Statistical Learning and Sequential Prediction

Alexander Rakhlin   and   Karthik Sridharan

DRAFT

October 16, 2014

Contents

Part I

Introduction

1

Introduction

This course will focus on theoretical aspects of *Statistical Learning* and *Sequential Prediction*. Until recently, these two subjects have been treated separately within the learning community. The course will follow a unified approach to analyzing learning in both scenarios. To make this happen, we shall bring together ideas from probability and statistics, game theory, algorithms, and optimization. It is this blend of ideas that makes the subject interesting for us, and we hope to convey the excitement. We shall try to make the course as self-contained as possible, and pointers to additional readings will be provided whenever necessary. Our target audience is graduate students with a solid background in probability and linear algebra.

"Learning" can be very loosely defined as the "ability to improve performance after observing data". Over the past two decades, there has been an explosion of both applied and theoretical work on machine learning. Applications of learning methods are ubiquitous: they include systems for face detection and face recognition, prediction of stock markets and weather patterns, speech recognition, learning user's search preferences, placement of relevant ads, and much more. The success of these applications has been paralleled by a well-developed theory. We shall call this latter branch of machine learning – "learning theory".

Why should one care about machine learning? The reality is that computers are now an integral part of our lives. Many tasks that we would like computers to perform cannot be hard-coded. The programs have to adapt. The goal then is to encode, for a particular application, as much of the domain-specific knowledge as needed, and leave *enough flexibility* for the system to improve upon observing data.

It is well-recognized that there is no single learning algorithm that will work universally (we will make this statement mathematically precise). It is *not* our goal to make computers learn everything at once: each application requires a lot of prior knowledge from the expert. The goal of learning theory then is to develop general guidelines and algorithms, and prove guarantees about learning performance under various natural assumptions.

A number of interesting learning models have been studied in the literature, and a glance at the proceedings of a learning conference can easily overwhelm a newcomer. Hence, we start this course by describing a few of the frameworks. We feel that the differences and similarities between various learning scenarios become more apparent once viewed as minimax problems. The minimax framework also makes it clear where the "prior knowledge" of the practitioner should be encoded. We will emphasize the minimax approach throughout the course.

What separates Learning from Statistics? Both look at data and have similar goals. Indeed, nowadays it is difficult to draw a line. Let us briefly sketch a few *historical* differences. According to [55], in the 1960's it became apparent that classical statistical methods are poorly suited for certain prediction tasks, especially those characterized by high dimensionality. Parametric statistics, as developed by Fisher, worked well if the statistician could *model the underlying process generating the data*. However, for many interesting problems (e.g. face detection, character recognition) the associated high-dimensional modeling problem was found to be intractable computationally, and the analysis given by classical statistics – inadequate. In order to avoid making assumptions on the data-generating mechanism, a new *distribution-free* approach was suggested. The goal within the machine learning community has therefore shifted from being *model-centric* to being *algorithm-centric*. An interested reader is referred to the (somewhat extreme) point of view of Breiman [13] for more discussion on the two cultures, but let us say that in the past 10 years both communities benefited from sharing of ideas. In the next lecture, we shall make the distinctions concrete by formulating the goals of nonparametric estimation and statistical learning as minimax problems. Further in the course, we will show that these goals are not as different as it might first appear.

Over the past 30 years, the development of Statistical Learning Theory has been intertwined with the study of uniform Laws of Large Numbers. The theory provided an understanding of the inherent complexities of distribution-free

learning, as well as *finite sample* and *data-dependent* guarantees. Besides the well-established theory, the algorithms developed by the learning community (e.g. Support Vector Machines and AdaBoost) are often considered to be state-of-the art methods for prediction problems. These methods adhere to the philosophy that, for instance, for classification problems one should not model the distributions but rather model the decision boundary. Arguably, this accounts for success of many learning methods, with the downside that interpretability of the results is often more difficult. The term "learning" itself is a legacy of the field's strong connection to computer-driven problems, and points to the fact that the goal is not necessarily that of "estimating the true parameter", but rather that of improving performance with more data.

In the past decade, research in learning theory has been shifting to sequential problems, with a focus on relaxing any distributional assumptions on the observed sequences. A rigorous analysis of sequential problems is a large part of this course. Interestingly, most research on sequential prediction (or, ***online learning***) has been algorithmic: given a problem, one would present a method and prove a guarantee for its performance. In this course, we present a thorough study of inherent complexities of sequential prediction. The goal is to develop it in complete parallel with the classical results of Statistical Learning Theory. As an added (and unexpected!) bonus, the online learning problem will give us an algorithmic toolkit for attacking problems in Statistical Learning.

We start the course by presenting a fun bit prediction problem. We then proceed to list in a rather informal way a few different learning settings, some of which are not "learning" per se, but quite closely related. We will not cover all these in the course, but it is good to see the breadth of problems anyway. In the following lecture, we will go through some of these problems once again and will look at them through the lens of minimax. As we go through the various settings, we will point out three key aspects: **(a) how data are generated; (b) how the performance is measured; and (c) where we place prior knowledge.**

Before proceeding, let us mention that we will often make over-simplified statements for the sake of clarity and conciseness. In particular, our definitions of research areas (such as Statistical Learning) are bound to be more narrow than they are. Finally, these lecture notes reflect a personal outlook and may have only a thin intersection with the reality.[1]

---

[1]For a memorable collection of juicy quotes, one is advised to take a course with J. Michael

Notation: A set $\{z_1, \ldots, z_t\}$ is variably denoted by either $z_{1:t}$ or z^t. A t-fold product of $\mathcal{Z}$ is denoted by $\mathcal{Z}^t$. The set $\{1, \ldots, n\}$ of natural numbers is denoted by $[n]$, and the set of all distributions on some set $\mathcal{A}$ by $\Delta(\mathcal{A})$.

Deviating from the standard convention, we sometimes denote random variables by lower-case letters, but we do so only if no confusion can arise. This is done for the purposes of making long equations appear more tidy.

Expectation with respect to a random variable z with distribution p is denoted by $\mathbb{E}_z$ or $\mathbb{E}_{z \sim p}$. We caution that in the literature $\mathbb{E}_z$ is sometimes used to denote a conditional expectation; our notation is more convenient for the problems we have in mind.

The inner product between two vectors is written variably as $a \cdot b$, or $\langle a, b \rangle$, or as $a^\top b$. The set of all functions from $\mathcal{X}$ to $\mathcal{Y}$ is denoted by $\mathcal{Y}^{\mathcal{X}}$. The unit L_p ball in $\mathbb{R}^d$ will be denoted by B_p^d and the unit ℓ_p ball by B_p.

2
An Appetizer: A Bit of Bit Prediction

We start our journey by describing the simplest possible scenario – that of "learning" with binary-valued data. We put "learning" in quotes simply because various research communities use this term for different objectives. We will now describe several such objectives with the aim of drawing parallels between them later in the course. Granted, the first three questions we ask are trivial, but the last one is not – so read to the end!

What can be simpler than the Bernoulli distribution? Suppose we observe a sample $y_1, \ldots, y_n$ drawn i.i.d. from such a distribution with an unknown bias $p \in (0, 1)$. The goal of **estimation** is to provide a guess of the population parameter p based on these data. Any kindergartner (raised in a Frequentist family) will happily tell us that a reasonable estimate of p is the empirical proportion of ones

$$\bar{y}_n \triangleq \frac{1}{n} \sum_{t=1}^{n} y_i,$$

while a child from a Bayesian upbringing will likely integrate over a prior and add a couple of extra 0's and 1's to regularize the solution for small n. What can we say about the quality of $\bar{y}_n$ as an estimate of p? From the Central Limit Theorem (CLT), we know that[1] $|p - \bar{y}_n| = O_P\left(n^{-1/2}\right)$, and in particular

$$\mathbb{E}|p - \bar{y}_n| = O(n^{-1/2}) \, .$$

For the **prediction** scenario, suppose again that we are given $y_1, \ldots, y_n$ drawn independently from the Bernoulli distribution with an unknown bias p, yet the

---

[1] For a sequence of random variables $y_1, \ldots, y_n, \ldots$ and positive numbers $a_1, \ldots, a_n, \ldots$, the notation $y_n = O_P(a_n)$ means that for any $\delta > 0$, there exists an $R > 0$ such that $\mathbf{P}\left(|y_n| > R a_n\right) < \delta$ for all n.

aim is to make a good binary forecast $\widehat{y} \in \{0, 1\}$ rather than to estimate p. The objective is the performance of the forecast on an independent draw y from the same distribution as measured by the indicator of a mistake $\mathbf{I}\{\widehat{y} \neq y\}$. Since y is a random variable itself, the decision $\widehat{y}$ incurs the expected cost of $\mathbb{E}\mathbf{I}\{\widehat{y} \neq y\}$. We may compare this cost to the cost of the best decision

$$\mathbb{E}\mathbf{I}\{\widehat{y} \neq y\} - \min_{y' \in \{0,1\}} \mathbb{E}\mathbf{I}\{y' \neq y\}$$

and observe that the minimum is attained at $y^* = \mathbf{I}\{p \geq 1/2\}$ and equal to

$$\min_{y' \in \{0,1\}} \mathbb{E}\mathbf{I}\{y' \neq y\} = \min\{p, 1 - p\}.$$

Also note that the minimum can only be calculated with the knowledge of p. However, since $\bar{y}_n$ is a good estimate of p, we can approximate the minimizer quite well. It is rather clear that we should predict with the majority vote $\widehat{y} = \mathbf{I}\{\bar{y}_n \geq 1/2\}$. Why?

Our third problem is that of **sequential prediction with i.i.d. data**. Suppose we observe the i.i.d. Bernoulli draws $y_1, \ldots, y_n, \ldots$ in a stream. At each time instant t, having observed $y_1, \ldots, y_{t-1}$, we are tasked with making the t-th prediction. It shouldn't come as a surprise that by going with the majority vote

$$\widehat{y}_t = \mathbf{I}\{\bar{y}_{t-1} \geq 1/2\}$$

once again, the average prediction cost

$$\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\mathbf{I}\{\widehat{y}_t \neq y_t\} - \min\{p, 1 - p\}$$

can be shown to be $O(n^{-1/2})$ once again. Another powerful statement can be deduced from the strong Law of Large Numbers (LLN):

$$\limsup_{n \to \infty} \left(\frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{\widehat{y}_t \neq y_t\} - \min\{\bar{y}_n, 1 - \bar{y}_n\} \right) \leq 0 \quad \text{almost surely} . \qquad (2.1)$$

That is to say, for almost all sequences (under the probabilistic model), the average number of mistakes is (asymptotically) no more than the smallest between the proportion of zeros and proportion of ones in the sequence.

We now leave the comfortable world of i.i.d. data where all the aforementioned results immediately followed from CLT or LLN. The fourth setting is that of **prediction of individual sequences**. Let us start with what should be a surprising statement:

13

> There exists a method for predicting the sequence that guarantees (2.1)
> *without any assumptions on the way the sequence is generated.*

Ponder for a minute on the meaning of this statement. It says that whenever the proportion of 1's (or 0's) in the sequence is, say, 70%, we should be able to correctly predict at least roughly 70% of the bits. It is not obvious that such a strategy even exists without any assumptions on the generative process of the data!

It should be observed that the method of predicting $\mathbf{I}\{\bar{y}_{t-1} \geq 1/2\}$ at step t no longer works. In particular, it fails on the alternating sequence $0, 1, 0, 1, \dots$ since $\mathbf{I}\{\bar{y}_{t-1} \geq 1/2\} = 1 - y_t$. Such an unfortunate sequence can be found for any deterministic algorithm: simply let y_t be the opposite of what the algorithm outputs given $y_1, \dots, y_{t-1}$. The only remaining possibility is to search for a randomized algorithm. The "almost sure" part of (2.1) will thus be with respect to algorithm's randomization, while the sequences are now deterministic. The roles have magically switched!

Let $q_t \in [0, 1]$ denote the bias of the distribution from which we draw the randomized prediction $\hat{y}_t$. Let us present two methods that achieve the goal (2.1). First method is defined with respect to a horizon n, which is subsequently doubled upon reaching (the details will be provided later), and the distribution is defined as

$$q_t = \frac{\exp\{-n^{-1/2}\sum_{s=1}^{t-1}(1-y_s)\}}{\exp\{-n^{-1/2}\sum_{s=1}^{t-1}y_s\} + \exp\{-n^{-1/2}\sum_{s=1}^{t-1}(1-y_s)\}}$$

We do not expect that this randomized strategy means anything to the reader at this point. And if it does – the next one should not. Here is a method due to D. Blackwell. Let L_{t-1} be the point in $[0, 1]^2$ with coordinates $(\bar{y}_{t-1}, \bar{c}_{t-1})$ where $\bar{c}_{t-1} = 1 - \frac{1}{t-1}\sum_{s=1}^{t-1}\mathbf{I}\{\hat{y}_s \neq y_s\}$ is the proportion of correct predictions of the algorithm thus far. If L_{t-1} is in the left or the right of the four triangles composing $[0, 1]^2$ (see figure below), choose q_t to be 0 or 1; otherwise draw a line from the center through L_{t-1} and let q_t be the value when this line intersects the x-axis. Why does this method work? Does it come from some principled way of solving such problems? We defer the explanation of the method to Chapter 21, but, meanwhile, we hope that these brief algorithmic sketches piqued readers' interest.

Foreshadowing the developments in the later part of the course, let us ask one more question: what other statements of the form (2.1) can we expect to get in the individual sequence framework? For instance, for which functions $\phi_n : \{0, 1\}^n \to \mathbb{R}$
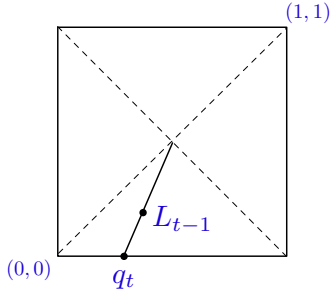
can one hope to find a prediction method that achieves

$$\forall y_1, \ldots, y_n, \quad \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\{\widehat{\boldsymbol{y}}_t \neq y_t\}\right] \leq \phi_n(y_1, \ldots, y_n) \tag{2.2}$$

In other words, what types of functions of the sequence upper bound the average number of mistakes an algorithm makes on that sequence? Of course, the smaller we can make this function, the better. But clearly $\phi_n \equiv 0$ is an impossibly difficult task (forcing us to make zero mistakes on any sequence) and $\phi_n \equiv 1$ is a trivial requirement achieved by any method. Hence, ϕ_n should be somewhere in-between. If we guess the bit by flipping a coin, the expected number of mistakes is $1/2$, and thus $\phi_n \equiv 1/2$ is feasible too. What about the more interesting (non-constant) functions?

Let's only focus only on those functions which are "stable" with respect to a coordinate flip:

$$|\phi_n(a) - \phi_n(a')| \leq 1/n \quad \text{for all } a, a' \in \{0, 1\}^n \text{ with } \|a - a'\|_1 = 1 \tag{2.3}$$

For such functions, the answer (due to T. Cover) might come as a surprise:

Proposition 2.1. *For a stable (in the sense of* (2.3)*) function* ϕ_n, *there exists an algorithm achieving* (2.2) *for all sequences* if and only if

$$\mathbb{E}\phi_n(y_1, \ldots, y_n) \geq 1/2 \tag{2.4}$$

where the expectation is under the uniform *distribution on* $\{0, 1\}^n$.

Let us show the easy direction. Fix an algorithm that enjoys (2.2). Suppose now that $y_1, \ldots, y_n$ are taken to be unbiased coin flips. Since the decision $\widehat{\boldsymbol{y}}_t$ is made before y_t is revealed, the expected loss is clearly $\mathbb{E}\mathbf{I}\{\widehat{\boldsymbol{y}}_t \neq y_t\} = 1/2$ for any

algorithm. To guarantee (2.2), it better be the case that (2.4) holds, as claimed. The other direction is rather unexpected: for any ϕ_n with the aforementioned property, there exists an algorithm that enjoys (2.2). We will show this in Section 25.2.

The above characterization is quite remarkable, as we only need to lower bound the expected value under the uniform distribution to ensure *existence* of an algorithm. Roughly speaking, the characterization says that a function ϕ_n can be smaller than $1/2$ for some sequences, but it then must be compensated by allowing for more errors on other sequences. This opens up the possibility of targeting those sequences we expect to observe in the particular application. If we can engineer a function ϕ_n that is small on those instances, the prediction algorithm will do well in practice. Of course, if we do have the knowledge that the sequence is i.i.d., we may simply choose $\min\{\bar{y}_n, 1 - \bar{y}_n\}$ as the benchmark. The ability to get good performance for non-i.i.d. sequences appears to be a powerful statement, and it foreshadows the development in these notes. We refer the reader to the exercises below to gain more intuition about the possible choices of ϕ_n.

So far, we have considered prediction of binary sequences in a "vacuum": the problem has little to do with any phenomenon that we might want to study in the real world. While it served as a playground for the introduction of several concepts, such a prediction problem is not completely satisfying. One typically has some *side information* and *prior knowledge* about the situation, and these considerations will indeed give rise to the complexity and applicability of the methods discussed in the course.

At this point, we hope that the reader has more questions than answers. Where do these prediction algorithms come from? How does one develop them in a more complicated situation? Why doesn't the simple algorithm from the i.i.d. world work? Is there a real difference in terms of learning rates between the individual sequence prediction and prediction with i.i.d. data? How far can the individual sequence setting be pushed in terms of applicability? These and many more questions will be addressed in the course.

We strongly encourage you to attempt these problems. If you solved all three, you are in a good shape for the course!

✎ **Exercise 2.1** (⋆)**.** From the above characterization of ϕ_n, conclude existence

of an algorithm that guarantees

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\{\widehat{y}_t \neq y_t\}\right] \leq \min\{\bar{y}_n, 1 - \bar{y}_n\} + Cn^{-1/2}$$

for any sequence. Can we take any $C > 0$? Find a good (or nearly best) constant C. (Hint: consult known bounds on lengths of random walks). Observe that $\mathbb{E}\min\{\bar{y}_n, 1 - \bar{y}_n\}$ by itself is less than $1/2$ and thus the necessary additional term is a compensation for "fluctuations".

✎ **Exercise 2.2** ($\star\star$)**.** Suppose that for each $i = 1, \ldots, k$, $\phi_n^i : \{0,1\}^n \to \mathbb{R}$ satisfies (2.4) as well as the stability condition (2.3). What penalty should we add to

$$\min_{i \in \{1,\ldots,k\}} \phi_n^i$$

to make sure the new "best of k" complexity satisfies (2.4)? Verify (2.3) for the new function and conclude that there must exist an algorithm that behaves not much worse than the given k prediction algorithms. (Hint: Use the stability property (2.3) together with McDiarmid inequality (see Appendix, Lemma A.1) to conclude subgaussian tails for $|\mathbb{E}\phi_n^i - \phi_n^i|$. Use union bound and integrate the tails to arrive at the answer.)

✎ **Exercise 2.3** ($\star \star \star$)**.** Suppose you have a hunch that the sequence $y_1, \ldots, y_n$ you will encounter can be partitioned into k parts with an imbalance of 0's and 1's within each part, but the endpoints of the segments are not known a priori. How can we leverage this information to get a better prediction method (if your hunch is correct)? Design a function ϕ_n that captures this prior knowledge for the best possible k-partition. Using the characterization above, prove that there exists an algorithm with overall prediction accuracy bounded by this function. (Hint: first, assume the partition is known and find the appropriate function that compensates for fluctuations within each interval.)

What are the Learning Problems?

Statistical Learning

Let us start with the so-called *supervised learning*. Within this setting, data are represented by pairs of input and output (also called predictor and response) variables, belonging to some sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. A *training set*, or a *batch of data*, will be denoted by

$$\{(\mathrm{X}_t, \mathrm{Y}_t)\}_{t=1}^{n} = (\mathrm{X}^n, \mathrm{Y}^n) \in \mathcal{X}^n \times \mathcal{Y}^n .$$

It is after observing this set that we hope to learn something about the relationship between elements of $\mathcal{X}$ and $\mathcal{Y}$.

For instance, X_t can be a high-dimensional vector of gene expression for the t-th patient and Y_t can stand for the presence or absence of diabetes. Such *classification* problems focus on binary "labels" $\mathcal{Y} = \{0, 1\}$. *Regression*, on the other hand, focuses on real-valued outputs, while *structured prediction* is concerned with more complex spaces of outcomes. The main (though not exclusive) goal of Statistical Learning is in *prediction*; that is, "learning" is equated with the ability to better predict the y-variable from the x-variable after observing the training data.

