Fix an arbitrary learning algorithm A. Thus, for any $n \ge 1$, and n samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, the algorithm produces a classifier $\widehat{f}_n \in \mathcal{F}$.

Let P_X denote the distribution on X that assigns probability $\frac{1}{N}$ to each of the points $\overline{x}_1, \ldots, \overline{x}_N$. For any $b \in \{0,1\}^N_D$, let P_b denote the joint probability distribution on $\mathsf{X} \times \{0,1\}$

such that the marginal on X is P_X, and given X, the label Y is equal to $f(X)$ with probability $\frac{1+h}{2}$ and to $1-f(X)$ with probability $\frac{1-h}{2}$. Note that

$$L_b(f) = \frac{1-h}{2} + \frac{h}{N}\sum_{v=1}^{N} \mathbf{1}_{\{f(\overline{x}_v)\neq b_v\}}.$$

The classifier f_b achieves minimum loss for P_b: $L_b^* = L_b(f_b) = \frac{1-h}{2}$. The set of all such distributions satisfies $\mathcal{Q} = \{P_b : b \in \{0,1\}^N_D\} \subset \mathcal{P}(h)$, so it suffices to find a lower bound on $\max_{b\in\{0,1\}^N_D} E_b^n[L_b(\widehat{f})] - L_b^*$. Let $\overline{P}_n$ be the measure on $\{0,1\}^N_D \times (\mathsf{X}\times\{0,1\})^n$ as in the proof of Theorem 15.1. We find (15.16) holds (with the minor change mentioned above – there is no parameter p, and all points $\overline{x}_v$ are sampled with equal probability):

$$\sup_{P\in\mathcal{P}(h)} (E^n[L(\widehat{f}_n)] - L^*) \geq \max_{b\in\{0,1\}^N_D} (E_b^n[L_b(\widehat{f}_n)] - L_b^*) \geq h\epsilon \tag{15.24}$$

where

$$\epsilon \triangleq \frac{1}{N}\sum_{v=1}^{N} \overline{P}^n\{\widehat{f}_n(\overline{x}_v) \neq B_v\}.$$

In other words, ϵ is the average bit error probability, for estimation of $B = (B_1, \ldots, B_N)$ by $\widehat{f}_n = A((X_1, Y_1), \ldots, (X_n, Y_n))$, such that the joint distribution of $B, (X_1, Y_1), \ldots, (X_n, Y_n)$ is $\overline{P}^n$.

Thus, $R_n(h) \geq h \min_A \epsilon$. Therefore, the remainder of the proof is to obtain a lower bound on ϵ for an arbitrary algorithm A. The proof differs from the proof of Theorem 15.1 for two reasons. First, here the bits B_v are not independent – exactly D of them are equal to one with probability one. Secondly, while each bit is Bernoulli distributed under $\overline{P}^n$, it has parameter $\frac{D}{N}$ instead of $\frac{1}{2}$.

The idea of the proof is a variation of the idea of Fano, but it focuses directly on the average probability of bit errors, rather than on the probability of estimating the entire block of bits B correctly. Let $\widehat{B} = (\widehat{f}_n(\overline{x}_1), \ldots, \widehat{f}_n(\overline{x}_N))$. It is useful to think of B as the vector of hidden bits and $\widehat{B}$ is the estimate of B produced by the learning algorithm A based on the data. Let $B \oplus \widehat{B}$ denote the bit-by-bit modulo two sum (i.e. XOR sum) of the hidden bit vector B and its estimate. Thus, $B \oplus \widehat{B}$ indicates the location of estimation errors. In particular, $\epsilon = \frac{1}{N}\overline{E}^n[w_H(B \oplus \widehat{B}]$, where w_H denotes Hamming weight. Lemma 15.2 implies that $H(B \oplus \widehat{B}) \leq N\mathsf{h}(\epsilon)$. Also, $H(B) = \log\binom{N}{D}$. Therefore,

$$\begin{aligned} I(B, \widehat{B}) &= H(B) - H(B|\widehat{B}) \\ &= H(B) - H(B\oplus\widehat{B}|\widehat{B}) \leq H(B) - H(B\oplus\widehat{B}) \\ &= \log\binom{N}{D} - H(B\oplus\widehat{B}) \\ &\geq D\log\frac{N}{D} - N\mathsf{h}(\epsilon). \end{aligned}$$

To get an upper bound on $I(B; \widehat{B})$ we use the fact that the label for each data sample is noisy, limiting the information about B in each sample, and thus limiting the total information about B available to the learning algorithm. Note that

$$\begin{aligned}
I(B;\widehat{B}) &\overset{(a)}{\leq} I(B;(X^n,Y^n)) \\
&\overset{(b)}{=} I(B;X^n) + I(B;Y^n|X^n) \\
&\overset{(c)}{=} I(B;Y^n|X^n) \\
&\leq I(B;Y^n|X^n) + I(X^n;Y^n) \\
&\overset{(d)}{=} I((X^n,B);Y^n)
\end{aligned}$$

where (a) follows from the fact $\widehat{B}$ is a function of (X^n,Y^n) and the data processing theorem, (b) and (d) follow from properties of conditional mutual information or, equivalently, conditional entropy, and (c) follows by the independence of B and X^n.

To continue, we shall use the bound (15.23) based on the geometric properties of divergence. If $\frac{D}{N}$ is fairly small, then most of the B_v's are zero, so a reasonably good approximation for the Y's is obtained by assuming they are independent with each having the Bernoulli$\left(\frac{1-h}{2}\right)$ distribution. So let's choose Q to be the probability distribution on $\{0,1\}^n$ corresponding to independent Bernoulli$\left(\frac{1-h}{2}\right)$ random variables. That is, Q is the n-fold tensor product of the Bern$\left(\frac{1-h}{2}\right)$ distribution. By (15.23),

$$I((B,X^n);Y^n) \leq D(P_{Y|X^n,B}\|Q|P_{X^n,B}). \tag{15.25}$$

To compute the right-hand side of (15.25), consider a fixed possible value (x^n,b) of (X^n,B). The random variables $Y_1,\ldots,Y_n$ are conditionally independent given $(X^n,B)=(x^n,b)$ and they are also independent under distribution Q, so

$$D(P_{Y|(X^n,B)=(x^n,b)}\|Q) = \sum_{i=1}^{n} D\left(P_{Y_i|(X^n,B)=(x^n,b)}\middle\|\text{Bern}\left(\frac{1-h}{2}\right)\right). \tag{15.26}$$

The terms in the sum on the right-hand side of (15.26) fall into one of two groups. For each i, $x_i=\overline{x}_v$ for some value of v, and if $b_v=0$ then the conditional distribution of Y_i is the Bern$\left(\frac{1-h}{2}\right)$ distribution and the i^{th} term is zero. In contrast, if $x_i=\overline{x}_v$ and $b_v=1$, then the conditional distribution of Y_i is the Bern$\left(\frac{1+h}{2}\right)$ distribution, and the i^{th} term is the divergence distance between two Bernoulli distributions with parameters $\frac{1\pm h}{2}$, denoted by $d\left(\frac{1+h}{2}\|\frac{1-h}{2}\right)$. Thus, the right-hand side of (15.26) is $md\left(\frac{1+h}{2}\|\frac{1-h}{2}\right)$, where m is the number of times points $\overline{x}_v$ were sampled such that b_v is one. Since a fraction D/N of the bits in B are ones, and the points $\overline{x}_1,\ldots,\overline{x}_N$ are sampled uniformly to produce $X_1,\ldots,X_n$, when we average over the distribution of (X^n,B) as in the right-hand side of (15.25) we obtain:

$$\begin{aligned}
I((B,X^n),Y^n) &\leq D(P_{Y|X^n,B}\|Q|P_{X^n,B}) \\
&= \frac{nD}{N} d\left(\frac{1+h}{2}\middle\|\frac{1-h}{2}\right) \\
&= \frac{nDh}{N}\log\frac{1+h}{1-h} \\
&\leq \left(\frac{nD}{N}\right)\left(\frac{2h^2}{1-h}\right),
\end{aligned} \tag{15.27}$$

where the last step uses the fact $\log t \le t-1$ for $t>0$. Combining the upper and lower bounds on $I(B;\widehat{B})$ yields

$$\left(\frac{nD}{N}\right)\left(\frac{2h^2}{1-h}\right) \ge D\log\frac{N}{D} - N\mathsf{h}(\epsilon).$$

or equivalently,

$$\mathsf{h}(\epsilon) \ge \frac{D}{N}\log\frac{N}{D} - \left(\frac{nD}{N^2}\right)\left(\frac{2h^2}{1-h}\right). \tag{15.28}$$