More specifically, *Statistical Learning* will refer to the following set of assumptions. We posit that the relationship between x and y is encoded by a fixed unknown distribution $P_{X \times Y}$. The observed data $\{(\mathrm{X}_t, \mathrm{Y}_t)\}_{t=1}^{n}$ are assumed to be drawn i.i.d. from this distribution. Furthermore, when time comes to evaluate our learning "progress", it is assumed that the world "will not be different". That is, our predictions are compared against the same fixed distribution from which the initial data were sampled (or observed). Both the i.i.d. and the "stationary world"

assumptions are rather strong, and we will spend the second part of this course on relaxing them.

So, how is our learning performance evaluated? Suppose, after observing the data, we can summarize the "learned" relationship between X and Y via a hypothesis $\widehat{\boldsymbol{y}} : \mathcal{X} \to \mathcal{Y}$. We may think of $\widehat{\boldsymbol{y}}$ as a "decision" from the set $\mathcal{D}$ of all functions mapping inputs to outputs. We then consider the average error in predicting Y from X based on this decision:

$$\mathbb{E}\left[\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, (X, Y)) \mid X^n, Y^n\right] \tag{3.1}$$

where the expectation is over the random draw of (X, Y) according to $P_{X \times Y}$ and independently of (X^n, Y^n), and $\boldsymbol{\ell} : \mathcal{D} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ is some *loss function* which measures the gravity of the mistake. As an example, if we are learning to classify spam emails, the measure of performance is how well we predict the spam label of an email drawn at random from the population of all emails.

If $\mathcal{Y} = \{0, 1\}$, we typically consider the *indicator loss*

$$\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, (X, Y)) = \mathbf{I}\left\{\widehat{\boldsymbol{y}}(X) \neq Y\right\}$$

or its surrogates (defined in later lectures), and for regression problems it is common to study the *square loss*

$$\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, (X, Y)) = (\widehat{\boldsymbol{y}}(X) - Y)^2$$

and the *absolute loss*

$$\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, (X, Y)) = |\widehat{\boldsymbol{y}}(X) - Y|.$$

Since the training data $\{(X_t, Y_t)\}_{t=1}^n$ are assumed to be a random draw, these examples might be misleading by chance. In this case, we should not penalize the learner. Indeed, a more reasonable goal is to ask that the average

$$\mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, (X, Y)) = \mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, (X, Y)) \mid X^n, Y^n\right\}\right\} \tag{3.2}$$

of (3.1) under the draw of training data be small. Alternatively, the goal might be to show that (3.1) is small with high probability. It is important to keep in mind that $\widehat{\boldsymbol{y}}$ is random, as it depends on the data. We will not write this dependence explicitly, but the hat should remind us of the fact.

If $P_{X \times Y}$ were known, finding $\widehat{\boldsymbol{y}}$ that minimizes $\mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, (X, Y))$ would be a matter of numerical optimization. But then there is no learning to be done, as the sought

after relationship between X and Y is known. It is crucial that the distribution is not available to us and the only information we receive about it is through the training sample. What do we do then? Well, as stated, the task of producing $\widehat{y}$ with a small error (3.2) is unsurmountable in general. Indeed, we mentioned that there is no universal learning method, and we just asked for exactly that! Assumptions or prior knowledge are needed.

Where do we encode this prior knowledge about the task? This is where Statistical Learning (historically) splits from classical Statistics. The latter would typically posit a particular form of the distribution $P_{X \times Y}$, called a *statistical model.* The goal is then to estimate the parameters of the model. Of course, one can make predictions based on these estimates, and we discuss this (plug-in) approach later. However, it is not necessary to perform the estimation step if the end-goal is prediction. Indeed, it can be shown quite easily (see [21]) that the goal of estimation is at least as hard as the goal of prediction for the problem of classification. For regression, the relationship between estimation and prediction with squared loss amounts to the study of well-specified and misspecified models and will be discussed in the course.

In contrast to the modeling approach, the classical work of Vapnik and Chervonenkis is of a *distribution-free* nature. That is, the prior knowledge does not enter in the form of an assumption on $P_{X \times Y}$, and our learning method should be "successful" for all distributions. Instead, practitioner's "inductive bias" is realized by selecting a class $\mathcal{F}$ of hypotheses $\mathcal{X} \to \mathcal{Y}$ that is believed to explain well the relationship between the variables. There is not much room to circumvent the problem of no universal learning algorithm! Requiring the method to be successful for all distributions means that the very notion of "successful" has to include the prior knowledge $\mathcal{F}$. A natural requirement is to minimize the loss *relative* to $\mathcal{F}$:

$$\mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, (X, Y)) - \inf_{f \in \mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f, (X, Y)) \tag{3.3}$$

Such a performance metric is called *regret*, and the loss function downshifted by the best in class $\mathcal{F}$ is called *excess loss.* Hence, "learning" is defined as the ability to provide a summary $\widehat{\boldsymbol{y}} : \mathcal{X} \to \mathcal{Y}$ of the relationship between X and Y such that the expected loss is competitive with the loss of the best "explanation" within the class $\mathcal{F}$. Generally, we do not require $\widehat{\boldsymbol{y}}$ itself to lie in $\mathcal{F}$, yet the learning algorithm that produces $\widehat{\boldsymbol{y}}$ will certainly depend on the benchmark set. We shall sometimes call the class $\mathcal{F}$ the "comparator class" and the term we subtract off in (3.3) – the

"comparator term". Such competitive framework is popular in a variety of fields and presents an alternative to the modeling assumption.

It can be argued that the distribution-free formulation in (3.3) should be used with a "growing" set $\mathcal{F}$, and idea formalized by Structural Risk Minimization and the method of sieves. A related idea is to estimate from data the parameters of $\mathcal{F}$ itself. Such methods will be discussed in the lecture on *model selection* towards the end of the course. In the next lecture, we will discuss at length the merits and drawbacks of the distribution-free formulation. For further reading on Statistical Learning Theory, we refer the reader to [12, 3, 38].

Let us also mention that independently of the developments by Vapnik and Chervonenkis, a learning framework was introduced in 1984 by Valiant within the computer science community. Importantly, the emphasis was made on polynomially computable learning methods, in the spirit of Theoretical Computer Science. This field of *Computational Learning Theory* started as a distribution-*dependent* (that is, non-distribution-free) study of classification. Fix a collection $\mathcal{F}$ of mappings $\mathcal{X} \rightarrow \{0, 1\}$. Suppose we can make a very strong prior knowledge assumption that one of the functions in $\mathcal{F}$ in fact exactly realizes the dependence between X and the label Y. Since (X, Y) is a draw from $P_{X \times Y}$, the assumption of

$$Y = f(X) \text{ for some } f \in \mathcal{F} \tag{3.4}$$

translates into the assumption that the conditional distribution $P_{Y|X=a} = \delta_{f(a)}$. It is then possible to characterize classes $\mathcal{F}$ for which the probability of error

$$P(\widehat{\boldsymbol{y}}(X) \neq f(X)) = \mathbb{E}|\widehat{\boldsymbol{y}}(X) - f(X)| \tag{3.5}$$

can be made small for some learning mechanism $\widehat{\boldsymbol{y}}$. This setting is termed *realizable* and forms the core of the original PAC (Probably Approximately Correct) framework of Valiant [53]. Of course, this is not a distribution-free setting. It was the papers of Haussler [25] and then Kearns et al [32] that extended the PAC framework to be distribution-free (termed *agnostic*).

In an intermediate setting between realizable and agnostic, one assumes *label noise*; that is, given X, the binary label Y is "often" equal to $f(X)$, except for some cross-over probability:

$$\exists f \in \mathcal{F} \text{ such that for any } x \in \mathcal{X}, \ P_{Y|X}(Y \neq f(a)|X = a) < \eta < 1/2 \tag{3.6}$$

This setting is very close to a modeling assumption made in Statistics, as discussed below. Under the condition (3.6), the probability of error (3.5) is a reasonable quantity to consider.

General Setting of Learning

Let us briefly mention a more general framework. Let us remove the assumption that data are of the form $\mathcal{X} \times \mathcal{Y}$ and instead write it as some abstract set $\mathcal{Z}$. This setting includes the so-called *unsupervised learning* tasks such as clustering and density estimation. Given the data $\{z_t\}_{t=1}^n$, the learning method is asked to summarize what it had learned by some element $\widehat{\boldsymbol{y}} \in \mathcal{D}$, yet we do not require $\mathcal{D}$ to be a class of functions $\mathcal{X} \to \mathcal{Y}$, and instead treat it as an abstract set. The quality of the prediction is assessed through a loss function $\boldsymbol{\ell} : \mathcal{D} \times \mathcal{Z} \to \mathbb{R}$, and we assume a fixed unknown distribution P_Z on $\mathcal{Z}$. As before, the data are assumed to be an i.i.d. sample from P_Z, and the performance measure

$$\mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, z)$$

is evaluated under the random draw $z \sim P_Z$.

The setting we described is almost too general. Indeed, the problem of finding $\widehat{\boldsymbol{y}}$ such that $\mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, z)$ is small is considered in such areas as Statistical Decision Theory, Game Theory, and Stochastic Optimization. One can phrase many different frameworks under any one of these umbrellas, and our aim is simply to make the connections clear.

Nonparametric Regression

Nonparametric regression with *fixed design* assumes that the data $\{(X_t, Y_t)\}_{t=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ consist of predictor-response pairs, where X_t's are fixed a priori (e.g. evenly spaced on the interval) and $Y_t = f(X_t) + \epsilon_t$, for independent zero-mean noise ϵ_t and some function $f : \mathcal{X} \to \mathcal{Y}$. For *random design*, we assume (as in Statistical Learning Theory) that X_t's are random and i.i.d. from some marginal distribution P_X, either known or unknown to us. Then, given X_t, we assume $Y_t = f(X_t) + \epsilon_t$, and this defines a joint distribution $P_{X \times Y}^f = P_X \times P_{Y|X}^f$ parametrized by f.

Given the data, let $\widehat{\boldsymbol{y}}$ summarize what has been learned. In statistical language, we construct an estimate $\widehat{\boldsymbol{y}}$ of the unknown f parametrizing the distribution $P_{X \times Y}^f$. The typical performance measure is the loss $\mathbb{E}(\widehat{\boldsymbol{y}}(X) - f(X))^2$ or some other p-norm. Instead of integrating with respect to the marginal P_X, the fixed-design setting only measures $(\widehat{\boldsymbol{y}}(X) - f(X))^2$ on the design points.

More generally, the notion of loss can be defined through the function $\boldsymbol{\ell} : \mathcal{D} \times \mathcal{F} \to \mathbb{R}$, such as some notion of a distance between functions or parameters [58].

The goal then is to ensure that

$$\mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, f) = \mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, f) \mid \mathbf{X}^n, \mathbf{Y}^n\right\}\right\} \tag{3.7}$$

is small.

Clearly, the study of (3.7) needs to depend on properties of $\mathcal{F}$, which embodies the prior assumptions about the problem. This prior knowledge is encoded as a distributional assumption on $P_{X \times Y}^f$ in a similar way to the "label noise setting" of PAC learning discussed above. In the next lecture, we will make the distinction between this setting and Statistical Learning even more precise.

Of course, this simple sketch cannot capture the rich and interesting field of nonparametric statistics. We refer to [52] and [57] for thorough and clear expositions. These books also study density estimation, which we briefly discuss next.

Density Estimation

Suppose that we observe i.i.d. data $\{z_t\}_{t=1}^n \in \mathcal{Z}^n$ from a distribution P_Z^f with a density f. The goal is to construct an estimate $\widehat{\boldsymbol{y}} : \mathcal{Z} \to \mathbb{R}$ of this unknown density. The error of the estimate is measured by, for instance, the integrated mean squared error (under the Lebesgue measure)

$$\mathbb{E}\left\{\int (\widehat{\boldsymbol{y}}(z) - f(z))^2 dz\right\}$$

or via the Kullback-Leibler divergence

$$\mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, f) = \int f(z) \log \frac{f(z)}{\widehat{\boldsymbol{y}}(z)} dz = \mathbb{E}\left\{\log \frac{f(\mathbf{Z})}{\widehat{\boldsymbol{y}}(\mathbf{Z})}\right\} = \mathbb{E}\left\{(-\log \widehat{\boldsymbol{y}}(\mathbf{Z})) - (-\log f(\mathbf{Z}))\right\}.$$

The KL divergence as a measure of quality of the estimator corresponds to a (negative) log loss $\boldsymbol{\ell}(f, z) = -\log f(z)$ which is central to problems in information theory.

Once again, construction of optimal estimators depends on the particular characteristics that are assumed about f, and these are often embodied via the distribution-dependent assumption $f \in \mathcal{F}$. The characteristics that lead to interesting statements are often related to the smoothness of densities in $\mathcal{F}$.

Summary: Let us discuss the main characteristics of the settings considered so far. First, they all assume a probabilistic nature of data. Moreover, the probability

distribution is typically assumed to be fixed in time. An i.i.d. sample is assumed to be available all at once as a batch, and the "performance" is assessed through $\hat{\boldsymbol{y}}$ with respect to the unknown distribution.

We now move beyond these "static" scenarios. In sequential problems we typically "learn" continuously as we observe more and more data, and there is a greater array of problems to consider. One important issue to look out for is whether the performance is measured throughout the sequence or only at the end. Problems also differ according to the impact the learner has through his actions. Yet another aspect is the amount of information or feedback the learner receives. Indeed, sequential problems offers quite a number of new challenges and potential research directions.

Universal Data Compression and Conditional Probability Assignment

In this scenario, suppose that a stream of $z_1, z_2, \ldots$ is observed. For simplicity, suppose that all z_t take on values in a discrete set $\mathcal{Z}$. There are two questions we can ask: (1) how can we sequentially predict each z_t having observed the prefix $\{z_s\}_{s=1}^{t-1}$? and (2) how can we compress this sequence? It turns out that these two questions are very closely related.

As in the setting of density estimation, suppose that the sequence is generated according to some distribution f on finite (or infinite) sequences in $\mathcal{Z}^n$ (or $\mathcal{Z}^\infty$), with f in some given set of distributions $\mathcal{F}$.[1] Once again, $\mathcal{F}$ embodies our assumption about the data-generating process, and the problem presented here is distribution-dependent. Later, we will introduce the analogue of universal prediction for *individual sequences*, where the prior knowledge $\mathcal{F}$ will be moved to the "comparator" in a way similar to distribution-free Statistical Learning.

So, how should we measure the quality of the "learner" in this sequential problem? Since z_t's are themselves random draws, we might observe unusual sequences, and should not be penalized for not being able to predict them. Hence, the measure of performance should be an expectation over the possible sequences. On

---

[1]Note the slight abuse of the notation: previously, we have used P^f do denote the distribution on (x, y) with the mean function f, then we used f to denote the unknown density of z, and now we use f to stand for the distribution itself. Hopefully, this should not cause much confusion.

each round, we can summarize what we learned so far from observing z^{t-1} via a conditional distribution $\widehat{\boldsymbol{y}}_t \in \mathcal{D} = \Delta(\mathcal{Z})$ — hence the name "conditional probability assignment". A natural measure of error at time t is then the expected log-ratio

$$\mathbb{E}\left\{\log\frac{f(z|z^{t-1})}{\widehat{\boldsymbol{y}}_t(z)} \,\middle|\, z^{t-1}\right\}$$

or some other notion of distance between the actual and the predicted distribution. When the mistakes are averaged over a finite-horizon time n, the measure of performance becomes

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left\{\log\frac{f(z|z^{t-1})}{\widehat{\boldsymbol{y}}_t(z)} \,\middle|\, z^{t-1}\right\} = \frac{1}{n}\mathbb{E}\left\{\log\frac{f(z^n)}{\widehat{\boldsymbol{y}}(z^n)}\right\}, \tag{3.8}$$

where $\widehat{\boldsymbol{y}}$ can be thought of either as a joint distribution over sequences, or as a collection of n conditional distributions for all possible prefixes. If data are i.i.d., the measure of performance becomes an averaged version of the one introduced for density estimation with KL divergence.

There are two main approaches to ensuring that the expected loss in (3.8) is small: (1) the plug-in approach involves estimating at each step the unknown distribution $f \in \mathcal{F}$ and using it to predict the next element of the sequence; (2) the mixture approach involves a Bayesian-type averaging of all distributions in $\mathcal{F}$. It turns out that the second approach is superior, a fact that is known in statistics as suboptimality of selectors, or suboptimality of plug-in rules. We refer the interested reader to the wonderful survey of Merhav and Feder [40]. In fact, the reader will find that many of the questions asked in that paper are echoed (and sometimes answered) throughout the present manuscript.

Prediction of Individual Sequences

Let us discuss a general setting of sequential prediction, and then particularize it to an array of problems considered within the fields of Statistics, Game Theory, Information Theory, and Computer Science.

At an abstract level, suppose we observe a sequence $z_1, z_2, \ldots \in \mathcal{Z}$ of data and need to make decisions $\widehat{\boldsymbol{y}}_1, \widehat{\boldsymbol{y}}_2, \ldots \in \mathcal{D}$ on the fly. Suppose we would like to lift the assumption that the sequence being observed is generated from some distribution from a given family, an assumption we made in the setting of universal lossless data compression. When we go to such a distribution-free setting for individual sequences, it actually no longer makes much sense to define per-round loss

as $\mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z)$ with the expectation over z according to some conditional distribution. In fact, let us assume that *there is no distribution governing the evolution of* $z_1, z_2, \ldots$. The sequence is then called *individual*[2]. This surprising twist of events might be difficult for statisticians to digest, and we refer to the papers of Phil Dawid on prequential statistics for more motivation. The (weak) prequential principle states that "any criterion for assessing the agreement between Forecaster and Nature should depend only on the actual observed sequences and not further on the strategies which might have produced these" [20].

If there is only one sequence, then how do we evaluate learner's performance? A sensible way is to score learner's instantaneous loss on round t by $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t)$, where the decision $\widehat{\boldsymbol{y}}_t$ must be chosen on the basis of z^{t-1}, but not z_t. Averaging over n, the overall measure of performance on the given sequence is

$$\frac{1}{n} \sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t). \tag{3.9}$$

Just as in our earlier exposition on statistical learning, it is easy to show that making the above expression small is an impossible goal even in simple situations. Some prior knowledge is necessary, and two approaches can be taken to make the problem reasonable. As we describe them below, notice that they echo the assumptions of "correctness of the model" and the alternative competitive analysis of statistical learning.

Assumption of Realizability   Let us consider the supervised setting: $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We make an assumption that the sequence $z_1, \ldots, z_n$ is only "in part" individual. That is, the sequence $x_1, \ldots, x_n$ is indeed arbitrary, yet y_t is given by $y_t = f(x_t)$ for some $f \in \mathcal{F}$. Additional "label noise" assumption has also been considered in the literature. Under the "realizability" assumption, the goal of minimizing (3.9) is feasible, as will be discussed later in the course. However, a much richer setting is the following.

Competitive Analysis   This is the most studied setting, and the philosophy is in a sense similar to that of going from distribution-dependent to distribution-free

---

[2]Strictly speaking, the term *individual sequence*, coming from information theory, refers to binary or finite-valued sequences. We use it more liberally for any sequences taking values in some abstract set $\mathcal{Z}$.

learning. Instead of assuming that some $f \in \mathcal{F}$ governs the evolution of the sequence, we push the assumption into the *comparator term* (the term we subtract off):

$$\text{Regret} = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{\ell}(f, z_t) \tag{3.10}$$

That is, the average loss of the learner is judged by the yardstick, which is the best fixed decision from the set $\mathcal{F}$ that we could have "played" for the n rounds. The reader probably noticed that we started to use a game-theoretic lingo. Indeed, the scenario has a very close relation to game theory.

More generally, we can define regret with respect to a set Π of strategies, where each strategy $\pi \in \Pi$ is a sequence of mappings π_t from the past to the set of decisions $\mathcal{F}$:

$$\text{Regret} = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{\pi \in \Pi} \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{\ell}(\pi_t(z^{t-1}), z_t) \tag{3.11}$$

From this point of view, (3.10) is regret with respect to a set of constant strategies.

The reader should take a minute to ponder upon the meaning of regret. If we care about our average loss (3.9) being small and we are able to show that regret as defined in (3.11) is small, we would be happy if we knew that the comparator loss is in fact small. But when is it small? When the sequences are well predicted by some strategy from the set Π. Crucially, we are *not* assuming that the data are generated according to a process related in any way to Π! All we are saying is that, *if* we can guarantee smallness of regret for all sequences and *if* the comparator term captures the nature of sequences we observe, our average loss is small. Nevertheless, it is important that the bound on regret that is proved for all individual sequences will hold... well... for all sequences. Just the interpretation might be unsatisfactory.