A nice thing about the bound (15.28) is that it holds whenever n, D, N are positive integers with $1 \le D \le N$, and $0 \le h < 1$. To complete the proof we use (15.28) and the assumptions to get a lower bound on ϵ, and multiply that bound by h to get the lower bound on $R_n(h,\mathcal{F})$.

Select N so that the second term on the right-hand side of (15.28) is less than or equal to half the first term:

$$\left(\frac{nD}{N^2}\right)\left(\frac{2h^2}{1-h}\right) \le \frac{1}{2}\frac{D}{N}\log\frac{N}{D} \quad \text{or equivalently,}$$

$$\frac{4nh^2}{1-h} \le N\log\frac{N}{D}. \tag{15.29}$$

Specifically, let

$$N = \left\lfloor \frac{9nh^2}{(1-h)\left(1+\log\frac{nh^2}{D}\right)}\right\rfloor,$$

which satisfies (15.29) provided $h \ge \sqrt{D/n}$.[5]

Then

$$\mathsf{h}(\epsilon) \ge \frac{1}{2}\frac{D}{N}\log\frac{N}{D}.$$

An application of Lemma 15.3 then shows that, under the assumption $\frac{D}{N} \le 0.25$,

$$\epsilon \ge \frac{D}{8N}. \tag{15.30}$$

Since the algorithm A is arbitrary, it follows that $R_n(h,\mathcal{F}) \ge \frac{Dh}{8N}$. Using the explicit formula for N then yields (15.9) for $c = 1/72$.

Remark 15.1. *The ratio $\frac{D}{N}$ is the fraction of bits that have value 0, so an estimator that estimates all the bits to equal zero has bit error probability $\frac{D}{N}$. The bound (15.30) shows that using the observations, the estimation error is reduced from that level by at most a constant factor in the context of the proof.*

[5]The fact $\frac{nh^2}{D} \ge 1$ and $\log u \le u-1$ for $u \ge 1$ imply $nh^2/\left(1+\log\frac{nh^2}{D}\right) \ge D \ge 1$, and hence, that $N \ge \widetilde{N} \triangleq \frac{8nh^2}{(1-h)\left(1+\log\frac{nh^2}{D}\right)}$. Since the right-hand side of (15.29) is increasing in N, it suffices to prove (15.29) holds with N replaced by $\widetilde{N}$. The inequality follows easily if it can be shown that $\frac{\log\left(\frac{8}{1+s}\right)+s}{1+s} \ge \frac{1}{2}$, where $s \triangleq \log\frac{nh^2}{D} \ge 0$. The inequality is true for all $s \ge 0$ because $\log(1+s) \le \log 2 - \frac{1}{2} + \frac{s}{2}$.

APPENDIX A

Probability and random variables

Probability theory is the foundation of statistical learning theory. This appendix gives a quick overview of the main concepts and sets up the notation that will be used consistently throughout the notes. This is by no means intended as a substitute for a serious course in probability; as a good introductory reference, see the text by Gray and Davisson [**GD04**], which is geared towards beginning graduate students in electrical engineering.

Let Ω be a set. A nonempty collection $\mathcal{F}$ of subsets of Ω is called a *σ-algebra* if it has the following two properties:

(1) If $A \in \mathcal{F}$, then $A^c \equiv \Omega \backslash A \in \mathcal{F}$
(2) For any sequence of sets $A_1, A_2 \ldots \in \mathcal{F}$, their union belongs to $\mathcal{F}$: $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

In other words, any σ-algebra is closed under complements and countable unions. This implies, in particular, that the empty set $\varnothing$ and the entire set Ω are contained in any σ-algebra $\mathcal{F}$, and that such an $\mathcal{F}$ is closed under countable intersections. A pair $(\Omega, \mathcal{F})$ consisting of a set and a σ-algebra is called a *measurable space*. A *probability measure* on $(\Omega, \mathcal{F})$ is a function $\mathbf{P} : \mathcal{F} \to [0, 1]$, such that

(1) $\mathbf{P}(\Omega) = 1$
(2) Given any countably infinite sequence $A_1, A_2, \ldots \mathcal{F}$ of pairwise disjoint sets, i.e., $A_i \cap A_j = \varnothing$ for every pair $i \neq j$,

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbf{P}(A_i).$$

The triple $(\Omega, \mathcal{F}, \mathbf{P})$ is called a *probability space*.