Let's see why the above discussion is similar to passing from distribution-dependent statistical learning to distribution-independent learning (with the redefined goal as in (3.3)). If we start a paper with the line "Assume that data are $Y_t = f(X_t) + \epsilon$ for some $f \in \mathcal{F}$", then whatever we prove is invalidated if the model is not correct (in the statistical language: *misspecified*). It is then unclear how badly the performance degrades as the assumption starts to become violated. On the other hand, putting the prior knowledge into the comparator term and asking for the method to hold for all distributions does not suffer from the problem of starting

with a wrong assumption. There is sometimes a price to pay, however, for moving the prior knowledge into the comparator term. The upper bounds one gets can be more conservative in general. How much more conservative will be a subject of interest in this course. Furthermore, the duality between assuming a form of a data-generating process versus competing with the related class of predictors is very intriguing and has not been studied much. We will phrase this duality precisely and show a number of positive results.

Summarizing, it should now be clear what the advantages and disadvantages of the regret formulation are: we have a setup of sequential decision-making with basically no assumptions that would invalidate our result, yet smallness of regret is "useful" whenever the comparator term is small. On the downside, protection against all models leads to more conservative results than one would obtain making a distributional assumption.

For a large part of the course we will study the regret notion (3.10) defined with respect to a single fixed decision. Hence, the upper bounds we prove for regret will hold for all sequences, but will be more useful for those sequences on which a single decision is good on all the rounds. What are such sequences? There is definitely a flavor of stationarity in this assumption, and surely the answer depends on the particular form of the loss ℓ. We will also consider regret of the form

$$\text{Regret} = \frac{1}{n} \sum_{t=1}^{n} \ell(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{(g_1,\dots,g_n)} \frac{1}{n} \sum_{t=1}^{n} \ell(g_t, z_t) \tag{3.12}$$

where the infimum is taken over "slowly-changing" sequences. This can be used to model an assumption of a non-stationary but slowly changing environment, without ever assuming that sequences come from some distribution. It should be noted that it is impossible in general to have an average loss comparable to that of the best unrestricted sequence $(g_1,\dots,g_n)$ of optimal decisions that changes on every round.[3]

The comprehensive book of Cesa-Bianchi and Lugosi [16] brings together many of the settings of prediction of individual sequences, and should be consulted as a supplementary reading. It is, in fact, after picking up this book in 2007 that our own interests in the field sparked. The following question was especially bothersome: why do upper bounds on regret defined in (3.10) look similar to those for

---

[3]Some algorithms studied in computer science do enjoy *competitive-ratio* type bounds that involve comparison with the best offline method — more on this later.

28

statistical learning with i.i.d. data? After all, there are no probabilistic assumptions placed on the individual sequences! This course will, in particular, address this question. We will show the connections between the two scenarios, as well as the important differences.

Let us mention that we will study an even more difficult situation than described so far: the sequence will be allowed to be picked not before the learning process, but *during* the process. In other words, the sequence will be allowed to change depending on the intermediate "moves" (or decisions) of the learner. This is a good model for learning in the environment on which learner's actions have an effect. We need to carefully define the rules for this, as to circumvent the so-called Cover's impossibility result. Since the sequence will be allowed to evolve in the worst-case manner, we will model the process as a game against an *adaptive adversary*. A good way to summarize the interaction between the learner and the adversary (environment, Nature) is by specifying the protocol:

Sequential prediction with adaptive environment

At each time step $t = 1$ to n,
- Learner chooses $\widehat{\boldsymbol{y}}_t \in \mathcal{D}$ and Nature simultaneously chooses $z_t \in \mathcal{Z}$
- Player suffers loss $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t)$ and both players observe $(\widehat{\boldsymbol{y}}_t, z_t)$

In the case of a non-adaptive adversary, the sequence $(z_1, \ldots, z_n)$ is fixed before the game and is revealed one-by-one to the player.

We remark that the name *individual sequence* typically refers to non-adaptive adversaries, yet we will use the name for the both adaptive and non-adaptive scenarios. There is yet another word we will abuse throughout the course: "prediction". Typically, prediction refers to the supervised setting when we are trying to predict some target response Y. We will use prediction in a rather loose sense, and synonymously with "decision making": abstractly, we "predict" or "play" or "make decision" or "forecast" $\widehat{\boldsymbol{y}}_t$ on round t even if the decision space $\mathcal{D}$ has nothing to do with the outcome space $\mathcal{Z}$. Finally, we remark that "online learning" is yet another name often used for sequential decision-making with the notion of regret.

Online Convex Optimization

Let us focus on regret against the best fixed comparator, as defined in (3.10), and make only one additional assumption: $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, z)$ is convex in the first argument.

Since no particular dependence on the second argument is assumed, we might as well slightly abuse the notation and equivalently rewrite $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, z)$ as $\boldsymbol{\ell}_z(\widehat{\boldsymbol{y}})$ or even $z(\widehat{\boldsymbol{y}})$, where $z : \mathcal{D} \to \mathbb{R}$ is a convex function and $\mathcal{D} = \mathcal{F}$ is a convex set. The latter notation makes it a bit more apparent that we are stepping into the world of convex optimization.

Regret can now be written as

$$\text{Regret} = \frac{1}{n} \sum_{t=1}^{n} z_t(\widehat{\boldsymbol{y}}_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} z_t(f) \tag{3.13}$$

where the sequence $z_1, \ldots, z_n$ is individual (also called *worst-case*), and each $z_t(f)$ is a convex function.

At the first sight, the setting seems restricted, yet it turns out that the majority of known regret-minimization scenarios can be phrased this way. We briefly mention that methods from the theory of optimization, such as gradient descent and mirror descent, can be employed to achieve small regret, and these methods are often very computationally efficient. These algorithms are becoming the methods of choice for large scale problems, and are used by companies such as Google. When data are abundant, it is beneficial to process them in a stream rather than as a batch, which means we are in the land of sequential prediction. Being able to relax distributional assumptions is also of great importance for present-day learning problems. It is not surprising that online convex optimization has been a "hot" topic over the past 10 years.

The second part of this course will focus on methods for studying regret in convex and non-convex situations. Some of our proofs will be algorithmic in nature, some – nonconstructive. The latter approach is particularly interesting because the majority of the results in the literature so far have been of the first type.

Multiarmed Bandits and Partial Information Games

The *exploration-exploitation dilemma* is a phenomenon usually associated with situation where an action that brings the most "information" does not necessarily yield the best performance. This dilemma does not arise when the aim is regret minimization and the outcome z_t is observed after predicting $\widehat{\boldsymbol{y}}_t$. Matters become more complicated when only partial information about z_t is observed, and a proper exploration-exploitation tradeoff is often key. The setting is often called "bandit" or "partial feedback".

In the original formulation of Robbins [47], the learner is faced with k decisions (arms of a multi-armed bandit), each producing a stochastic reward if chosen. A choice of arm i at time step t results in a random reward r_t drawn independently from the distribution p_i with support on $[0, 1]$ and mean μ_i. The goal is to minimize regret of not knowing the best arm in advance: $\max_{i \in \{1...k\}} \mu_i n - \sum_{t=1}^n r_t$. An optimal algorithm for this problem has been exhibited by Lai and Robbins [34]. Switching to losses instead of rewards, the goal can be equivalently written as minimization of expected regret

$$\mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^n \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t)\right\} - \inf_{f \in \mathcal{F}} \mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^n \boldsymbol{\ell}(f, z_t)\right\}$$

where $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, z) = \langle \widehat{\boldsymbol{y}}, z \rangle$, $\mathcal{D} = \mathcal{F} = \{e_1, \ldots, e_k\}$ the set of standard basis vectors, and expectation is over i.i.d. draws of $z_t \in [0, 1]^k$ according to the product of one-dimensional reward distributions $p_1 \times \ldots \times p_k$. Crucially, the decision-maker only observes the loss $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, z)$ (that is, one coordinate of z) upon choosing an arm. The setting is the most basic regret-minimization scenario for i.i.d. data where the exploration-exploitation dilemma arises. The dilemma would not be present had we defined the goal as that of providing the best hypothesis at the end, or had we observed the full reward vector z_t at each step.

Interestingly, one can consider the individual sequence version of the multi-armed bandit problem. It is not difficult to formulate the analogous goal: minimize regret

$$\mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^n \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^n \boldsymbol{\ell}(f, z_t)\right\}$$

where $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, z) = \langle \widehat{\boldsymbol{y}}, z \rangle$, the learner only observes the value of the loss, and the sequence $z_1, \ldots, z_n$ is arbitrary. Surprisingly, it is possible to develop a strategy for minimizing this regret. Generalizations of this scenario have been studied recently, and we may formulate the following protocol:

Bandit online linear optimization

At each time step $t = 1$ to n,
- Learner chooses $\widehat{\boldsymbol{y}}_t \in \mathcal{D}$ and Nature simultaneously chooses $z_t \in \mathcal{Z}$
- Player suffers loss $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) = \langle \widehat{\boldsymbol{y}}_t, z_t \rangle$ and the learner only observes this value.

One can also move from online linear to online convex optimization for even more generality, yet the question of optimal rates here is still open. In the stochastic setting with a fixed i.i.d. distribution for rewards, however, we may formulate the problem as minimizing regret

$$\mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^{n} z(\widehat{\boldsymbol{y}}_t) - \inf_{f\in\mathcal{F}} z(f)\right\} \tag{3.14}$$

where z is an unknown convex function and the learner receives a random draw from a fixed distribution with mean $z(\widehat{\boldsymbol{y}}_t)$ upon playing $\widehat{\boldsymbol{y}}_t$. An optimal (in terms of the dependence on n) algorithm for this problem has been recently developed.

Note that (3.14) basically averages the values of the unknown convex function z at the trajectory $\widehat{\boldsymbol{y}}_1,\ldots,\widehat{\boldsymbol{y}}_n$ and compares it to the minimum value of the function over the set $\mathcal{F}$. The reader will recognize this as a problem of optimization, but with the twist of measuring average error instead of the final error. This twist leads to the exploration-exploitation tradeoff which is absent if we can spend n iterations gathering information about the unknown function and then output the final answer based on all the information.

Convex Optimization with Stochastic and Deterministic Oracles

Convex optimization is concerned with finding the minimum of an unknown convex function $z(\widehat{\boldsymbol{y}})$ over a set $\mathcal{F}$, which we assume to be a convex set. The optimization process consists of repeatedly querying $\widehat{\boldsymbol{y}}_1,\ldots,\widehat{\boldsymbol{y}}_n \in \mathcal{D} = \mathcal{F}$ and receiving some information (from the Oracle) about the function at each query point. For deterministic optimization, this information is noiseless, while for stochastic optimization it is noisy. For the zero-th order optimization, the noiseless feedback to the learner (optimizer) consists of the value $z(\widehat{\boldsymbol{y}}_t)$, while for stochastic optimization the feedback is typically a random draw from a distribution with mean $z(\widehat{\boldsymbol{y}}_t)$. Similarly, first order noiseless information consists of a (sub)gradient of z at $\widehat{\boldsymbol{y}}_t$, while the stochastic feedback provides this value only on average. The deterministic optimization goal is to minimize $z(\widehat{\boldsymbol{y}}_n) - \inf_{f\in\mathcal{F}} z(f)$. In the stochastic case the goal is, for instance, in expectation:

$$\mathbb{E}\left\{z(\widehat{\boldsymbol{y}}_n) - \inf_{f\in\mathcal{F}} z(f)\right\}.$$

Once again, the difference between stochastic bandit formulation in (3.14) and stochastic optimization is in the way the goal is defined.

A particularly interesting stochastic optimization problem is to minimize the convex function

$$\tilde{z}(\widehat{\boldsymbol{y}}) = \mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, z) \tag{3.15}$$

Even if distribution is known, the integration becomes computationally difficult whenever $\mathcal{Z} \subset \mathbb{R}^d$ with even modest-size d [41]. The idea then is to generate an i.i.d. sequence $z_1, \ldots, z_n$ and use it in place of the difficult-to-compute integral. Since given the random draws we still need to perform optimization, we suppose there is an oracle that returns random (sub)gradients $G(\widehat{\boldsymbol{y}}, z)$ given the query $(\widehat{\boldsymbol{y}}, z)$ such that

$$\mathbb{E}G(\widehat{\boldsymbol{y}}, z) \in \partial \tilde{z}(\widehat{\boldsymbol{y}}) = \mathbb{E}\partial_{\widehat{\boldsymbol{y}}}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, z).$$

Given access to such information about the function, two approaches can be considered: stochastic approximation (SA) and sample average approximation (SAA). The SAA approach involves directly minimizing

$$\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, z_t)$$

as a surrogate for the original problem in (3.15). The SA approach consists of taking gradient descent steps of the type $\widehat{\boldsymbol{y}}_{t+1} = \widehat{\boldsymbol{y}}_t - \eta G(\widehat{\boldsymbol{y}}_t, z_t)$ with averaging of the final trajectory. We refer to [41] for more details.

Note that we have come a full circle! Indeed, (3.15) is the problem of Statistical Learning that we started with, except we did not assume convexity of $\boldsymbol{\ell}$ or $\mathcal{F}$. But the connections go even further: we will see in a few lectures that the SAA approach is a natural method called "Empirical Risk Minimization", while the SA method will make its appearance when we talk about sequential prediction problems.

4

Example: Linear Regression

Linear regression is, arguably, the most basic problem that can be considered within the scope of statistical learning, classical regression, and sequential prediction. Yet, even in this setting, obtaining sharp results without stringent assumptions proves to be a challenge. Below, we will sketch similar guarantees for methods in these three different scenarios, and then make some unexpected connections. Using the terminology introduced earlier, we consider the *supervised* setting, that is $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \mathbb{R}$. Let $\mathcal{F} = B_2^d$ be the unit ℓ_2 ball in $\mathbb{R}^d$. Each $f \in \mathbb{R}^d$ can be thought of as a vector, or as a function $f(x) = f \cdot x$, and we will move between the two representations without warning. Consider the square loss

$$\ell(f, (x, y)) = (f(x) - y)^2 = (f \cdot x - y)^2$$

as a measure of prediction quality. In order to show how various settings differ in their treatment of linear regression, we will make many simplifying assumptions along the way in order to present an uncluttered view. In particular, we assume that n is larger than d.

Classical Regression Analysis

We start with the so-called *fixed design* setting, where $x_1, \ldots, x_n$ are assumed to be fixed and only Y_t's are randomly distributed. Assume a *linear model*: there exists a $g^* \in \mathbb{R}^d$ such that

$$Y_t = g^* \cdot x_t + \epsilon_t, \tag{4.1}$$

34

for some independent zero-mean noise ϵ_t with, say, bounded variance σ^2. The model is frequently written in the matrix form as

$$Y = \mathbf{X}g^* + \epsilon$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ the matrix with rows x_t, Y and ϵ the vectors with coordinates Y_t and ϵ_t, respectively. Let

$$\hat{\Sigma} \triangleq \frac{1}{n} \sum_{t=1}^{n} x_t x_t^\mathsf{T}$$

be the covariance matrix of the design. Denoting by $\mathcal{D}$ the set of all functions $\mathcal{X} \to \mathcal{Y}$, the goal is to come up with an estimate $\hat{y} \in \mathcal{D}$ such that

$$\mathbb{E}\left\{ \frac{1}{n} \sum_{t=1}^{n} (\hat{y}(x_t) - g^*(x_t))^2 \right\} \tag{4.2}$$

is small. With a slight abuse of notation, whenever $\hat{y} \in \mathbb{R}^d$, we will write the above measure as

$$\mathbb{E}\|\hat{y} - g^*\|_{\hat{\Sigma}}^2.$$

In particular, let

$$\hat{y} = \operatorname*{argmin}_{g \in \mathbb{R}^d} \frac{1}{n} \sum_{t=1}^{n} (Y_t - g \cdot x_t)^2 = \operatorname*{argmin}_{g \in \mathbb{R}^d} \frac{1}{n} \left\| Y - \mathbf{X}g \right\|^2 \tag{4.3}$$

be the *ordinary least squares estimator*, or the *empirical risk minimizer*, over $\mathbb{R}^d$. Then

$$\hat{y} = \hat{\Sigma}^{-1}\left(\frac{1}{n} \sum_{t=1}^{n} Y_t x_t \right) = \frac{1}{n} \hat{\Sigma}^{-1} \mathbf{X}^\mathsf{T} Y$$

assuming the inverse exists (and use pseudo-inverse otherwise). Multiplying (4.1) on both sides by x_t and averaging over t, we find that

$$g^* = \hat{\Sigma}^{-1}\left(\frac{1}{n} \sum_{t=1}^{n} (Y_t - \epsilon_t) x_t \right),$$

and hence

$$\mathbb{E}\left\| \hat{y} - g^* \right\|_{\hat{\Sigma}}^2 = \mathbb{E}\left\| \hat{\Sigma}^{-1}\left(\frac{1}{n} \sum_{t=1}^{n} \epsilon_t x_t \right) \right\|_{\hat{\Sigma}}^2 = \frac{1}{n} \mathbb{E}\left[\epsilon \left(\frac{1}{n} \mathbf{X}\hat{\Sigma}^{-1}\mathbf{X}^\mathsf{T} \right) \epsilon \right]$$

Observe that the projection (or the *hat*) matrix $\frac{1}{n}\mathbf{X}\hat{\Sigma}^{-1}\mathbf{X}^{\mathsf{T}}$ can be written as UU^{T} where U has orthonormal columns (orthonormal basis of the column space of $\mathbf{X}$). Then the measure of performance (4.2) for ordinary least squares is

$$\frac{1}{n}\mathbb{E}\left\|U^{\mathsf{T}}\epsilon\right\|^2 \le \frac{\sigma^2 d}{n}$$

The random design analysis is similar, and the $O(d/n)$ rate can be obtained under additional assumptions on the distribution of X (see [28]).

CONCLUSION: In the setting of linear regression, we assume a linear relationship between predictor and response variables, and obtain an $O(d/n)$ rate for both fixed and random design.

Statistical Learning

In this setting we assume that $\{(X_t, Y_t)\}_{t=1}^n$ are i.i.d. from some unknown distribution $P_{X \times Y}$. We place *no prior assumption* on the relationship between X and Y. In particular, $\eta(a) \triangleq \mathbb{E}[Y|X = a]$ is not necessarily a linear function in $\mathbb{R}^d$. Recall that the goal is to come up with $\hat{\boldsymbol{y}}$ such that

$$\mathbb{E}\left(\hat{\boldsymbol{y}} \cdot X - Y\right)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f \cdot X - Y)^2 \tag{4.4}$$

is small.

   Before addressing this problem, we would like to motivate the goal of comparing prediction performance to that in $\mathcal{F}$. Why not simply consider unconstrained minimization as in the previous section? Let

$$g^* = \operatorname*{argmin}_{g \in \mathbb{R}^d} \mathbb{E}(g \cdot X - Y)^2 \tag{4.5}$$

be the minimizer of the expected error over all of $\mathbb{R}^d$, and $\hat{\boldsymbol{y}}$ be the ordinary least squares, as in (4.3). For any $g, g' \in \mathbb{R}^d$, using the fact that $\eta(X) = \mathbb{E}[Y|X]$,

$$\mathbb{E}(g \cdot X - Y)^2 - \mathbb{E}(g' \cdot X - Y)^2 = \mathbb{E}(g \cdot X - \eta(X))^2 - \mathbb{E}(g' \cdot X - \eta(X))^2. \tag{4.6}$$

Further,

$$\begin{aligned}
\mathbb{E}(g \cdot X - Y)^2 - \mathbb{E}(g^* \cdot X - Y)^2 &= \mathbb{E}(g \cdot X - g^* \cdot X + g^* \cdot X - Y)^2 - \mathbb{E}(g^* \cdot X - Y)^2 \\
&= \mathbb{E}((g - g^*) \cdot X)^2 + 2\mathbb{E}\left[(g^* \cdot X - Y)X^{\mathsf{T}}(g - g^*)\right]. \tag{4.7}
\end{aligned}$$

The cross-term in (4.7) is zero since because $\mathbb{E}\left[(XX^\top)g^* - YX\right] = 0$ is the optimality condition for g^*. Denoting the norm $\|f\|_X \triangleq (\mathbb{E}f(X)^2)^{1/2}$ for any $f : \mathcal{X} \to \mathbb{R}$, equations (4.6) and (4.7) give, for any $g \in \mathbb{R}^d$, the Pythagoras relationship

$$\|g - \eta\|_X^2 - \|g^* - \eta\|_X^2 = \|g - g^*\|_X^2 \, .$$

Attempting to control excess loss over all of $\mathbb{R}^d$ is equivalent to finding an estimator $\widehat{\boldsymbol{y}}$ that ensures convergence of $\|\widehat{\boldsymbol{y}} - g^*\|_X^2$ (which is simply $\|\widehat{\boldsymbol{y}} - g^*\|_\Sigma^2$) to zero, a task that seems very similar to the one in the previous section. However, the key difference is that y is no longer a random variable necessarily centered at $g^* \cdot X$. This makes the goal of estimating g^* difficult, and is only possible under some conditions. Hence, we would not be able to get a distribution-free result. Let us sketch the argument in [28]: we may write

$$\Sigma^{1/2}(\widehat{\boldsymbol{y}} - g^*) = \Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}\hat{\mathbb{E}}\left[\Sigma^{-1/2}X(\eta(X) - g^* \cdot X)\right] + \Sigma^{1/2}\hat{\Sigma}^{-1/2}\hat{\mathbb{E}}\left[\hat{\Sigma}^{-1/2}X(Y - \eta(X))\right]$$

A few observations can be made. First, by optimality of g^*, $\mathbb{E}(X(\eta(X) - g^* \cdot X)) = 0$. Of course, we also have $\mathbb{E}(Y - \eta(X)) = 0$. Further, $\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}$ can be shown to be tightly concentrated around identity. We refer the reader to [28] for a set of conditions under which the above decomposition gives rise to $O(d/n)$ rates.

Instead of making additional distributional assumptions to make the unrestricted minimization problem feasible, we can instead turn to the goal of proving upper bounds on (4.4). Interestingly, this objective requires somewhat different techniques. Let us denote by $\hat{f}$ the empirical minimizer constrained to lie within the set $\mathcal{F}$ and f^* be the minimizer of the expected loss within $\mathcal{F}$. As depicted in Figure **??**, we may view $\hat{f}$ and f^* as projections (although with respect to different norms) of the unconstrained optima $\widehat{\boldsymbol{y}}$ and g^*, respectively, onto $\mathcal{F}$. The projections may be closer than the unconstrained minimizers, although by itself this argument is not enough since the inequalities are in the opposite direction:

$$\|f - f^*\|_X^2 \le \|f - \eta\|_X^2 - \|f^* - \eta\|_X^2 = \mathbb{E}\left(f \cdot X - Y\right)^2 - \mathbb{E}(f^* \cdot X - Y)^2$$

for any $f \in \mathcal{F}$ (Exercise). With tools developed towards the end of the course, we will be able to show that excess loss (4.4) in the distribution-free setting of Statistical Learning is indeed upper bounded by $O(d/n)$.

We conclude this section by observing that the problem (4.4) falls under the purview of Stochastic Optimization. By writing

$$\mathbb{E}(g \cdot X - Y)^2 = g^\top \Sigma g - 2g\mathbb{E}(XY) + Y^2$$

we notice that the condition number of Σ plays a key role. In particular, if $\Sigma = I$, we expect a $O(1/n)$ convergence of the ordinary least squares, without the dependence on d.

CONCLUSION: In the setting of Statistical Learning, without assuming any particular form of the relationship between X and Y, we obtain an $O(d/n)$ rate for the prediction error of the least squares estimator restricted to $\mathcal{F}$ relative to the best linear predictor in $\mathcal{F}$.

Prediction of Individual Sequences

In the individual sequence scenario, we are tasked with designing an estimator (forecaster) $\widehat{\boldsymbol{y}}_t$ on round t in an attempt to predict well on the next observation (x_t, y_t). Recall that the goal is to achieve small regret, defined as

$$\frac{1}{n}\sum_{t=1}^{n}(\widehat{\boldsymbol{y}}_t(x_t) - y_t)^2 - \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}(f \cdot x_t - y_t)^2 \ . \tag{4.8}$$

So, how do we choose $\widehat{\boldsymbol{y}}_{t+1}$ based on the data $\{(x_s, y_s)\}_{s=1}^{t}$ observed so far? It might be tempting to simply define it as a linear function $\widehat{\boldsymbol{y}}_t(x) = \widehat{\boldsymbol{y}}_t \cdot x$ via the least squares solution (4.3) on the prefix of data. However, to avoid suffering huge regret in the beginning of the sequence, we need to add a small regularization term:

$$\widehat{\boldsymbol{y}}_t = \operatorname*{argmin}_{f \in \mathbb{R}^d} \sum_{s=1}^{t-1}(f \cdot x_s - y_s)^2 + \|f\|^2 \tag{4.9}$$

The influence of the extra term diminishes as t increases, thus approaching the least squares solution. The method is readily recognized as *ridge regression* or *regularized least squares.* Unfortunately, the method suffers from a major drawback: the norm of the output $\widehat{\boldsymbol{y}}_t$ (and, hence, the prediction $\widehat{\boldsymbol{y}}_t \cdot x_t$ can grow with t. If an a priori bound on y_t's is known, one may clip the predictions to that interval; otherwise, there does not seem to be a clean way to analyze the method.

Observe that our prediction $\widehat{\boldsymbol{y}}_t$ only enters through the product $\widehat{\boldsymbol{y}}_t \cdot x_t$. This suggests that one may define a different $\widehat{\boldsymbol{y}}_t$ for each possible x_t thus making the function $\widehat{\boldsymbol{y}}_t$ non-linear. It turns out, the following simple modification of (4.9) works beautifully:

$$\widehat{\boldsymbol{y}}_t(x) = w_t^x \cdot x, \quad \text{where} \quad w_t^x = \operatorname*{argmin}_{f \in \mathbb{R}^d} \sum_{s=1}^{t-1}(f \cdot x_s - y_s)^2 + \|f\|^2 + (f \cdot x)^2 \tag{4.10}$$

The method is called the Vovk-Azoury-Warmuth forecaster. Denoting $w_t = w_t^{x_t}$, observe that the closed form is simply

$$w_t = \left(I + \sum_{s=1}^{t} x_s x_s^\top\right)^{-1} \left(\sum_{s=1}^{t-1} y_s x_s\right). \tag{4.11}$$

Denoting $\Sigma_t = \left(I + \sum_{s=1}^{t} x_s x_s^\top\right)$, the update from w_t to w_{t+1} can be written as

$$w_{t+1} = \Sigma_{t+1}^{-1} \left(\Sigma_t w_t + y_t x_t\right)$$

or, in an incremental form, as

$$w_{t+1} = w_t - \Sigma_t^{-1}(x_t x_t^\top w_{t+1} - x_t y_t) \quad \text{or} \quad w_{t+1} = w_t - \Sigma_{t+1}^{-1}(x_{t+1} x_{t+1}^\top f_t - x_t y_t) \tag{4.12}$$

Using the definitions of w_t, w_{t+1}, the identity

$$(w_t \cdot x_t - y_t)^2 - (f \cdot x_t - y_t)^2 = \|f - w_t\|_{\Sigma_t}^2 - \|f - w_{t+1}\|_{\Sigma_{t+1}}^2 + \|w_t - w_{t+1}\|_{\Sigma_{t+1}}^2$$
$$+ (w_t \cdot x_t)^2 - (w_t \cdot x_{t+1})^2 + (f \cdot x_{t+1})^2 - (f \cdot x_t)^2$$

can be shown to hold for any f. Further, with the help of the closed form for updates in (4.12),

$$\|w_t - w_{t+1}\|_{\Sigma_{t+1}}^2 = x_t^\top \Sigma_t^{-1} x_t (y_t)^2 - x_{t+1}^\top \Sigma_t^{-1} x_{t+1} (w_{t+1} \cdot x_{t+1})^2 + (w_t \cdot x_{t+1})^2 - (w_{t+1} \cdot x_{t+1})^2$$

When we sum over n time steps, most terms telescope leaving us with

$$\sum_{t=1}^{n} (w_t \cdot x_t - y_t)^2 - \sum_{t=1}^{n} (f \cdot x_t - y_t)^2 = \|f - w_1\|_{\Sigma_1}^2 - \|f - w_{n+1}\|_{\Sigma_{n+1}}^2 + \sum_{t=1}^{n} (y_t^2) x_t^\top \Sigma_t^{-1} x_t + R \tag{4.13}$$

where the remainder R is equal to

$$R = -\sum_{t=1}^{n} (w_{t+1} \cdot x_{t+1})^2 x_{t+1}^\top \Sigma_t^{-1} x_{t+1} + (w_1 \cdot x_1)^2 - (w_{n+1} \cdot x_{n+1})^2 + (f \cdot x_{n+1})^2 - (f \cdot x_1)^2$$

Since x_{n+1} is a phantom quantity used only for analysis, we may set it to 0. Together with $\hat{f}_1 = 0$, we conclude that $R \leq 0$. Let $B^2 = \max_{t \in [n]} y_t^2$. We now use the identity

$$x^\top (\Sigma + x x^\top)^{-1} x = 1 - \frac{\det(\Sigma)}{\det(\Sigma + x x^\top)} \tag{4.14}$$

39

to obtain the following bound on the third term the upper bound (4.13):

$$\sum_{t=1}^{n}(y_t)^2 x_t^\top \Sigma_t^{-1} x_t \le B^2 \sum_{t=1}^{n}\left(1 - \frac{\det(\Sigma_{t-1})}{\det(\Sigma_t)}\right) \le B^2 \ln\frac{\det(\Sigma_n)}{\det(\Sigma_0)} = B^2 \sum_{j=1}^{d} \ln(1 + \lambda_j) \quad (4.15)$$

where λ_i are the eigenvalues of $\sum_{t=1}^{n} x_t x_t^\top$. The sum of eigenvalues cannot be more than nD^2, where $D^2 = \max_{t\in[n]}\|x_t\|^2$, and thus the above upper bound is maximized when all these eigenvalues are equal to nD^2/d. Hence, the third term in (4.13) is upper bounded by $dB^2 \ln(1 + nD^2/d)$. Recalling that $\widehat{y}_t(x_t) = w_t^{x_t} \cdot x_t$ it holds that

$$\frac{1}{n}\sum_{t=1}^{n}(\widehat{y}_t(x_t) - y_t)^2 - \inf_{f\in\mathbb{R}^d}\left\{\frac{1}{n}\sum_{t=1}^{n}(f\cdot x_t - y_t)^2 + \frac{1}{n}\|f\|^2\right\} \le \frac{dB^2\ln(1 + nD^2/d)}{n}$$

$$(4.16)$$

In fact, we derived more than just a regret bound for (4.8) with respect to a set $\mathcal{F}$: the bound holds for all of $\mathbb{R}^d$, appropriately penalizing those f with large norm. Of course, we may pass to the usual regret bound by taking $\mathcal{F}$ to be a bounded set. Typically, a bounded set will be easier to work with, and penalized versions, such as the one above, will be obtained by "stitching together" results for increasing set sizes.

CONCLUSION: In the setting of individual sequence prediction, with no assumptions on the mechanism generating the data, we obtained an $O\left(\frac{d\log n}{n}\right)$ bound on regret, only a factor of $\log n$ worse than those under probabilistic assumptions! Moreover, the proof is direct (the majority of steps are *equalities*), the constants are explicit, and the argument can be easily modified to hold in infinite-dimensional Reproducing Kernel Hilbert spaces with an appropriate decay of the kernel eigenvalues. The proof, however, seems magical and it is unclear why it works. The ideas of using deviations of empirical and expected quantities, as successfully employed in the previous sections, give us a hint of how to approach more complex problems with i.i.d. data. But does the individual sequence proof generalize to other scenarios? A major part of this course is on understanding how such algorithms and upper bounds arise.

From Individual Sequences to I.I.D

The bound (4.16) on regret holds for any sequence $\{(x_t, y_t)\}_{t=1}^{n}$, which can even be adapted on the fly by a malicious adversary. The only requirement is that (x_t, y_t)

is not known to the learner when choosing $\widehat{\boldsymbol{y}}_t$. Now, suppose that the data are actually i.i.d., drawn from some unknown distribution $P_{X \times Y}$. Let

$$\widehat{\boldsymbol{y}} = \frac{1}{n} \sum_{t=1}^{n} \widehat{\boldsymbol{y}}_t \tag{4.17}$$

be the average of the trajectory of the learner. We now claim that $\widehat{\boldsymbol{y}}$ enjoys a guarantee of the type studied in Statistical Learning:

Lemma 4.1. *Suppose* $\mathcal{X} = \mathcal{F} = B_2^d$ *and* $\mathcal{Y} = [-1, 1]$. *The estimator defined in* (4.17) *satisfies*

$$\mathbb{E}(\widehat{\boldsymbol{y}}(X) - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f \cdot X - Y)^2 \le \frac{1 + d\ln(1 + n/d)}{n} \tag{4.18}$$

for any distribution $P_{X \times Y}$.

Proof. To prove this simple fact, take the expectation in (4.16) with respect to the (i.i.d.) data $\{(X_t, Y_t)\}_{t=1}^{n}$:

$$\mathbb{E}\left\{\frac{1}{n} \sum_{t=1}^{n} (\widehat{\boldsymbol{y}}_t(X_t) - Y_t)^2\right\} \le \mathbb{E}\left\{\inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} (f \cdot X_t - Y_t)^2\right\} + \frac{1 + d\ln(1 + n/d)}{n} \tag{4.19}$$

Observe that by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\left\{\mathbb{E}\left\{\left(\widehat{\boldsymbol{y}}(X) - Y\right)^2 \mid X^n, Y^n\right\}\right\} &\le \mathbb{E}\left\{\mathbb{E}\left\{\frac{1}{n} \sum_{t=1}^{n} \left(\widehat{\boldsymbol{y}}_t(X) - Y\right)^2 \mid X^n, Y^n\right\}\right\} \\
&= \mathbb{E}\left\{\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left\{\left(\widehat{\boldsymbol{y}}_t(X) - Y\right)^2 \mid X^{t-1}, Y^{t-1}\right\}\right\} \\
&= \mathbb{E}\left\{\frac{1}{n} \sum_{t=1}^{n} (\widehat{\boldsymbol{y}}_t(X_t) - Y_t)^2\right\} \tag{4.20}
\end{aligned}$$

and

$$\mathbb{E}\left\{\inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} (f \cdot X_t - Y_t)^2\right\} \le \inf_{f \in \mathcal{F}} \mathbb{E}\left\{\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}(f \cdot X_t - Y_t)^2\right\} = \inf_{f \in \mathcal{F}} \mathbb{E}(f \cdot X - Y)^2 \tag{4.21}$$

Putting all the pieces together, we conclude the proof. □

41

DISCUSSION   It is indeed remarkable that the bound for *all* sequences can be easily converted into a bound for the distribution-free Statistical Learning scenario. Such a bound also holds for the well-specified model assumption discussed in the beginning of this lecture. The bounds for the estimator $\widehat{y}$ have an extra $\log n$ factor and require machinery that is markedly different from the one used for analyzing the least squares solution $\widehat{y}$. The former estimator can be viewed as arising from stochastic approximation (SA), while the latter – from sample average approximation (SAA). What is quite curious, the Central Limit Theorem, which played a key role in the first two i.i.d. results, does not seem to appear in the proof of Lemma 4.1! We will show in this course that the CLT is implicitly contained in a regret guarantee. But how is this possible? After all, regret is defined on deterministic sequences! We hope that there is now enough motivation to proceed.

Part II

Theory

5

Minimax Formulation of Learning Problems

Statistical Decision Theory, introduced by A. Wald [56], unified statistical problems – such as estimation and hypothesis testing – under the same umbrella. In the abstract formulation of statistical decision theory, we may think of Nature and the Statistician as playing a zero-sum game. Suppose we have a statistical model, that is a set of possible distributions $\mathcal{P}_0 = \{P^f : f \in \mathcal{F}\}$ on $\mathcal{W}$. We assume that Nature chooses $f \in \mathcal{F}$ and the statistician observes data from the distribution P^f. These data are represented abstractly as a random variable W with distribution P^f (in case of i.i.d. data, think of W as taking values in the product space). Based on W, the Statistician is asked to make a decision – e.g. reject the null hypothesis, compute a particular estimate, or, in the language of the previous lecture, summarize the relationship between X and Y. Let $\mathcal{D}$ bet the set of possible decisions, and let $\widehat{\boldsymbol{y}}$ be either a non-randomized decision function $\widehat{\boldsymbol{y}} : \mathcal{W} \to \mathcal{D}$, or a randomized decision $\widehat{\boldsymbol{y}} : \mathcal{W} \to \Delta(\mathcal{D})$. Fix a loss function $\bar{\boldsymbol{\ell}} : \mathcal{D} \times \mathcal{F} \to \mathbb{R}$. The expected cost (or, *risk*) is

$$\mathbb{E}\bar{\boldsymbol{\ell}}(\widehat{\boldsymbol{y}}(\mathrm{W}), f) \tag{5.1}$$

where the expectation is over the data from the distribution given by Nature's choice of parameter $f \in \mathcal{F}$. This framework is abstract enough to capture a large array of "statistical games".

As an example, consider the realizable or label noise setting introduced in the previous lecture in (3.6). We may set

$$\bar{\boldsymbol{\ell}}\left(\widehat{\boldsymbol{y}}(\mathrm{X}^n, \mathrm{Y}^n), f\right) \triangleq P\left(\widehat{\boldsymbol{y}}(\mathrm{X}) \neq f(\mathrm{X}) \mid \mathrm{X}^n, \mathrm{Y}^n\right),$$

44

with W being the training set (X^n, Y^n). Then (5.1) becomes the expected loss in (3.5).

The games modeled by statistical decision theory can, in fact, be sequential in nature, whereby the statistician makes a sequence of intermediate decisions as well as a decision to terminate the process. We refer to Blackwell and Girshick [10] for a wonderful exposition that brings together the theory of games and statistical decision theory. For another detailed exposition see the book of Berger [7]. In this course, we will be interested in extending (5.1) to the setting of individual sequences.

The introductory lecture was a wind-whirl trip of a dozen topics in learning, estimation, and optimization. We will now formalize some of these problems through the lens of minimax in a way similar to statistical decision theory. However, we will also discuss distribution-free scenarios alongside those based on statistical models. Of special interest is the minimax formulation of sequential prediction with individual sequences, as the blend of ideas from Game Theory and Statistical Learning Theory will yield some new tools.

5.1   Minimax Basics

Let us start by introducing the idea of games and minimax values. We only focus on *zero-sum* games, where the loss of Player I is the reward of Player II and vice-versa. Let $\mathcal{A}$ be the set of moves of Player I and $\mathcal{B}$ the set of available moves of Player II. Let $\boldsymbol{\ell}(a, b)$ be the loss of the first player when she chooses an action $a \in \mathcal{A}$ while the opponent chooses $b \in \mathcal{B}$. Here, $\boldsymbol{\ell} : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ is a known loss function.

As an example, take the penny-matching game, with $\mathcal{A} = \mathcal{B} = \{0, 1\}$ and $\boldsymbol{\ell}(a, b) = \mathbf{I}\{a \neq b\}$. That is, Player I suffers a loss of 1 if her chosen bit is not equal to the one chosen by the opponent. This game has a clear symmetry. If the players make their moves simultaneously, nobody has an advantage, and 1/2 should be the reasonable "expected" loss of either of the players. But how does 1/2 surface? To which question is this an answer?

Let us consider the optimal moves that the players should choose. Since Player I is trying to minimize the loss (and the opponent – maximize), we can write down an expression

$$\min_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} \boldsymbol{\ell}(a, b).$$

Unfortunately, something is wrong with this expression, and it has to do with the

fact that the inner maximization can be done as a function of a. In the penny-matching game, b can always be chosen opposite of a to ensure $\ell(a, b) = 1$. If we instead write $\max_b \min_a$, Player I now has the advantage and guarantees zero loss. This does not match our intuition that the game is symmetric.

But the problem is really with the fact that "$\min \max$" breaks simultaneity of the two moves and instead puts them in order of their occurrence. (This order will play an important role for sequential problems we will be discussing.) So, is there a way to write down simultaneous moves as a $\min \max$? The answer is yes, but under some conditions.

Let Q be the set of distributions on $\mathcal{A}$ and P the set of distributions on $\mathcal{B}$, and let us write the expression

$$V^+ \triangleq \min_{q \in Q} \max_{b \in \mathcal{B}} \mathbb{E}_{a \sim q} \ell(a, b). \tag{5.2}$$

That is, the first player chooses his *mixed strategy q* and tells the second player "hey, you can choose b based on this q". It seems that we have gained nothing: the second player still has the advantage of responding to the first player move, and simultaneity is not preserved. It turns out that, under some conditions (which definitely hold for finite sets $\mathcal{A}$ and $\mathcal{B}$), the above minimax expression is equal to the maximin variant

$$V^- \triangleq \max_{p \in P} \min_{a \in \mathcal{A}} \mathbb{E}_{b \sim p} \ell(a, b). \tag{5.3}$$

Moreover, both of the expressions are equal to $V = \mathbb{E}_{a \sim q^*, b \sim p^*} \ell(a, b)$ for some (q^*, p^*). It becomes irrelevant who makes the first move, as the minimax value is equal to the maximin value! In such a case, we say that the game *has a value* $V = V^+ = V^-$. We will be interested in conditions under which this happens, proving a generalization of the von Neumann Minimax Theorem. The "minimax swap", as we call it, will hold for infinite sets of moves $\mathcal{A}$ and $\mathcal{B}$. To avoid the issue of attainability of min and max, we shall always talk about inf and sup.