Let $(\mathsf{X}, \mathcal{B})$ be some other measurable space. A *random variable* on Ω with values in X is any function $X : \Omega \to \mathsf{X}$ with the property that, for any $B \in \mathcal{B}$, the set

$$X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\}$$

lies in $\mathcal{F}$ X is said to be a *measurable* mapping from $(\Omega, \mathcal{F})$ into $(\mathsf{X}, \mathcal{B})$). Together, X and $\mathbf{P}$ induce a probability measure P_X on $(\mathsf{X}, \mathcal{B})$ by setting

$$P_X(B) := \mathbf{P}\left(X^{-1}(B)\right) \equiv \mathbf{P}\left(\{\omega \in \Omega : X(\omega) \in B\}\right),$$

which is called the *distribution* of X. (A mildly challenging character-building exercise is to try and prove that P_X is indeed a valid probability measure.) Once P_X is defined, we can typically forget all about $(\Omega, \mathcal{F}, \mathbf{P})$ and just work with P_X. Here are two standard examples to keep in mind. One is when X is a finite set, $\mathcal{B}$ is σ-algebra consisting of all subsets of X, and $x \mapsto P_X\{x\}$ is the probability mass function (pmf) of X, so for any $B \subseteq \mathsf{X}$,

$$P_X(B) = \sum_{x \in B} P_X(x). \tag{A.1}$$

The other is when X is the real line $\mathbb{R}$, $\mathcal{B}$ is the set of Borel subset of $\mathbb{R}$ (the smallest σ-algebra containing all open sets), and P_X has a probability density function (pdf) p_X, giving

$$P_X(B) = \int_B p_X(x)\mathrm{d}x. \tag{A.2}$$

for any $B \in \mathcal{B}$. We will use a more abstract notation that covers these two cases (and much more besides):

$$P_X(B) = \int_B P_X(\mathrm{d}x), \qquad \forall B \in \mathcal{B}. \tag{A.3}$$

When seeing something like (A.3), just think of (A.1) or (A.2).

If $f : \mathsf{X} \to \mathbb{R}$ is a real-valued function on X, the *expected value* of $f(X)$ is

$$\mathbf{E}[f(X)] = \int_{\mathsf{X}} f(x)P_X(\mathrm{d}x);$$

again, think of either

$$\mathbf{E}[f(X)] = \sum_{x\in\mathsf{X}} f(x)P_X(x)$$

in the case of discrete X, or

$$\mathbf{E}[f(X)] = \int_{\mathbb{R}} f(x)p_X(x)\mathrm{d}x$$

in the case of $\mathsf{X} = \mathbb{R}$ and a random variable with a pdf p_X.

For any two jointly distributed random variables $X \in \mathsf{X}$ and $Y \in \mathsf{X}$, we have their joint distribution P_{XY}, the marginals

$$P_X(A) := P_{XY}(A \times \mathsf{Y}) \equiv \int_{A\times\mathsf{Y}} P_{XY}(\mathrm{d}x, \mathrm{d}y)$$
$$P_Y(B) := P_{XY}(\mathsf{X} \times B) \equiv \int_{\mathsf{X}\times B} P_{XY}(\mathrm{d}x, \mathrm{d}y)$$

for all measurable sets $A \subseteq \mathsf{X}, B \subseteq \mathsf{Y}$, and the conditional distribution

$$P_{Y|X}(B|A) := \frac{P_{XY}(A \times B)}{P_X(A)}$$

of Y given that $X \in A$. Neglecting technicalities and considerations of rigor, we can define the conditional distribution of Y given $X = x$, denoted by $P_{Y|X}(\cdot|x)$, implicitly through

$$P_{XY}(A \times B) = \int_A P_X(\mathrm{d}x)\left(\int_B P_{Y|X}(\mathrm{d}y|x)\right).$$

Here, it is helpful to think of the conditional pmf

$$P_{Y|X}(y|x) = \frac{P_{XY}(x, y)}{P_X(x)}$$

in the discrete case, and of the conditional pdf

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)}$$

in the continuous case. The *conditional expectation* of any function $f : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ given X, denoted by $\mathbf{E}[f(X,Y)|X]$, is a random variable $g(X)$ that takes values in $\mathbb{R}$, such that for any bounded measurable function $b : \mathsf{X} \to \mathbb{R}$, [1]

$$\mathbf{E}[f(X,Y)b(X)] = \mathbf{E}[g(X)b(X)].$$

In other words, $\mathbf{E}[(f(X,Y) - g(X))b(X)] = 0$, or $f(X,Y) - g(X)$ is orthogonal to all bounded measurable functions of X. In particular, taking $b \equiv 1$, we get the *law of iterated expectations*: $\mathbf{E}[\mathbf{E}[f(X,Y)|X]] = \mathbf{E}[f(X,Y)]$.

Once again, think of

$$\mathbf{E}[f(X,Y)|X = x] = \sum_{y \in \mathsf{Y}} f(x,y) P_{Y|X}(y|x)$$

if both X and Y are discrete sets, and of

$$\mathbf{E}[f(X,Y)|X = x] = \int_{\mathsf{Y}} f(x,y) p_{Y|X}(y|x) \mathrm{d}y$$

if both X and Y are subsets of $\mathbb{R}$.

THEOREM A.1. *(Jensen's inequality) Let φ be a convex function and let X be a random variable such that $\mathbf{E}[X]$ is finite. Then $\mathbf{E}[\varphi(X)] \geq \varphi(\mathbf{E}[X])$.*

For example, Jensen's inequality implies that $\mathbf{E}[X^2] \geq \mathbf{E}[X]^2$, which also follows from the fact $\mathrm{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$.

PROOF. Since φ is convex, there is a tangent to the graph of φ at $\mathbf{E}[X]$. Equivalently, there is a subgradient g of φ at $\mathbf{E}[X]$, meaning that

$$\varphi(x) \geq \varphi(\mathbf{E}[X]) + \langle g, x - \mathbf{E}[X] \rangle \tag{A.4}$$

for all x. Replacing x by X and taking the expectation on each side of (A.4) yields the result. □

[1] As usual, we are being rather cavalier with the definitions here, since the choice of g is not unique; one typically speaks of different *versions* of the conditional expectation, which, properly speaking, should be defined w.r.t. the σ-algebra generated by X.

Bibliography

[Ang88] D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1988.

[Bar98] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[BCN16] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. arXiv preprint 1606.04838, 2016.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik–Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[BFT17] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[BHM18] Mikhail Belkin, Daniel J. Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 2300–2311, 2018.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BM02] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[BRT19] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In *Proceedings of Machine Learning Research*, volume 89, pages 1611–1619, 2019.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CBCG04] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

[CFZ09] B. Clarke, E. Fokoué, and H. H. Zhang. *Principles and Theory for Data Mining and Machine Learning*. Springer, 2009.

[Che52] H. Chernoff. A meausre of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[CZ07] F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[DL01] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

[Dud78] R. M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1978.

[Dwo06] C. Dwork. Differential privacy. In *Proceedings of the International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 1–12, 2006.

[FM09] J.C. Ferreira and V.A. Menegatto. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, 64(1):61–81, 2009.

[Fra87] P. Frankl. The shifting technique in extremal set theory. In *Surveys in Combinatorics, 1987 (New Cross, 1987), vol. 123 of London Math Soc. Lecture Note Ser.*, volume 123, pages 81–110. Cambridge University Press, 1987.

[Fra91] P. Frankl. Shadows and shifting. *Graphs and Combinatorics*, 7:23–29, 1991.

[FS97] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[GD04] R. M. Gray and L. D. Davisson. *An Introduction to Statistical Signal Processing.* Cambridge University Press, 2004.

[GG92] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression.* Kluwer, 1992.