Let us mention a few facts which are easy to verify. First, the inner maximization (minimization) is often achieved at a pure strategy if it is followed by an expectation:

$$\min_{q \in Q} \max_{p \in P} \mathbb{E}_{a \sim q, b \sim p} \ell(a, b) = \min_{q \in Q} \max_{b \in \mathcal{B}} \mathbb{E}_{a \sim q} \ell(a, b)$$

and

$$\max_{p \in P} \min_{q \in Q} \mathbb{E}_{a \sim q, b \sim p} \ell(a, b) = \max_{p \in P} \min_{a \in \mathcal{A}} \mathbb{E}_{b \sim p} \ell(a, b)$$

because the minimum (maximum) of a linear functional is achieved at a vertex of the simplex (or set of distributions). The second observation is:

Lemma 5.1. *If $\ell(a, b)$ is convex in a and $\mathcal{A}$ is a convex set, then the outer minimization*

$$\min_{q \in Q} \max_{b \in \mathcal{B}} \mathbb{E}_{a \sim q} \ell(a, b) = \min_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} \ell(a, b)$$

is achieved at a pure strategy.

Proof. We have

$$\min_{a' \in \mathcal{A}} \max_{b \in \mathcal{B}} \ell(a', b) = \min_{q \in Q} \max_{b \in \mathcal{B}} \ell(\mathbb{E}_{a \sim q} a, b) \leq \min_{q \in Q} \max_{b \in \mathcal{B}} \mathbb{E}_{a \sim q} \ell(a, b) \leq \min_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} \ell(a, b) \ .$$

$\square$

As a rather trivial observation, note that *upper bounds* on the minimax value in (5.2) can be obtained by taking a particular choice q for Player I, while *lower bounds* arise from any choice $b \in \mathcal{B}$. Indeed, this is the approach that we take. Throughout the course, we will take the side of Player I, associating her with the Statistician (player, learner), whereas Player II will be Nature (or, adversary, opponent).

Now, let us consider a two-stage game where both players make the first moves $a_1 \in \mathcal{A}$ and $b_1 \in \mathcal{B}$, learn each other's moves, and then proceed to the second round, choosing $a_2 \in \mathcal{A}$ and $b_2 \in \mathcal{B}$. Suppose the payoff is a function $\ell(a_1, b_1, a_2, b_2)$. We can then write down the minimax expression

$$\min_{q_1 \in Q} \max_{b_1 \in \mathcal{B}} \mathbb{E}_{a_1 \sim q_1} \min_{q_2 \in Q} \max_{b_2 \in \mathcal{B}} \mathbb{E}_{a_2 \sim q_2} \ell(a_1, b_1, a_2, b_2) \tag{5.4}$$

We will study such long expressions in great detail throughout the course, so it is good to ponder upon its form for a minute. Notice that the first-stage moves are written first, followed by the second-stage. This is indeed correct: by writing the terms in such an order we allow, for instance, the choice of q_2 to depend on both a_1 and b_1. This corresponds to the protocol that was just described: both players observe each other's moves. In fact, any minimum or maximum in the sequence is calculated with respect to all the variables that have been "chosen" so far. Informally, by looking at the above expression, we can say that when Player II makes the second move, she "knows" the moves a_1 and b_1 at the first time step, as well as the mixed strategy q_1 of Player I at the second step. Based on these,

the maximization over b_2 can be performed. As the expressions in future lectures involve many more than two stages, it is good to go through the sequence and make sure each player "knows" not more and not less than he is supposed to.

Another way to think of the long minimax expression is via operators. To illustrate, $\ell(a_1, b_1, a_2, b_2)$ is a function of 4 variables. The inner expectation over a_2 is an operator (defined by q_2), mapping the function ℓ to a function of 3 variables. The maximization over b_2 maps it to a function of 2 variables, and so forth. When each stage involves the same sequence of operators (e.g. $\min\max\mathbb{E}$), we collapse the long sequence of operators and write

$$\left\langle\!\!\left\langle \min_{q_i \in Q}\max_{b_i \in \mathcal{B}}\mathbb{E}_{a_i \sim q_i} \right\rangle\!\!\right\rangle_{i=1,2} \{\ell(a_1, b_1, a_2, b_2)\} \ .$$

There is yet another way to write the multiple-stage minimax value, such as (5.4). Consider the second move of the two players. For different a_1, b_1, the best moves a_2, b_2 might be different, and so we will write the second-stage choice as a function of a_1, b_1. To avoid the issue of simultaneity, assume that both players choose mixed strategy $q_2 = \pi_2(a_1, b_1)$ and $p_2 = \tau_2(a_1, b_1)$, respectively. The first-stage decisions are trivial constant mappings $q_1 = \pi_1(), p_1 = \tau_1()$, but the later-stage decisions depend on the past. We can write $\pi = (\pi_1, \pi_2)$ and $\tau = (\tau_1, \tau_2)$ as two-stage *strategies* of the players. We can then rewrite (5.4) as

$$\min_{\pi}\max_{\tau}\mathbb{E}_{a_1 \sim \pi_1}\mathbb{E}_{b_1 \sim \tau_1}\mathbb{E}_{a_2 \sim \pi_2(a_1, b_1)}\mathbb{E}_{b_2 \sim \tau_2(a_1, b_1)}\ell(a_1, b_1, a_2, b_2) \ ,$$

or, more succinctly as

$$\min_{\pi}\max_{\tau}\mathbb{E}\ell(a_1, b_1, a_2, b_2) \ . \tag{5.5}$$

We say that the above expression is written in terms of *strategies*, while (5.4) is in the *extensive form*. As we already observed, the maximization over τ will be achieved at a non-randomized strategy, resulting in the simplified stochastic process that evolves according to the randomized choice of Player I and the deterministic choice of Player II. Finally, we remark that π can be thought of as a joint distribution over sequences (a_1, a_2).

5.2   Defining Minimax Values for Learning Problems

In statistical decision theory formulation with risk defined in (5.1), the Statistician's move consists of a decision rule $\widehat{y}$, while Nature's moves are the possible

distributions P^f parametrized by $f \in \mathcal{F}$. The minimax value can be written as

$$\inf_{\widehat{\boldsymbol{y}}} \sup_{f \in \mathcal{F}} \mathbb{E}\bar{\ell}(\widehat{\boldsymbol{y}}(\mathrm{W}), f) \tag{5.6}$$

where the infimum is over all decision rules, and expectation is over $\mathrm{W} \sim P^f$. If ℓ is convex in $\widehat{\boldsymbol{y}}$, the minimization can be taken over deterministic decision rules. This stems from a useful fact that a randomized rule can be represented as a random draw from a distribution over deterministic rules, together with Lemma 5.1.

Statistical Learning Theory

We assume that data are an i.i.d. sample of n pairs $\{(\mathrm{X}_t, \mathrm{Y}_t)\}_{t=1}^{n} \in (\mathcal{X} \times \mathcal{Y})^n$. A *learning algorithm* (or, a *prediction rule*) is a mapping $\widehat{\boldsymbol{y}} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{D}$, where $\mathcal{D} = \mathcal{Y}^{\mathcal{X}}$ is the space of all measurable functions $\mathcal{X} \to \mathcal{Y}$. We either write $\widehat{\boldsymbol{y}}(x; \mathrm{X}^n, \mathrm{Y}^n)$ to make the dependence on data explicit, or simply $\widehat{\boldsymbol{y}}(x)$ if the dependence is understood. Let $\mathcal{P}$ denote the set of *all* distributions on $\mathcal{X} \times \mathcal{Y}$. Consider the case of regression with squared loss. For the distribution-free setting of Statistical Learning Theory, define the minimax value is

$$\mathcal{V}^{iid,sq}(\mathcal{F}, n) \triangleq \inf_{\widehat{\boldsymbol{y}}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}(\widehat{\boldsymbol{y}}(\mathrm{X}) - \mathrm{Y})^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f(\mathrm{X}) - \mathrm{Y})^2 \right\} \tag{5.7}$$

$$= \inf_{\widehat{\boldsymbol{y}}} \sup_{P \in \mathcal{P}, f \in \mathcal{F}} \left\{ \mathbb{E}(\widehat{\boldsymbol{y}}(\mathrm{X}) - \mathrm{Y})^2 - \mathbb{E}(f(\mathrm{X}) - \mathrm{Y})^2 \right\},$$

where the expected value in the first term is over $n + 1$ i.i.d. random variables $(\mathrm{X}_1, \mathrm{Y}_1), \ldots, (\mathrm{X}_n, \mathrm{Y}_n), (\mathrm{X}, \mathrm{Y})$. Note that the supremum ranges over *all* distributions on $\mathcal{X} \times \mathcal{Y}$. The minimax objective $\mathcal{V}^{iid,sq}(\mathcal{F}, n)$ is defined above in terms of *predictive risk* relative to the risk of a reference class $\mathcal{F}$. Alternatively, it can be re-written as follows:

$$\mathcal{V}^{iid,sq}(\mathcal{F}, n) = \inf_{\widehat{\boldsymbol{y}}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}\|\widehat{\boldsymbol{y}} - f_P\|^2 - \inf_{f \in \mathcal{F}} \|f - f_P\|^2 \right\} \tag{5.8}$$

$$= \inf_{\widehat{\boldsymbol{y}}} \sup_{P \in \mathcal{P}, f \in \mathcal{F}} \left\{ \mathbb{E}\|\widehat{\boldsymbol{y}} - f_P\|^2 - \|f - f_P\|^2 \right\}$$

where $f_P(a) = \mathbb{E}[\mathrm{Y}|\mathrm{X} = a]$ is the mean function associated with P, and the norm $\|\cdot\| = \|\cdot\|_{L_2(P_X)}$. Recalling that $\|g\|_{L_2(P_X)}^2 = \int g^2(x) P_X(dx) = \mathbb{E}g^2(\mathrm{X})$, we can easily

verify the equivalence of (5.7) and (5.8). For absolute loss, let us define the analogue of (5.7) as

$$\mathcal{V}^{iid,ab}(\mathcal{F},n) \triangleq \inf_{\widehat{\boldsymbol{y}}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}\left|\widehat{\boldsymbol{y}}(X) - Y\right| - \inf_{f \in \mathcal{F}} \mathbb{E}\left|f(X) - Y\right| \right\} \tag{5.9}$$

and for general losses as

$$\mathcal{V}^{iid}(\mathcal{F},n) \triangleq \inf_{\widehat{\boldsymbol{y}}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}},(X,Y)) - \inf_{f \in \mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f,(X,Y)) \right\} \tag{5.10}$$

Let us now consider the distribution-dependent PAC framework for classification and write down its minimax value:

$$\mathcal{V}^{pac}(\mathcal{F},n) \triangleq \inf_{\widehat{\boldsymbol{y}}} \sup_{P_f} P(\widehat{\boldsymbol{y}}(X) \neq f(X)) = \inf_{\widehat{\boldsymbol{y}}} \sup_{P_f} \mathbb{E}\left|\widehat{\boldsymbol{y}}(X) - f(X)\right| \tag{5.11}$$

where P_f ranges over distributions given by $P_X \times P^f_{Y|X}$ with $P^f_{Y|X=a} = \delta_{f(a)}$ for $f \in \mathcal{F}$, a class of $\{0,1\}$-valued functions. In the label noise scenario, the distribution $P^f_{Y|X=a}$ puts some mass on $1 - f(a)$.

As we go forward, it is important to think of $\mathcal{V}(\mathcal{F},n)$ as a measure of complexity of $\mathcal{F}$. If $\mathcal{F} = \{f\}$ contains only one function, the values defined above are zero since we can simply set $\widehat{\boldsymbol{y}} = f$. On the opposite end of the spectrum, the complexity $\mathcal{V}(\mathcal{Y}^{\mathcal{X}},n)$ of the set of all possible functions is, in general, impossible to make small. One goal of this course is to understand how this value fares for function classes between these two extremes, and to understand what other (easier-to-grasp) measures of complexity of $\mathcal{F}$ are related to $\mathcal{V}(\mathcal{F},n)$.

Nonparametric Regression

We would like to compare the minimax problems studied in nonparametric regression and in statistical learning theory. Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathcal{Y}$. Let P_X denote a marginal distribution on $\mathcal{X}$. For $f \in \mathcal{F}$, let P_f denote the distribution on $\mathcal{X} \times \mathcal{Y}$ obtained as a product of some P_X and a conditional distribution of Y given X $= a$ being a normal $N(f(a),\sigma^2)$ with mean $f(a)$ and variance σ^2. This corresponds to the model Y $= f(X) + \epsilon$, where $\epsilon \sim N(0,\sigma^2)$.

Let $\mathcal{P}_0 = \mathcal{P}_{\mathcal{F}} = \{P_f : f \in \mathcal{F}\} \subset \mathcal{P}$ be a subset of all probability distributions whose regression function $f(a) = \mathbb{E}[Y|X = a]$ is a member of $\mathcal{F}$ and the marginal distribu-

tion P_X is either arbitrary or fixed (if it is part of the design). A goal in nonparametric regression is to study the following minimax value

$$N(\mathcal{F}, n) \triangleq \inf_{\widehat{\boldsymbol{y}}} \sup_{P_f \in \mathcal{P}_\mathcal{F}} \mathbb{E}\|\widehat{\boldsymbol{y}} - f\|^2 \qquad (5.12)$$

for the norm $\|\cdot\| = \|\cdot\|_{L_2(P_X)}$ and some σ. Once again, one can view $N(\mathcal{F}, n)$ as a measure of complexity of the set of distributions parametrized by $\mathcal{F}$, with the two extremes (a single distribution and all possible distributions) having, respectively, complexities of 0 and a positive constant.

Let us show that $N(\mathcal{F}, n) \leq \mathcal{V}^{iid,sq}(\mathcal{F}, n)$, that is, the goal of statistical learning is more difficult than that of nonparametric regression (at least in the particular setting we mentioned). To see this, replace the supremum in (5.8) by the supremum over the smaller class $\mathcal{P}_\mathcal{F} \subset \mathcal{P}$:

$$\mathcal{V}^{iid,sq}(\mathcal{F}, n) \geq \inf_{\widehat{\boldsymbol{y}}} \sup_{P_g \in \mathcal{P}_\mathcal{F}} \left\{ \mathbb{E}\|\widehat{\boldsymbol{y}} - f_{P_g}\|^2 - \inf_{f \in \mathcal{F}} \|f - f_{P_g}\|^2 \right\}$$

$$= \inf_{\widehat{\boldsymbol{y}}} \sup_{P_g \in \mathcal{P}_\mathcal{F}} \mathbb{E}\|\widehat{\boldsymbol{y}} - g\|^2$$

$$= N(\mathcal{F}, n)$$

where it is understood in the first equality that data is distributed according to P_g for some $g \in \mathcal{F}$. The second term is then clearly zero.

A question of interest is to study the gap between $N(\mathcal{F}, n)$ and $\mathcal{V}^{iid,sq}(\mathcal{F}, n)$. The advantage of studying the latter quantity is that tools from empirical process theory (which we will cover in the next few lectures) can be used to get upper bounds.

One can interpolate between the values defined in statistical learning and in nonparametric regression by placing prior knowledge both in the comparator and in the set of possible distributions for nature:

$$\mathcal{V}^{iid,sq}(\mathcal{F}, \mathcal{P}_0, n) \triangleq \inf_{\widehat{\boldsymbol{y}}} \sup_{P \in \mathcal{P}_0} \left\{ \mathbb{E}\|\widehat{\boldsymbol{y}} - f_P\|^2 - \inf_{f \in \mathcal{F}} \|f - f_P\|^2 \right\} \qquad (5.13)$$

where $\mathcal{P}_0 \subseteq \mathcal{P}$ and does not have to necessarily match $\mathcal{F}$.

Universal Data Compression and Conditional Probability Assignment

In the non-sequential minimax settings considered so far, the move of the Statistician is a strategy $\widehat{\boldsymbol{y}}$ is a mapping from data (e.g. (X^n, Y^n)) to the space of functions

$\mathcal{X} \to \mathcal{Y}$. Another way to write this fact is $\widehat{\boldsymbol{y}}(\cdot | X^n, Y^n) : \mathcal{X} \to \mathcal{Y}$. For sequential problems, we talk of strategies as a sequence of mappings $\widehat{\boldsymbol{y}}_t$ which depend on the prefix of data. A good example of this is universal data compression where each $\widehat{\boldsymbol{y}}_t : \mathcal{Z}^{t-1} \to \Delta(\mathcal{Z})$ is a conditional distribution. Specifying the sequence of $\widehat{\boldsymbol{y}}_t$'s is the same as specifying a joint probability distribution $\widehat{\boldsymbol{y}}$ on sequences $(z_1, \ldots, z_n)$. Hence, we can write the minimax value (without the normalization by n) via the n-fold KL divergence as

$$R^+(\mathcal{F}, n) = \inf_{\widehat{\boldsymbol{y}}} \sup_{f \in \mathcal{F}} \mathbb{E} \left\{ \log \frac{f(Z^n)}{\widehat{\boldsymbol{y}}(Z^n)} \right\} = \inf_{\widehat{\boldsymbol{y}}} \sup_{f \in \mathcal{F}} D(f, \widehat{\boldsymbol{y}})$$

where the infimum is over joint distributions and the expectation is over $Z^n = (Z_1, \ldots, Z_n)$ distributed according to $f \in \mathcal{F}$. This minimax value is called the *minimax redundancy* in universal coding. The *maximin redundancy* defined as

$$R^-(\mathcal{F}, n) = \sup_{p \in \Delta(\mathcal{F})} \inf_{\widehat{\boldsymbol{y}}} \mathbb{E}_{f \sim p} D(f, \widehat{\boldsymbol{y}})$$

is related to an information-theoretic notion of a capacity of a channel.

Sequential Prediction of Individual Sequences

In this setting, the move of the learner at time t is based on the knowledge of its own previous moves and the outcomes $z_1, \ldots, z_{t-1}$. Let us use π_t to denote the deterministic mapping $\pi_t : \mathcal{Z}^{t-1} \to \mathcal{D}$, where $\mathcal{D}$ denotes the space of moves of the player. Note that we omitted the dependence on strategy's own past moves: they can be calculated from the moves of the opponent whenever the strategy is deterministic.

We write $\pi = \{\pi_t\}_{t=1}^n$ to denote a n-stage strategy. While $\widehat{\boldsymbol{y}}_t = \pi_t(z^{t-1})$ is the move of the player at time t, we should be thinking of $\pi \in \Pi$ as the single move in the n-stage game. This is the analogue of providing an estimator $\widehat{\boldsymbol{y}}$ for i.i.d. learning. Let Π be the space of all such deterministic strategies and $\Delta(\Pi)$ a distribution on this space (suppose for now that such an object exists). Then, minimax regret against a single comparator and an oblivious adversary can be defined as

$$\mathcal{V}^{obliv}(\mathcal{F}, n) = \inf_{p \in \Delta(\Pi)} \sup_{(z_1, \ldots, z_n) \in \mathcal{Z}^n} \mathbb{E}_{\pi \sim p} \left\{ \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(\pi_t(z^{t-1}), z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(f, z_t) \right\}$$

(5.14)

If $\mathcal{D} = \mathcal{F}$, the set of our moves coincides with the set of moves against which the loss is to be compared. In this case, the algorithm is said to be *proper*, and it is called *improper* if $\mathcal{D} \neq \mathcal{F}$. Improper learning is often useful for computational purposes, though it is not known (at least to us) whether improper learning is strictly more powerful than proper learning in terms of regret bounds. In our notation, $\mathcal{F}$ in $\mathcal{V}(\mathcal{F}, n)$ denotes the comparator term. For improper learning, we will point out the set of moves of the player.

What do we do if sequences are generated by an adaptive (non-oblivious) adversary? We have to define strategies $\tau = \{\tau_t\}_{t=1}^n$ with $\tau_t : \mathcal{D}^{t-1} \to \mathcal{Z}$ for the adversary and interleave it with the strategy of the learner. The minimax value can then be written as

$$\mathcal{V}^{seq}(\mathcal{F}, n) = \inf_{p \in \Delta(\Pi)} \sup_\tau \mathbb{E}_{\pi \sim p} \left\{ \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(\pi_t, \tau_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(f, \tau_t) \right\}$$

where it is understood that π_t and τ_t are functions defined recursively. Unfortunately, the above expression hides too much interdependence which we would like to bring out. We argue that the above value is equal to

$$\mathcal{V}^{seq}(\mathcal{F}, n) = \inf_{q_1 \in \Delta(\mathcal{F})} \sup_{z_1 \in \mathcal{Z}} \underset{\widehat{\boldsymbol{y}}_1 \sim q_1}{\mathbb{E}} \ldots \inf_{q_n \in \Delta(\mathcal{F})} \sup_{z_n \in \mathcal{Z}} \underset{\widehat{\boldsymbol{y}}_n \sim q_n}{\mathbb{E}} \left\{ \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(f, z_t) \right\} .$$
$$(5.15)$$

The fact that these expressions are equal requires a proof, already motivated for $n = 2$ in the earlier section on minimax basics. As discussed previously, our homegrown notation for the long sequence in (5.15) is

$$\mathcal{V}^{seq}(\mathcal{F}, n) = \left\langle\!\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{F})} \sup_{z_t \in \mathcal{Z}} \underset{\widehat{\boldsymbol{y}}_t \sim q_t}{\mathbb{E}} \right\rangle\!\!\!\right\rangle_{t=1}^n \left\{ \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(f, z_t) \right\} \quad (5.16)$$

In particular, for *proper* supervised learning with absolute loss, the value can be written as

$$\mathcal{V}^{seq,ab}(\mathcal{F}, n) = \left\langle\!\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{F})} \sup_{(x_t, y_t)} \underset{\widehat{\boldsymbol{y}}_t \sim q_t}{\mathbb{E}} \right\rangle\!\!\!\right\rangle_{t=1}^n \left\{ \frac{1}{n} \sum_{t=1}^n |\widehat{\boldsymbol{y}}_t(x_t) - y_t| - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n |f(x_t) - y_t| \right\} .$$
$$(5.17)$$

For improper learning,

$$\mathcal{V}^{seq,ab}(\mathcal{F}, n) = \left\langle\!\!\!\left\langle \sup_{x_t} \inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t} \underset{\widehat{\boldsymbol{y}}_t \sim q_t}{\mathbb{E}} \right\rangle\!\!\!\right\rangle_{t=1}^n \left\{ \frac{1}{n} \sum_{t=1}^n |\widehat{\boldsymbol{y}}_t - y_t| - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n |f(x_t) - y_t| \right\} .$$
$$(5.18)$$

The set $\mathcal{F}$ in the last expression refers to the comparator term, and the distributions q_t are over the outcome space $\mathcal{Y}$. To see why this setting is equivalent to (5.17) with an improper choice of $\widehat{\boldsymbol{y}}_t \in \mathcal{Y}^{\mathcal{X}}$, observe that we can decide on $\widehat{\boldsymbol{y}}_t$ for each possible x_t before actually observing it. This amounts to choosing a function $\widehat{\boldsymbol{y}}_t \in \mathcal{Y}^{\mathcal{X}}$ instead of $\mathcal{F}$ as in (5.17).

With an argument identical to that of Lemma 4.1, we see that

$$\mathcal{V}^{iid}(\mathcal{F}, n) \leq \mathcal{V}^{seq}(\mathcal{F}, n)$$

for loss $\boldsymbol{\ell}$ that is convex in the first argument.

Finally, let us mention what would happen to (5.16) if instead of *regret* we considered the sum of our losses without the comparator term:

$$\left\langle\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{F})} \sup_{z_t \in \mathcal{Z}} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^n \left\{ \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) \right\} = \frac{1}{n} \sum_{t=1}^n \inf_{q_t \in \Delta(\mathcal{F})} \sup_{z_t \in \mathcal{Z}} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) = V^+$$

(5.19)

where V^+ is the minimax value (5.2) for the two player game with the payoff $\boldsymbol{\ell}(f, z)$. The first equality comes from the fact that the rounds are completely decoupled. There is no learning to do, as both players will play the optimal move. Not so in the setting of regret minimization: the comparator term forces us to "learn" the strategy of the opponent. It is for this reason that we refer to regret minimization as "learning in games" and study it within the realm of Learning Theory.

Online Convex Optimization

Using the observation that the infimum is achieved at pure strategies when the cost of the learner's decision is a convex function, we can rewrite all the minimax values of the previous general setting. In particular, the oblivious case can be written as

$$\mathcal{V}^{obliv} = \inf_{\pi} \sup_{(z_1, \dots, z_n) \in \mathcal{Z}^n} \left\{ \frac{1}{n} \sum_{t=1}^n z_t(\pi_t(z^{t-1})) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n z_t(f) \right\}$$

where we remind the convention that $\boldsymbol{\ell}(f, z)$ is written as $z(f)$.

The extended form with non-oblivious adversary becomes

$$\mathcal{V}^{oco} = \inf_{\widehat{\boldsymbol{y}}_1 \in \mathcal{F}} \sup_{z_1 \in \mathcal{Z}} \inf_{\widehat{\boldsymbol{y}}_2 \in \mathcal{F}} \sup_{z_2 \in \mathcal{Z}} \dots \inf_{\widehat{\boldsymbol{y}}_n \in \mathcal{F}} \sup_{z_n \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{t=1}^n z_t(\widehat{\boldsymbol{y}}_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n z_t(f) \right\} \qquad (5.20)$$

$$= \left\langle\!\!\left\langle \inf_{\widehat{\boldsymbol{y}}_t \in \mathcal{F}} \sup_{z_t \in \mathcal{Z}} \right\rangle\!\!\right\rangle_{t=1}^n \left\{ \frac{1}{n} \sum_{t=1}^n z_t(\widehat{\boldsymbol{y}}_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n z_t(f) \right\}. \qquad (5.21)$$

5.3   No Free Lunch Theorems

The so-called "No Free Lunch Theorems" say that one cannot achieve uniform (in a certain sense) rates under no assumptions on the problem. We will first demonstrate such a statement in the context of statistical learning theory and nonparametric regression.

5.3.1   Statistical Learning and Nonparametric Regression

As discussed before, the difference between the statistical learning formulation in (5.8) and the nonparametric regression definition in (5.12) is in the way the prior assumption $\mathcal{F}$ appears in the objective. Absence of an assumption corresponds to taking $\mathcal{F} = \mathcal{Y}^{\mathcal{X}}$, the space of all (measurable) functions from $\mathcal{X}$ to $\mathcal{Y}$. In this case,

$$\mathcal{V}^{iid,sq}\left(\mathcal{Y}^{\mathcal{X}}, n\right) = \inf_{\widehat{\boldsymbol{y}}} \sup_{P \in \mathcal{P}} \mathbb{E} \|\widehat{\boldsymbol{y}} - f_P\|^2 \tag{5.22}$$

where the supremum is over all distributions, while

$$N(\mathcal{Y}^{\mathcal{X}}, n) = \inf_{\widehat{\boldsymbol{y}}} \sup_{P_f} \mathbb{E} \|\widehat{\boldsymbol{y}} - f\|^2 \tag{5.23}$$

where P_f ranges over all distributions with an arbitrary mean function $f \in \mathcal{Y}^{\mathcal{X}}$, but Y being distributed as a gaussian with mean $f(X)$. Both values $\mathcal{V}^{iid,sq}(\mathcal{Y}^{\mathcal{X}}, n)$ and $N(\mathcal{Y}^{\mathcal{X}}, n)$ can be lower bounded via a simple argument. Let us focus on the former value, and the latter will follow. We employ the following well-known lower bound, based on binary-valued functions.

Theorem 5.2. *If $|\mathcal{X}| \geq 2n$, it holds that*

$$\mathcal{V}^{iid,sq}\left(\mathcal{Y}^{\mathcal{X}}, n\right) \geq \frac{1}{8} \tag{5.24}$$

Proof. To lower-bound $\mathcal{V}^{iid,sq}(\mathcal{Y}^{\mathcal{X}}, n)$, let us pass to a smaller set of distributions $\mathcal{P}_{\mathcal{F}} \subset \mathcal{P}$, where $\mathcal{F}$ is specified below. The distributions $P^f \in \mathcal{P}_{\mathcal{F}}$ will be defined by a uniform marginal distribution P_X supported on a subset $\mathcal{X}' \subset \mathcal{X}$ of size $|\mathcal{X}'| = 2n$, and the conditional $P^f_{Y|X}$ parametrized by $f \in \mathcal{F}$ via $P^f_{Y|X=a} = \delta_{f(a)}$. Let $\mathcal{F}$ consists of 2^{2n} functions taking on all possible configurations of values $\{0, 1\}$ on $\mathcal{X}'$ and de-

fined as $f(x) = 0$ on $\mathcal{X} \setminus \mathcal{X}'$. Then

$$
\begin{aligned}
\mathcal{V}^{iid,sq}\left(\mathcal{Y}^{\mathcal{X}}, n\right) &\geq \inf_{\widehat{\boldsymbol{y}}} \sup_{P_f \in \mathcal{P}_{\mathcal{F}}} \mathbb{E}(\widehat{\boldsymbol{y}}(\mathrm{X}) - f(\mathrm{X}))^2 \\
&\geq \inf_{\widehat{\boldsymbol{y}}} \mathbb{E}_q \left\{ \mathbb{E}\left\{ (\widehat{\boldsymbol{y}}(\mathrm{X}) - f(\mathrm{X}))^2 \mid f \right\} \right\} \\
&= \inf_{\widehat{\boldsymbol{y}}} \mathbb{E}_{\mathrm{X}^n, \mathrm{X}} \left\{ \mathbb{E}_f \left\{ \left(\widehat{\boldsymbol{y}}\left(\mathrm{X}; \mathrm{X}^n, f(\mathrm{X}^n)\right) - f(\mathrm{X})\right)^2 \mid \mathrm{X}, \mathrm{X}^n \right\} \right\}
\end{aligned}
$$

where q is a uniform distribution on $\mathcal{F}$ and f is distributed according to q. The equality follows by exchanging the order of integration, which is possible because the distribution of X_t's is independent of f. The notation $(\mathrm{X}^n, f(\mathrm{X}^n))$ stands for $\{(\mathrm{X}_t, f(\mathrm{X}_t))\}_{t=1}^n$. Consider the inner conditional expectation and let us further condition on the event $\{\mathrm{X} \notin \mathrm{X}^n\}$ which holds with probability at least $1/2$. For any $f \in \mathcal{F}$ there is $f' \in \mathcal{F}$ which agrees with f on all of $\mathcal{X}'$ except X. However, $\widehat{\boldsymbol{y}}(\mathrm{X}; (\mathrm{X}^n, f(\mathrm{X}^n))) = \widehat{\boldsymbol{y}}(\mathrm{X}; (\mathrm{X}^n, f'(\mathrm{X}^n)))$, and thus the expected loss is at least $1/2$, given that $\mathrm{X} \notin \mathrm{X}^n$. The statement follows. $\qquad\square$

Note that the simple construction of Theorem 5.2 corresponds to the setting of nonparametric regression with a gaussian noise of zero variance. This, of course, makes the regression only easier, and thus the lower bound of Theorem 5.2 is also applicable.

5.3.2   Sequential Prediction with Individual Sequences

We now turn to sequential prediction of individual sequences. The No-Free-Lunch Theorem is a simple variation on the expression (5.19) without the comparator. To start, consider the value defined in (5.18) for the supervised setting with absolute loss. Suppose we remove the prior knowledge of $\mathcal{F}$. That is, as before, $\mathcal{F} = \mathcal{Y}^{\mathcal{X}}$. Suppose for simplicity that $\mathcal{Y} = [-1, 1]$. Assuming x_t's are distinct, which we can do because we are exhibiting a strategy for the adversary in order to obtain a lower bound, we can always find a function that perfectly fits the data. The value then

becomes "comparator-less", just as in (5.19):

$$\mathcal{V}^{seq,ab}(\mathcal{Y}^{\mathcal{X}}, n) = \left\langle\!\!\left\langle \sup_{x_t} \inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \frac{1}{n} \sum_{t=1}^{n} |\widehat{\boldsymbol{y}}_t - y_t| \right\}$$

$$= \left\langle\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \frac{1}{n} \sum_{t=1}^{n} |\widehat{\boldsymbol{y}}_t - y_t| \right\}$$

$$\geq \left\langle\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{Y})} \mathbb{E}_{y_t} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ \frac{1}{n} \sum_{t=1}^{n} |\widehat{\boldsymbol{y}}_t - y_t| \right\}$$

where $y_t \in \{\pm 1\}$ are i.i.d. Rademacher random variables (fair coin flips). Writing $|\widehat{\boldsymbol{y}}_t - y_t| = 1 - \widehat{\boldsymbol{y}}_t \cdot y_t$, exchanging the order of expectations, and unwinding the minimax value

$$\left\langle\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{Y})} \mathbb{E}_{y_t} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left\{ 1 - \frac{1}{n} \sum_{t=1}^{n} \widehat{\boldsymbol{y}}_t y_t \right\} = \left\langle\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{Y})} \mathbb{E}_{y_t} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left\{ \inf_{q_n \in \Delta(\mathcal{Y})} \mathbb{E}_{\widehat{\boldsymbol{y}}_n \sim q_n} \mathbb{E}_{y_n} \left(1 - \frac{1}{n} \sum_{t=1}^{n} \widehat{\boldsymbol{y}}_t y_t \right) \right\}$$

$$= \left\langle\!\!\left\langle \inf_{q_t \in \Delta(\mathcal{Y})} \mathbb{E}_{y_t} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n-1} \left\{ 1 - \frac{1}{n} \sum_{t=1}^{n-1} \widehat{\boldsymbol{y}}_t y_t \right\}$$

we conclude that $\mathcal{V}^{seq,ab}\left(\mathcal{Y}^{\mathcal{X}}, n\right) \geq 1$.

The i.i.d. coin flip strategy y_t for the adversary is an example of an *equalizer* strategy: the move of the learner becomes irrelevant. It turns out many lower bounds in learning theory arise in this way, and we will revisit this idea several times in the course.

6

Learnability, Oracle Inequalities, Model Selection, and the Bias-Variance Trade-off

6.1 Statistical Learning

The No-Free-Lunch Theorem 5.2 for Statistical Learning says that, for any n, the value $\mathcal{V}^{iid,sq}(\mathcal{Y}^{\mathcal{X}}, n)$ cannot be made small. If we think of n as a fixed quantity, this lower bound should be taken seriously: no matter what the learning algorithm $\widehat{\boldsymbol{y}}$ is, there will be a distribution such that the expected loss $\mathbb{E}(\widehat{\boldsymbol{y}}(X) - Y)^2$ of the estimator is much greater than the best achievable over all measurable functions (which, for the case of square loss, is achieved at the regression (or, *Bayes*) function $f_P(a) = \mathbb{E}[Y|X = a]$).

The interpretation of Theorem 5.2 becomes quite murky when n is considered to be variable and increasing to infinity. In this case, the construction of the lower bound is somewhat unsatisfying because the "bad distribution" depends on n. The lower bound says that no matter what our learning rules $\{\widehat{\boldsymbol{y}}_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}\}$ are, for any n there is a distribution on which $\widehat{\boldsymbol{y}}_n$ performs unsatisfactory. One can argue, however, that such a lower bound is overly pessimistic. It might still be possible that for any particular distribution, with enough data the performance of our procedure will be good. Yes, for the given n our estimator might be bad on a particular distribution, but if we eventually get better, then why worry? In other words, the No Free Lunch Theorem does not exclude the possibility that for any particular distribution, we can drive the difference $\mathbb{E}(\widehat{\boldsymbol{y}}_n(X) - Y)^2 - \inf_{f \in \mathcal{Y}^{\mathcal{X}}} \mathbb{E}(f(X) - Y)^2$ to zero.

The subtlety between the two notions of learning is in the order of quantifiers.

Specifically, the distinction boils down to existence of rates that hold *uniformly* for all distributions. To illustrate this, let us first define a shorthand

$$\mathbf{L}(f) \triangleq \mathbb{E}\boldsymbol{\ell}(f, (\mathrm{X}, \mathrm{Y})) \tag{6.1}$$

We have the following two possible definitions of "learnability":

Uniform Consistency

There exists a sequence $\{\widehat{\mathbf{y}}_t\}_{t=1}^{\infty}$ of estimators, such that for any $\epsilon > 0$, there exists n_ϵ such that for any distribution $P \in \mathcal{P}$ and $n \geq n_\epsilon$,

$$\mathbb{E}\mathbf{L}(\widehat{\mathbf{y}}_n) - \inf_{f \in \mathcal{Y}^{\mathcal{X}}} \mathbf{L}(f) \leq \epsilon$$

The smallest value $n_\epsilon = n(\epsilon)$ is called *sample complexity*.

Universal Consistency

There exists a sequence $\{\widehat{\mathbf{y}}_t\}_{t=1}^{\infty}$ of estimators, such that for any distribution $P \in \mathcal{P}$ and any $\epsilon > 0$, there exists n_ϵ such that for $n \geq n_\epsilon$,

$$\mathbb{E}\mathbf{L}(\widehat{\mathbf{y}}_n) - \inf_{f \in \mathcal{Y}^{\mathcal{X}}} \mathbf{L}(f) \leq \epsilon$$

For a given P, the smallest value $n_\epsilon = n(\epsilon, P)$ is called *sample complexity*.

The first notion will be called *uniform consistency*, while the second will be called *universal consistency*.[1] The first can be written as

$$\limsup_{n \to \infty} \inf_{\widehat{\mathbf{y}}_n} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}\mathbf{L}(\widehat{\mathbf{y}}_n) - \inf_{f \in \mathcal{Y}^{\mathcal{X}}} \mathbf{L}(f) \right\} = 0$$

while the second as

$$\inf_{\{\widehat{\mathbf{y}}_t\}_{t=1}^{\infty}} \sup_{P \in \mathcal{P}} \limsup_{n \to \infty} \left\{ \mathbb{E}\mathbf{L}(\widehat{\mathbf{y}}_n) - \inf_{f \in \mathcal{Y}^{\mathcal{X}}} \mathbb{E}\mathbf{L}(f) \right\} = 0$$

The fact that uniform consistency is not possible without restricting the set $\mathcal{P}$ of distributions (as shown in our lower bound of Theorem 5.2) is well known: there is

---

[1]The almost-sure (rather than in-probability) version of this statement is called *universal strong consistency* [21].

no non-trivial uniform rate at which estimators would converge to the regression function for all distributions simultaneously. Lower bounds in the study of universal consistency are generally harder to obtain, and they are called *individual lower rates*. We refer to [24] for a detailed exposition in the regression setting.

Let us illustrate the difference with an example.

Example 1. Let $\mathcal{X} = \mathcal{Y} = [0,1]$ and $\mathcal{P}_0 \subset \mathcal{P}$ the set all distributions on $\mathcal{X} \times \mathcal{Y}$ given by a uniform marginal distribution P_X and $P_{Y|X=a}^f = \delta_{f(a)}$ for some continuous mean function $f : [0,1] \to [0,1]$. Given n samples $\{(X_t, f(X_t))\}_{t=1}^n$, an estimator $\widehat{\boldsymbol{y}}$ is a function $[0,1] \to [0,1]$. No matter what this function is, there is another distribution with a very different mean function f' and the same marginal P_X that passes through the data, yet differs from $\widehat{\boldsymbol{y}}$. Hence, there is no hope for uniform consistency. However, suppose an arbitrary $P \in \mathcal{P}$ is fixed. As we obtain more and more samples, we can approximate the mean function arbitrarily well via kernel or other well-known methods, implying universal consistency.

Uniform and universal consistency can be defined with respect to a smaller class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, in which case the sample complexities are defined, respectively, as $n(\epsilon, \mathcal{F})$ and $n(\epsilon, P, \mathcal{F})$.

We now show that by modifying the comparison yardstick in the definition of $\mathcal{V}^{iid,sq}(\mathcal{Y}^{\mathcal{X}}, n)$, we can pass from uniform to universal consistency. This observation will lead us directly to the ideas of model selection and oracle inequalities. Since

$$\mathcal{V}(\mathcal{Y}^{\mathcal{X}}, n) = \inf_{\widehat{\boldsymbol{y}}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E} \boldsymbol{L}(\widehat{\boldsymbol{y}}_n) - \inf_{f \in \mathcal{Y}^{\mathcal{X}}} \boldsymbol{L}(f) \right\} \tag{6.2}$$

cannot be made small (at least for the square loss) without restricting the set of models, let us redefine the comparator term by making it larger. A particularly interesting modification (which seems rather arbitrary right now, but will be explained in a few moments) is to consider

$$W(\mathcal{F}, n) = \inf_{\widehat{\boldsymbol{y}}_n} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E} \boldsymbol{L}(\widehat{\boldsymbol{y}}_n) - C \inf_k \left[\inf_{f \in \mathcal{F}_k} \boldsymbol{L}(f) + \text{pen}(k, n) \right] \right\} \tag{6.3}$$

where $\mathcal{F} = \cup_{k \geq 1} \mathcal{F}_k$ is either all of $\mathcal{Y}^{\mathcal{X}}$, or a very large set of functions. The subdivision of the large class $\mathcal{F}$ into "manageable" pieces $\{\mathcal{F}_k\}$ is called a *sieve*, and

inequalities of the type

$$\mathbb{E}L(\widehat{\boldsymbol{y}}_n) \le C \inf_k \left[\inf_{f \in \mathcal{F}_k} L(f) + \text{pen}(k, n) \right] \tag{6.4}$$

are called *oracle inequalities*. We assume that $\text{pen}(k, n) \to 0$ as n increases. If $C = 1$, the oracle inequality is called *exact*. It is typically assumed that $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$ and $\text{pen}(k, n)$ is increasing with k.