[GN98] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, October 1998.

[GRS17] N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. arXiv preprint 1712.06541, 2017.

[GZ84] E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12:929–989, 1984.

[HAK07] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[Hau92] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 95:129–161, 1992.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[HR90] T. Hagerup and C. Rüb. A guided tour of Chernoff bounds. *Information Processing Letters*, 33(6):305–308, 1990.

[HRS16] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1225–1234, 2016. arXiv preprint 1509.01240.

[KAA$^+$00a] V. Koltchinskii, C. T. Abdallah, M. Ariola, P. Dorato, and D. Panchenko. Improved sample complexity estimates for statistical learning control of uncertain systems. *IEEE Transactions on Automatic Control*, 45(12):2383–2388, 2000.

[KAA$^+$00b] V. Koltchinskii, C. T. Abdallah, M. Ariola, P. Dorato, and D. Panchenko. Statistical learning control of uncertain systems: it is better than it seems. Technical Report EECE-TR-00-001, University of New Mexico, April 2000.

[Kai67] Thomas Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology*, 15(1):52–60, 1967.

[KP02] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1):1–50, 2002.

[KV94] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, 1994.

[Lin01] T. Linder. Learning-theoretic methods in vector quantization. In L. Györfi, editor, *Principles of Nonparametric Learning.* Springer, 2001.

[LLZ94] T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, 40(6):1728–1740, November 1994.

[LS99] D. D. Lee and H. S. Seung. Learing the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer, 1991.

[McD89] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

[Men03] S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures in Machine Learning*, volume 2600 of *Lecture Notes in Computer Science*, pages 1–40. Springer, 2003.

[MN06] P. Massart and É. Nédélec. Risk bounds for statistical learning. *Annals of Statistics*, 34(5):2326–2366, 2006.

[MP69] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry.* MIT Press, 1969.

[MP10] A. Maurer and M. Pontil. K-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, November 2010.

[MT07] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 94–103, 2007.

[NS15] K. Nissim and U. Stemmer. On the generalization properties of differential privacy. arXiv preprint 1504.05800, April 2015.

[Paj85] A. Pajor. Sous-espaces l_1^n des espaces de Banach. In *Travaux en Course [Works in Progress]*, volume 16. Hermann, Paris, 1985.

[Pol82] D. Pollard. Quantization and the method of k-means. *IEEE Transactions on Information Theory*, IT-28(2):199–205, March 1982.

[Pol87] B. T. Polyak. *Introduction to Optimization*. Translation Series in Mathematics and Engineering. Optimization Software, 1987.

[Sau72] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13:145–147, 1972.

[SFBL98] R. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.

[She72] S. Shelah. A combinatorial problem: stability and order for models and theories in infinity languages. *Pacific Journal of Mathematics*, 41:247–261, 1972.

[SHS01] B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In D. Helmbold and B. Williamson, editors, *Computational Learning Theory*, volume 2111 of *Lecture Notes in Computer Science*, pages 416–426. Springer, 2001.

[Sle62] D. Slepian. The one-sided barrier problem for Gaussian noise. *Bell Systems Technical Journal*, 41:463–501, 1962.

[SSBD14] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[SSSS10] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability, and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

[Sti86] S. M. Stigler. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, 1986.

[Tal14] Michel Talagrand. *Upper and Lower Bounds of Stochastic Processes: Modern Methods and Classical Problems*. Springer, 2014.

[Val84] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[VC71] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16:264–280, 1971.

[Vid98] M. Vidyasagar. Statistical learning theory and randomized algorithms for control. *IEEE Control Systems Magazine*, 18(6):162–190, 1998.

[Vid01] M. Vidyasagar. Randomized algorithms for robust controller synthesis using statistical learning theory. *Automatica*, 37:1515–1528, 2001.

[Vid03] M. Vidyasagar. *Learning and Generalization*. Springer, 2 edition, 2003.

[WD81] R. S. Wencour and R. M. Dudley. Some special Vapnik–Chervonenkis classes. *Discrete Mathematics*, 33:313–318, 1981.

[Zha04] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004.

[Zin03] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proc. 20th Int. Conf. Machine Learning (ICML-2003)*, 2003.

Index