We may think of $\mathcal{F}_k$ as nested models. The smaller the class, the easier it is to learn, yet the worse is the comparison to the Bayes error. The first term in

$$\left\{ \mathbb{E}L(\widehat{\boldsymbol{y}}) - \inf_{f \in \mathcal{F}_k} L(f) \right\} + \left\{ \inf_{f \in \mathcal{F}_k} L(f) - \inf_{f \in \mathcal{Y}^{\mathcal{X}}} L(f) \right\} \tag{6.5}$$

is known as the *estimation error*, and the second – as the *approximation error*, associated with the choice of the model $\mathcal{F}_k$. This is precisely a bias-variance trade-off. The choice of a good k is known as the *model selection* problem. One may think of choosing among the set $\widehat{\boldsymbol{y}}_n^1, \widehat{\boldsymbol{y}}_n^2, \dots$ of estimators, where $\widehat{\boldsymbol{y}}_n^k$ is the prediction rule associated with $\mathcal{F}_k$.

If $\text{pen} = 0$ and $C = 1$, we get back (6.2), but for a nonzero penalty the goal (6.3) appears more viable, as we subtract a larger value. But what is the meaning of this expression? The idea is that we cannot compete with all functions in the large class $\cup_k \mathcal{F}_k$ with the same uniform rate: some functions are more complex and require more data. This "complexity" is captured in the penalty function $\text{pen}(k, n)$.

We now show that a control on $W(\mathcal{F}, n)$ guarantees universal consistency. Such a result is reassuring because we can focus on obtaining oracle inequalities.

Lemma 6.1. *If it holds that*

$$\limsup_{n \to \infty} W(\mathcal{F}, n) = 0$$

for $W(\mathcal{F}, n)$ defined as an exact oracle ($C = 1$), then there exists a universally consistent estimator $\{\widehat{\boldsymbol{y}}_t\}_{t=1}^{\infty}$.

Proof. Indeed,

$$0 = \limsup_{n \to \infty} W(\mathcal{F}, n) \ge \inf_{\{\widehat{\boldsymbol{y}}_t\}_{t=1}^{\infty}} \sup_{P \in \mathcal{P}} \limsup_{n \to \infty} \left\{ \mathbb{E}L(\widehat{\boldsymbol{y}}_n) - \inf_k \left[\inf_{f \in \mathcal{F}_k} L(f) + \text{pen}(k, n) \right] \right\}$$

The inequality holds by exchanging the order of $\sup_P$ and the lim sup. We can now reason conditionally on P. Assume that the minimizer $f^* = f^*(P) = \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f, (X, Y))$ belongs to some $\mathcal{F}_{k^*}$, where the expected loss is with respect to $(X, Y) \sim P$. Since $\lim_{n \to \infty} \text{pen}(k, n) = 0$, for any $\epsilon > 0$ there exists n_ϵ such that $\text{pen}(k^*, n) \le \epsilon$ for $n \ge n_\epsilon$. For any such $n \ge n_\epsilon$,

$$\mathbb{E}L(\widehat{\boldsymbol{y}}_n) - \inf_k \left[\inf_{f \in \mathcal{F}_k} L(f) + \text{pen}(k, n) \right] \ge \mathbb{E}L(\widehat{\boldsymbol{y}}_n) - \left[\inf_{f \in \mathcal{F}_{k^*}} L(f) + \text{pen}(k^*, n) \right] \ge \mathbb{E}L(\widehat{\boldsymbol{y}}_n) - L(f^*) - \epsilon,$$

and the statement follows. □

Oracle inequalities of the form (6.4) are interesting in their own right, without the implication of universal consistency. An oracle inequality ensures that we are able to do almost as well as if an Oracle told us the model class $\mathcal{F}_k$ to which f^* belongs.

Of course, it remains to be seen whether $W(\mathcal{F}, n)$ defined in (6.3) can be made small, and what kinds of penalties one should use. These penalties can be *distribution-independent*, i.e. of the form $\text{pen}(k, n)$, or *distribution-dependent*, i.e. of the form $\text{pen}(k, n, P)$. What is interesting, a distribution-independent penalty $\text{pen}(k, n)$ should roughly be on the order of the value $\mathcal{V}(\mathcal{F}_k, n)$. While not immediately apparent, it is indeed a good rule of thumb: the penalty should roughly correspond to a measure of "complexity" of $\mathcal{F}_k$. Of course, the penalty can be larger than that since it only makes it easier to show decay of $W(\mathcal{F}, n)$. But a penalty too large leads to less meaningful statements. A penalty too small will make it impossible to ensure smallness of $W(\mathcal{F}, n)$.

Let us sketch an argument showing that the penalty should be at least on the order of the value $\mathcal{V}(\mathcal{F}_k, n)$. Suppose that $\limsup_{n \to \infty} W(\mathcal{F}, n) = 0$, where $C = 1$ and $\text{pen}(k, n) = \mathcal{V}(\mathcal{F}_k, n) - \psi(k, n)$ for some nonnegative ψ. In other words, we are assuming that the penalty is smaller by some amount $\psi(k, n)$ than the difficulty of learning $\mathcal{F}_k$ in a distribution-free manner. We would like to argue that $\psi(k, n)$ cannot be an "interesting" function of k that captures any dependence beyond what is already captured by $\mathcal{V}(\mathcal{F}_k, n)$.

Lemma 6.2. *If it holds that*

$$\limsup_{n \to \infty} W(\mathcal{F}, n) = 0$$

for $C = 1$, and $pen(k, n) = \mathcal{V}(\mathcal{F}_k, n) - \psi(k, n)$ for some nonnegative ψ is a distribution-independent penalty function, then

$$\limsup_{n \to \infty} \sup_k \psi(k, n) \leq 0$$

Proof. Denoting $f_k^* \triangleq \arg\min_{f \in \mathcal{F}_k} L(f)$, we have

$$W(\mathcal{F}, n) = \inf_{\widehat{\boldsymbol{y}}_n} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}\boldsymbol{L}(\widehat{\boldsymbol{y}}_n) - \inf_k \left[\boldsymbol{L}(f_k^*) + \mathcal{V}(\mathcal{F}_k, n) - \psi(k, n) \right] \right\}$$

$$= \inf_{\widehat{\boldsymbol{y}}_n} \sup_k \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}\boldsymbol{L}(\widehat{\boldsymbol{y}}_n) - \boldsymbol{L}(f_k^*) - \inf_{\widehat{\boldsymbol{y}}_n^k} \sup_{P' \in \mathcal{P}} \left[\mathbb{E}\boldsymbol{L}(\widehat{\boldsymbol{y}}_n^k) - \boldsymbol{L}(f_k^*) \right] + \psi(k, n) \right\}$$

$$= \inf_{\widehat{\boldsymbol{y}}_n} \sup_{\widehat{\boldsymbol{y}}_n^k} \sup_k \left[\sup_{P \in \mathcal{P}} \left\{ \mathbb{E}\boldsymbol{L}(\widehat{\boldsymbol{y}}_n) - \boldsymbol{L}(f_k^*) \right\} - \sup_{P' \in \mathcal{P}} \left\{ \mathbb{E}\boldsymbol{L}(\widehat{\boldsymbol{y}}_n^k) - \boldsymbol{L}(f_k^*) \right\} + \psi(k, n) \right]$$

where $\widehat{\boldsymbol{y}}_n^k$ in the definition of $\mathcal{V}(\mathcal{F}_k, n)$ is given the knowledge of k. Clearly, this knowledge should not hurt, so the difference of the two suprema is non-negative. Since the limsup of $W(\mathcal{F}, n)$ is zero, the statement follows. Thus, $\psi(k, n)$ has to have a uniform rate of decay (for each k) in terms of n. Thus, in the asymptotic sense, $\psi(k, n)$ cannot capture any non-trivial dependence on k. □

In this somewhat hand-waving argument, we conclude that distribution-independent penalties used in $W(\mathcal{F}, n)$ should be upper bounds on $\mathcal{V}(\mathcal{F}_k, n)$. Can we show that $W(\mathcal{F}, n)$ is in fact controlled if the penalty is defined as $pen(k, n) = \mathcal{V}(\mathcal{F}_k, n)$? The answer is yes. Suppose that $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$. The particular method we will use is *penalized empirical risk minimization*:

$$\widehat{\boldsymbol{y}} = \arg\min_f \left\{ \frac{1}{n} \sum_{t=1}^n \ell(f, (X_t, Y_t)) + \widetilde{pen}(f) \right\} \tag{6.6}$$

where

$$\widetilde{pen}(f) \triangleq \left\{ \mathcal{V}(\mathcal{F}_k, n) : k = \inf\{k : f \in \mathcal{F}_k\} \right\}.$$

This method guarantees $W(\mathcal{F}, n) \to 0$ under some reasonable assumptions, but we will postpone the discussion until later in the course.

While $pen(k, n)$ is a distribution-independent penalty, one can obtain better bounds with distribution-dependent penalties $pen(k, n, P)$. These, in turn, lead to *data-dependent penalties* $\widehat{pen}$ that can be used in (6.6). There is a large body of literature on such penalties, and a formal understanding of the types of penalties

one can use is an interesting subject. We refer to [4, 37] and [33] for more insights. What is important to note for us is that the tools required for understanding data-dependent penalties in fact arise from the study of each $\mathcal{F}_k$, a "manageable" part of the larger set. These basic tools will be introduced in the next few lectures, and we will return to oracle inequalities towards the end of the course. We will also prove oracle inequalities for the sequential prediction setting with individual sequences.

✎ **Exercise 6.1** (⋆⋆⋆)**.** Suppose $\mathcal{F} = \cup_{i \in \mathbb{N}} \mathcal{F}_i$ is a countable union, and the learning problem with each $\mathcal{F}_i$ is Uniformly Consistent. Prove that the learning problem with respect to $\mathcal{F}$ is Universally Consistent.

6.2   Sequential Prediction

The definitions of Universal and Uniform Consistency in the setting of Statistical Learning have their analogues in the study of Sequential Prediction.

Let regret with respect to $f \in \mathcal{F}$ be defined as

$$\mathbf{Reg}_n(f) = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) - \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{\ell}(f, z_t)$$

Uniform Sequential Consistency

There exists a strategy for the learner such that for any $\epsilon > 0$, there exists n_ϵ such that for any $n \geq n_\epsilon$, irrespective of Nature's strategy, the regret $\mathbf{Reg}_n(f)$ with respect to any $f \in \mathcal{F}$ is at most ϵ in expectation, that is

$$\mathbb{E} \sup_{f \in \mathcal{F}} \mathbf{Reg}_n(f) \leq \epsilon$$

Universal Sequential Consistency

There exists a strategy for the learner such that **for any** $f \in \mathcal{F}$ and any $\epsilon > 0$, there exists n_ϵ such that for $n \geq n_\epsilon$, irrespective of Nature's strategy, the regret $\mathbf{Reg}_n(f)$ is at most ϵ in expectation, that is

$$\mathbb{E}\mathbf{Reg}_n(f) \leq \epsilon$$

We remark that the notion of *Hannan consistency* (as defined, for instance, in [16]), is equivalent to Uniform Sequential Consistency that holds *almost surely* instead of in expectation. For the sake of conciseness, we do not expand on the "almost sure" or "in probability" definitions.

Many of the results of the previous section on model selection and oracle inequalities can be extended in the straightforward way to the case of sequential prediction, and we will touch upon these at the end of the course, time permitting. An exercise on page 163 asks to prove an analogue of the last exercise in the previous section, now in the setting of sequential prediction.

6.3   Remarks

The minimax formulation is a succinct way to represent the problem. Once written down, it is clear what are the sets of moves of the Statistician and Nature, who makes the first move, and whether the distribution is allowed to be changed for each n. Furthermore, it is possible to compare different minimax values with different sets of moves of Nature and Statistician.

Let us mention that the minimax viewpoint is not always the most interesting object of study. In particular, much interest in statistical learning theory is in obtaining *data-dependent* bounds, a key step towards oracle inequalities and model selection discussed above. These would not be possible if we pass to the worst distribution. However, one can argue that the data-dependent bound can be incorporated into the minimax framework. To illustrate this point, let us consider the value defined in (5.9) for i.i.d. supervised learning. Instead of

$$\inf_{\widehat{\boldsymbol{y}}} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, (\mathrm{X}, \mathrm{Y})) - \inf_{f \in \mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f, (\mathrm{X}, \mathrm{Y})) \right\} \tag{6.7}$$

it is quite desirable to get data-dependent bounds of the form

$$\mathbb{E}\left\{ \boldsymbol{\ell}(\widehat{\boldsymbol{y}}, (\mathrm{X}, \mathrm{Y})) \mid \mathrm{X}^n, \mathrm{Y}^n \right\} - \inf_{f \in \mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f, (\mathrm{X}, \mathrm{Y})) \leq \Psi(\mathcal{F}, \mathrm{X}^n, \mathrm{Y}^n)$$

for some function Ψ. While it is often possible to include such a data-dependent bound in a minimax formulation through, for instance, a uniform bound on a ratio of the deviation to Ψ, we will avoid such complications. It is important to note that data-dependent upper bounds can often be isolated as intermediate steps in

proving minimax bounds. This will be the case when we study statistical learning in depth.

7

Stochastic processes, Empirical processes, Martingales, Tree Processes

Learning theory is intimately related to the study of stochastic processes. Statistical learning is concerned with empirical and Rademacher processes, while the study of sequential prediction, as we will see soon, involves a martingale-type processes. Think of this part of the course as a bag of tools we need to study the values of various learning problems introduced earlier.

7.1 Motivation

Let us first motivate the need to study stochastic processes in the settings of statistical learning and sequential prediction.

7.1.1 Statistical Learning

For the setting of statistical learning, the motivation is quite simple. Consider the value $\mathcal{V}^{iid}(\mathcal{F}, n)$ defined in (5.10). Let us take a particular estimator $\widehat{\boldsymbol{y}}$, namely the empirical risk minimizer (ERM)

$$\widehat{\boldsymbol{y}} = \arg\min_{f \in \mathcal{F}} \widehat{\boldsymbol{L}}(f), \quad \text{where} \quad \widehat{\boldsymbol{L}}(f) \triangleq \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{\ell}(f, (X_t, Y_t)).$$

Denoting

$$f^* = \arg\min_{f \in \mathcal{F}} \boldsymbol{L}(f)$$

67

we have for any (X^n, Y^n),

$$L(\widehat{\boldsymbol{y}}) - \inf_{f \in \mathcal{F}} L(f) = \left\{ L(\widehat{\boldsymbol{y}}) - \widehat{L}(\widehat{\boldsymbol{y}}) \right\} + \left\{ \widehat{L}(\widehat{\boldsymbol{y}}) - \widehat{L}(f^*) \right\} + \left\{ \widehat{L}(f^*) - L(f^*) \right\} . \tag{7.1}$$

The second term is negative by the definition of $\widehat{\boldsymbol{y}}$, and the third term is zero in expectation over $(\mathrm{X}^n, \mathrm{Y}^n)$. Hence

$$\mathbb{E} L(\widehat{\boldsymbol{y}}) - \inf_{f \in \mathcal{F}} L(f) \le \mathbb{E} \left\{ L(\widehat{\boldsymbol{y}}) - \widehat{L}(\widehat{\boldsymbol{y}}) \right\} \le \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ L(f) - \widehat{L}(f) \right\} \tag{7.2}$$

As we will see later in the lecture, the last quantity is the expected supremum of an empirical process.

7.1.2   Sequential Prediction

Instead of a full-blown n-round prediction problem, let us consider the two-round game discussed in the previous section. Recall that the players make moves (a_1, b_1) at the first step, and then (a_2, b_2) at the next step. Consider the minimax value defined in (5.4), and suppose it is equal to the maximin value

$$\sup_{p_1 \in P} \inf_{a_1 \in \mathcal{A}} \mathbb{E}_{b_1 \sim p_1} \sup_{p_2 \in P} \inf_{a_2 \in \mathcal{A}} \mathbb{E}_{b_2 \sim p_2} \boldsymbol{\ell}(a_1, b_1, a_2, b_2)$$

Since the object of sequential prediction studied so far is regret, we set

$$\boldsymbol{\ell}(a_1, b_1, a_2, b_2) = \boldsymbol{\ell}(a_1, b_1) + \boldsymbol{\ell}(a_2, b_2) - \inf_{a \in \mathcal{A}} \{ \boldsymbol{\ell}(a, b_1) + \boldsymbol{\ell}(a, b_2) \} .$$

We can now write the value of the two-stage game as

$$\sup_{p_1 \in P} \inf_{a_1 \in \mathcal{A}} \mathbb{E}_{b_1 \sim p_1} \sup_{p_2 \in P} \inf_{a_2 \in \mathcal{A}} \mathbb{E}_{b_2 \sim p_2} \left\{ \boldsymbol{\ell}(a_1, b_1) + \boldsymbol{\ell}(a_2, b_2) - \inf_{a \in \mathcal{A}} \{ \boldsymbol{\ell}(a, b_1) + \boldsymbol{\ell}(a, b_2) \} \right\}$$

$$= \sup_{p_1 \in P} \inf_{a_1 \in \mathcal{A}} \mathbb{E}_{b_1} \left[\boldsymbol{\ell}(a_1, b_1) + \sup_{p_2 \in P} \inf_{a_2 \in \mathcal{A}} \left\{ \mathbb{E}_{b_2} \boldsymbol{\ell}(a_2, b_2) - \mathbb{E}_{b_2} \inf_{a \in \mathcal{A}} \{ \boldsymbol{\ell}(a, b_1) + \boldsymbol{\ell}(a, b_2) \} \right\} \right]$$

$$= \sup_{p_1 \in P} \left[\inf_{a_1 \in \mathcal{A}} \mathbb{E}_{b_1} \boldsymbol{\ell}(a_1, b_1) + \mathbb{E}_{b_1} \sup_{p_2 \in P} \left\{ \inf_{a_2 \in \mathcal{A}} \mathbb{E}_{b_2} \boldsymbol{\ell}(a_2, b_2) - \mathbb{E}_{b_2} \inf_{a \in \mathcal{A}} \{ \boldsymbol{\ell}(a, b_1) + \boldsymbol{\ell}(a, b_2) \} \right\} \right]$$

$$= \sup_{p_1 \in P} \mathbb{E}_{b_1} \sup_{p_2 \in P} \mathbb{E}_{b_2} \left[\inf_{a_1 \in \mathcal{A}} \mathbb{E}_{b_1} \boldsymbol{\ell}(a_1, b_1) + \inf_{a_2 \in \mathcal{A}} \mathbb{E}_{b_2} \boldsymbol{\ell}(a_2, b_2) - \inf_{a \in \mathcal{A}} \{ \boldsymbol{\ell}(a, b_1) + \boldsymbol{\ell}(a, b_2) \} \right]$$

$$\le \sup_{p_1 \in P} \mathbb{E}_{b_1} \sup_{p_2 \in P} \mathbb{E}_{b_2} \sup_{a \in \mathcal{A}} \left[\mathbb{E}_{b_1} \boldsymbol{\ell}(a, b_1) + \mathbb{E}_{b_2} \boldsymbol{\ell}(a, b_2) - \{ \boldsymbol{\ell}(a, b_1) + \boldsymbol{\ell}(a, b_2) \} \right]$$

where in the last step we replaced the infima over a_1 and a_2 with the particular choice a from the third term. The operator $\sup_{p_1 \in P} \mathbb{E}_{b_1} \sup_{p_2 \in P} \mathbb{E}_{b_2}$ can be simply written as $\sup_p \mathbb{E}_{(b_1, b_2) \sim p}$ where p is a joint distribution. Hence, the value of the game is upper bounded by

$$\sup_p \mathbb{E}_{(b_1, b_2)} \sup_{a \in \mathcal{A}} \left\{ \mathbb{E}_{b_1 \sim p_1} \boldsymbol{\ell}(a, b_1) - \boldsymbol{\ell}(a, b_1) + \mathbb{E}_{b_2 \sim p_2(\cdot|b_1)} \boldsymbol{\ell}(a, b_2) - \boldsymbol{\ell}(a, b_2) \right\}, \quad (7.3)$$

and $p_2(\cdot|b_1)$ is the conditional distribution on b_2 given b_1.

We have arrived at an expression which can be recognized as the expected supremum of a stochastic process. Those familiar with the subject observe that the two differences in the last expression form a (rather short) martingale difference sequence. Unlike the statistical learning scenario, the process appears to be non-i.i.d. Next, we precisely define various stochastic processes, and then proceed to study their properties. Equipped with some basic inequalities, we will turn to the study of the suprema of the relevant processes.

Before proceeding, let us agree on some notation. Whenever talking about abstract processes, we will refer to $\mathcal{F}$ as the function class. For learning applications, however, the function class is really $\boldsymbol{\ell}(\mathcal{F}) \triangleq \{\boldsymbol{\ell}(f, \cdot) : f \in \mathcal{F}\}$. We can always switch between the two by thinking of $\boldsymbol{\ell} \circ f = \boldsymbol{\ell}(f, \cdot)$ as our functions on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. For the abstract study of stochastic processes, this distinction will be immaterial.

7.2   Defining Stochastic Processes

Definition 7.1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. A real-valued random variable U is a measurable mapping $\Omega \mapsto \mathbb{R}$. A stochastic process is a collection $\{U_s : s \in \mathcal{S}\}$ of random variables on Ω indexed by a set $\mathcal{S}$.

More generally, we can define $\mathcal{B}$-valued random variables for $\mathcal{B} = \mathbb{R}^d$ or some abstract Banach space $\mathcal{B}$. In such a case, a random variable is a measurable map from $(\Omega, \mathcal{A}, \mathbb{P})$ into $\mathcal{B}$ equipped with Borel σ-algebra generated by the open sets of $\mathcal{B}$ [35].

A stochastic process is defined through its state space (that is, $\mathbb{R}$ or $\mathcal{B}$), the index set $\mathcal{S}$, and the joint distributions of the random variables. If $\mathcal{S}$ is infinite, care should be taken to ensure measurability of events. In this course, we will omit these complications, and assume that necessary conditions hold to ensure measurability.

The first important stochastic process we study is the one that arises from averaging over i.i.d. data:

Definition 7.2. An *empirical process* is a stochastic process $\{\mathbb{G}_f\}$ indexed by a function class $f \in \mathcal{F}$ and defined as

$$\mathbb{G}_f \triangleq \frac{1}{n} \sum_{t=1}^{n} \big(\mathbb{E}f(\mathbb{Z}) - f(\mathbb{Z}_t) \big) = \mathbb{E}(f) - \hat{\mathbb{E}}(f)$$

where $\mathbb{Z}_1, \ldots, \mathbb{Z}_n, \mathbb{Z}$ are i.i.d. (Oftentimes in the literature, the normalization factor is $\frac{1}{\sqrt{n}}$ instead of $\frac{1}{n}$).

Definition 7.3. A random variable ϵ taking on values $\{\pm 1\}$ with equal probability is called a *Rademacher random variable.*[1]

Definition 7.4. Let $\epsilon_1, \ldots, \epsilon_n \in \{\pm 1\}$ be independent Rademacher random variables. A *Rademacher process* is a stochastic process $\{\mathbb{S}_a\}$ indexed by a set $F \subset \mathbb{R}^n$ of vectors $a \in F$ and defined as

$$\mathbb{S}_a \triangleq \frac{1}{n} \sum_{t=1}^{n} \epsilon_t a_t$$

Given $z^n = \{z_1, \ldots, z_n\} \in \mathcal{Z}^n$ and a class $\mathcal{F}$ of functions $\mathcal{Z} \to \mathbb{R}$, we define the Rademacher process on $\mathcal{F}$ as

$$\mathbb{S}_f \triangleq \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(z_t)$$

for $f \in \mathcal{F}$. Since z^n is fixed, we may think of $a = (f(z_1), \ldots, f(z_n))$ as a vector that corresponds to f, matching the earlier definition. From this point of view, the behavior of the functions outside z^n is irrelevant, and we may view the set

$$F = \mathcal{F}|_{(z_1, \ldots, z_n)} = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\} \tag{7.4}$$

as the finite-dimensional *projection* of the function class $\mathcal{F}$ onto z^n.

The processes defined so far are averages of functions of independent random variables. We now bring in the notion of temporal dependence, which will play an important role for the analysis of sequential prediction problems.

---

[1] Did you know that Hans Adolph Rademacher (1892-1969) was a professor here at UPenn?

Definition 7.5. Let $\mathbb{S} = \{0, 1, 2, \ldots, \}$. A stochastic process $\{U_s\}$ is a discrete-time *martingale* if

$$\mathbb{E}\{U_{s+1} \mid U_1, \ldots, U_s\} = U_s$$

and $\mathbb{E}|U_s| < \infty$ for all $s \in \mathbb{S}$. More generally, a stochastic process $\{U_s\}$ is a martingale with respect to another stochastic process $\{V_s\}$ if

$$\mathbb{E}\{U_{s+1} \mid V_1, \ldots, V_s\} = U_s$$

and $\mathbb{E}|U_s| < \infty$. A stochastic process $\{U_s\}$ is a *martingale difference sequence* (MDS) if

$$\mathbb{E}\{U_{s+1} \mid V_1, \ldots, V_s\} = 0$$

for some stochastic process $\{V_s\}$. Any martingale $\{V_s\}$ defines a martingale difference sequence $U_s = V_s - V_{s-1}$.

We now define a "dependent" version of the i.i.d. empirical process.

Definition 7.6. An *empirical process with dependent data* is a stochastic process $\{\mathbb{M}_f\}$ indexed by a function class $f \in \mathcal{F}$ and defined as

$$\mathbb{M}_f \triangleq \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E}\{f(\mathsf{z}_t) \mid \mathsf{z}_1, \ldots, \mathsf{z}_{t-1}\} - f(\mathsf{z}_t) \right)$$

where $(\mathsf{z}_1, \ldots, \mathsf{z}_n)$ is a discrete-time stochastic process with a joint distribution P.

Clearly, the sequence $\left\{ \mathbb{E}\{f(\mathsf{z}_t) \mid \mathsf{z}_1, \ldots, \mathsf{z}_{t-1}\} - f(\mathsf{z}_t) \right\}$ is a martingale-difference sequence for any f. Furthermore, the notion of an empirical process with dependent data boils down to the classical notion if $\mathsf{z}_1, \ldots, \mathsf{z}_n$ are i.i.d.

When specifying martingales, we can talk more generally about filtrations, defined as an increasing sequence of σ-algebras

$$\mathcal{A}_0 \subset \mathcal{A}_1 \subset \ldots \subset \mathcal{A}.$$

A martingale is then defined as a sequence of $\mathcal{A}_s$-measurable random variables U_s such that $\mathbb{E}\{U_{s+1} \mid \mathcal{A}_s\} = U_s$.

Of particular interest is the dyadic filtration $\{\mathcal{A}_t\}$ on $\Omega = \{-1, 1\}^{\mathbb{N}}$ given by $\mathcal{A}_t = \sigma(\epsilon_1, \ldots, \epsilon_t)$, where ϵ_t's are independent Rademacher random variables. Fix Z-valued functions $\mathbf{z}_t : \Omega^{t-1} \to \mathsf{Z}$ for all $t \geq 1$. Then the random variables $\mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1})$

are $\mathcal{A}_{t-1}$-measurable with respect to the dyadic filtration, and the discrete-time stochastic process

$$\left\{ \epsilon_t \mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1}) \right\}$$

is a martingale difference sequence. Indeed,

$$\mathbb{E}\{ \epsilon_t \mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1}) \mid \epsilon_1, \ldots, \epsilon_{t-1} \} = 0 \, .$$

A sequence $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ is called a $\mathcal{Z}$-*valued tree*.

Example 2. To give a bit of intuition about the tree and the associated martingale difference sequence, consider a scenario where we start with a unit amount of money and repeatedly play a fair game. At each stage, we flip a coin and either gain or lose half of our current amount. So, at the first step, we either lose 0.5 or gain 0.5. If we gain 0.5 (for the total of 1.5) the next differential will be ± 0.75. If, however, we lost 0.5 at the first step, the next coin flip will result in a gain or loss of 0.25. It is easy to see that this defines a complete binary tree $\mathbf{z}$. Given any prefix, such as $(1, -1, 1)$, the gain (or loss) $\mathbf{z}_4(1, -1, 1)$ at round 4 is determined. The sum $\sum_{t=1}^{n} \epsilon_t \mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1})$ determines the total payoff.

We may view the martingale $\{U_s\}$ with

$$U_s = \sum_{t=1}^{s} \epsilon_t \mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1})$$

as a random walk with symmetric increments $\pm \mathbf{z}_s$ which depend on the path that got us to this point. Such martingales are known as the *Walsh-Paley martingales*. Interestingly enough, these martingales generated by the Rademacher random variables are, in some sense, "representative" of all the possible martingales with values in $\mathcal{Z}$. We will make this statement precise and use it to our advantage, as these tree-based martingales are much easier to deal with than general martingales.

A word about the notation. For brevity, we shall often write $\mathbf{z}_t(\epsilon)$, where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$, but it is understood that $\mathbf{z}_t$ only depends on the prefix $(\epsilon_1, \ldots, \epsilon_{t-1})$.

Now, given a tree $\mathbf{z}$ and a function $f : \mathcal{Z} \to \mathbb{R}$, we define the composition $f \circ \mathbf{z}$ as a real-valued tree $(f \circ \mathbf{z}_1, \ldots, f \circ \mathbf{z}_n)$. Each $f \circ \mathbf{z}_t$ is a function $\{\pm 1\}^{t-1} \to \mathbb{R}$ and

$$\left\{ \epsilon_t f(\mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1})) \right\}$$

is also a martingale-difference sequence for any given f.

Definition 7.7. Let $\epsilon_1, \ldots, \epsilon_n \in \{\pm 1\}$ be independent Rademacher random variables. Given a tree $\mathbf{z}$, a stochastic process $\{\mathbb{T}_f\}$ defined as

$$\mathbb{T}_f \triangleq \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1}))$$

will be called *tree process* indexed by $\mathcal{F}$.

Example 2, continued Let f be a function that gives the level of excitement of a person observing his fortune U_s going up and down, as in the previous example. If the increment $\mathbf{z}_t$ at the next round is large, the person becomes very happy (large $+f(\mathbf{z}_t)$) upon winning the round, and very unhappy $-f(\mathbf{z}_t)$ upon losing. A person who does not care about the game might have a constant level $f(\mathbf{z}_t) = 0$ throughout the game. On the other extreme, suppose someone becomes agitated when the increments $\mathbf{z}_t$ become close to zero, thus having a large $\pm f(\mathbf{z}_t)$ ups and downs. Suppose $\mathcal{F}$ contains the profiles of a group of people observing the same outcomes. An interesting object of study is the largest cumulative level $\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}(\epsilon_1, \ldots, \epsilon_{t-1}))$ after n rounds.

We may view the tree process $\mathbb{T}_f$ as a generalization of the Rademacher process $\mathbb{S}_f$. Indeed, suppose $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ is a sequence of constant mappings such that $\mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1}) = z_t$ for any $(\epsilon_1, \ldots, \epsilon_{t-1})$. In this case, $\mathbb{T}_f$ and $\mathbb{S}_f$ coincide. Generally, however, the tree process can behave differently (in a certain sense) from the Rademacher process. Understanding the gap in behavior of the two processes will have an implication on the understanding of learnability in the i.i.d. and adversarial models.

7.3   Application to Learning

Turning to the setting of statistical learning theory, we see from (7.2) that the excess loss of the empirical minimizer is upper bounded by

$$\mathbb{E}\boldsymbol{L}(\widehat{\boldsymbol{y}}) - \inf_{f \in \mathcal{F}} \boldsymbol{L}(f) \leq \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \boldsymbol{L}(f) - \widehat{\boldsymbol{L}}(f) \right\} = \mathbb{E} \sup_{g \in \boldsymbol{\ell}(\mathcal{F})} \mathbb{G}_g, \tag{7.5}$$

the expected supremum of the empirical process indexed by the loss class. Thus, to obtain upper bounds on the excess loss $\mathbb{E}\boldsymbol{L}(\widehat{\boldsymbol{y}}) - \inf_{f \in \mathcal{F}} \boldsymbol{L}(f)$ of empirical risk minimizer it is sufficient to obtain upper bounds on the expected supremum of the

empirical process. Furthermore, a *distribution-independent* upper bound on the expected supremum leads to an upper bound on the minimax value.

Theorem 7.8. *Let $\hat{y}$ be the ERM algorithm. Then*

$$\mathbb{E}L(\hat{y}) - \inf_{f \in \mathcal{F}} L(f) \leq \mathbb{E} \sup_{g \in \ell(\mathcal{F})} \mathbb{G}_g, \tag{7.6}$$

and hence

$$\mathcal{V}^{iid}(\mathcal{F}, n) \leq \sup_{P} \left\{ \mathbb{E}L(\hat{y}) - \inf_{f \in \mathcal{F}} L(f) \right\} \leq \sup_{P} \left\{ \mathbb{E} \sup_{g \in \ell(\mathcal{F})} \mathbb{G}_g \right\} \tag{7.7}$$

Now, consider the setting of sequential prediction with individual sequences. The proof given in Section 7.1.2 for $n = 2$ can be readily generalized for any n, and we then arrive at the following upper bound on the value $\mathcal{V}^{seq}(\mathcal{F}, n)$ defined in (5.15) :

Theorem 7.9. *The value of the sequential prediction problem is upper bounded as*

$$\mathcal{V}^{seq}(\mathcal{F}, n) \leq \sup_{P} \left\{ \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E}\left\{ \ell(f, Z_t) \mid Z^{t-1} \right\} - \ell(f, Z_t) \right) \right\} = \sup_{P} \left\{ \mathbb{E} \sup_{g \in \ell(\mathcal{F})} \mathbb{M}_g \right\} \tag{7.8}$$

where the supremum is over all joint distributions P on sequences $(Z_1, \ldots, Z_n)$.

The next key step is to pass from the stochastic processes $\mathbb{G}_g$ and $\mathbb{M}_g$ to the simpler processes $\mathbb{S}_g$ and $\mathbb{T}_g$ which are generated by the Rademacher random variables. The latter two processes turn out to be much more convenient. To make this happen, we appeal to the so-called *symmetrization technique.*

7.4   Symmetrization

We now employ a symmetrization device to pass from the supremum of the empirical process to the supremum of the Rademacher process.

Theorem 7.10. *For a class $\mathcal{F}$ of functions bounded by C, the expected suprema of empirical and Rademacher processes satisfy*

$$\mathbb{E} \sup_{f \in \mathcal{F}} |\mathbb{G}_f| \leq 2\mathbb{E} \sup_{f \in \mathcal{F}} |\mathbb{S}_f|$$

and the same statement holds without absolute values. Furthermore,

$$\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{G}_f| \geq \frac{1}{2}\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{S}_f| - \frac{C}{2\sqrt{n}}$$

For symmetric classes, the statement without absolute values and without the negative term.

Proof. We prove the statement without the absolute values. Observe that

$$\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{G}_f = \mathbb{E}\sup_{f\in\mathcal{F}}\left\{\mathbb{E}f(\mathbb{Z}) - \frac{1}{n}\sum_{t=1}^{n}f(\mathbb{Z}_t)\right\}$$

$$= \mathbb{E}\sup_{f\in\mathcal{F}}\left\{\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}f(\mathbb{Z}_t')\right] - \frac{1}{n}\sum_{t=1}^{n}f(\mathbb{Z}_t)\right\}$$

where the "ghost sample" $\mathbb{Z}_1',\ldots,\mathbb{Z}_n'$ is i.i.d. and all $\mathbb{Z}_t'$'s have the same distribution as $\mathbb{Z}_t$'s. By exchanging supremum and the expectation over the ghost sample, we arrive at an upper bound

$$\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{G}_f \leq \mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}(f(\mathbb{Z}_t') - f(\mathbb{Z}_t))\right\}$$

where the expectation is now over the double sample. And now for the tricky part. Let us define a function h as

$$h(\mathbb{Z}_1,\ldots,\mathbb{Z}_n,\mathbb{Z}_1',\ldots,\mathbb{Z}_n') = \sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}(f(\mathbb{Z}_t') - f(\mathbb{Z}_t))\right\}.$$

Then for any fixed sequence $(\epsilon_1,\ldots,\epsilon_n) \in \{\pm 1\}^n$, the expression

$$\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(f(\mathbb{Z}_t') - f(\mathbb{Z}_t))\right\}$$

is simply a permutation of the coordinates of $h(\mathbb{Z}_1,\ldots,\mathbb{Z}_n,\mathbb{Z}_1',\ldots,\mathbb{Z}_n')$. To illustrate, consider

$$g(1,1,\ldots,1) = h(\mathbb{Z}_1,\mathbb{Z}_2,\ldots,\mathbb{Z}_n,\mathbb{Z}_1',\mathbb{Z}_2',\ldots,\mathbb{Z}_n')$$

while

$$g(-1,1,\ldots,1) = h(\mathbb{Z}_1',\mathbb{Z}_2,\ldots,\mathbb{Z}_n,\mathbb{Z}_1,\mathbb{Z}_2',\ldots,\mathbb{Z}_n').$$

The sequence of signs permutes the respective pairs of coordinates of the function h. But since all the random variables are independent and identically distributed,

any such permutation hardly changes the expectation with respect to the data. Hence,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}(f(\mathsf{Z}_t')-f(\mathsf{Z}_t))\right\}=\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(f(\mathsf{Z}_t')-f(\mathsf{Z}_t))\right\}$$

with the expectation over the double sample and over the Rademacher random variables $\epsilon_1,\dots,\epsilon_n$. We now split the expression into two suprema:

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(f(\mathsf{Z}_t')-f(\mathsf{Z}_t))\right\}\leq\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathsf{Z}_t')\right\}+\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}-\epsilon_t f(\mathsf{Z}_t)\right\}$$

$$=2\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathsf{Z}_t')\right\}$$

because the distributions of $-\epsilon_t$ and ϵ_t are the same. We have arrived at

$$\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{G}_f\leq\mathbb{E}\left\{\mathbb{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathsf{Z}_t)\;\Bigg|\;\mathsf{Z}_1,\dots,\mathsf{Z}_n\right\}\right\}$$

$$=2\mathbb{E}\left\{\mathbb{E}\left\{\sup_{f\in\mathcal{F}}\mathbb{S}_f\;\Bigg|\;\mathsf{Z}_1,\dots,\mathsf{Z}_n\right\}\right\}$$

$$=2\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{S}_f$$

The other direction also holds:

$$\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{S}_f|=\mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\Big(\epsilon_t(f(\mathsf{Z}_t)-\mathbb{E}f(\mathsf{Z}))+\epsilon_t\mathbb{E}f(\mathsf{Z})\Big)\right|\right\}$$

$$\leq\mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(f(\mathsf{Z}_t)-\mathbb{E}f(\mathsf{Z}))\right|\right\}+\mathbb{E}\left|\left(\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\right)\sup_{f\in\mathcal{F}}\mathbb{E}f(\mathsf{Z})\right|,$$

and the last term is upper bounded by $|\sup_{f\in\mathcal{F}}\mathbb{E}f|\sqrt{\frac{1}{n}}$. We now proceed to introduce a ghost sample and then eliminate the random signs in the same way as they were initially introduced.

$$\mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(f(\mathsf{Z}_t)-\mathbb{E}f(\mathsf{Z}))\right|\right\}\leq\mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(f(\mathsf{Z}_t)-f(\mathsf{Z}_t'))\right|\right\}$$

$$=\mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}(f(\mathsf{Z}_t)-f(\mathsf{Z}_t'))\right|\right\}$$

$$\leq 2\mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{t=1}^{n}(\mathbb{E}f(\mathsf{Z})-f(\mathsf{Z}_t))\right|\right\}$$

If the class $\mathcal{F}$ is symmetric, then performing the same steps without the absolute values we see that the term $\mathbb{E}\left\{\left(\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\right)\sup_{f\in\mathcal{F}}\mathbb{E}f(z)\right\} = 0$, and symmetry is invoked in the last step. $\qquad\square$

We now apply the symmetrization technique to the empirical process with dependent data. Note that any such process is an average of martingale difference sequences (MDS).

The symmetrization argument for MDS is more delicate than the one for the i.i.d. case, as swapping z_t and z'_t changes every history for martingale differences with index greater than t. The very same problem has been previously experienced by the characters of "The End of Eternity" (by Isaac Asimov): by traveling into the past and changing it, the future is changed as well. One way to go around this conundrum is to pass to the worst-case future, a pessimistic yet instructive approach. For this purpose, we will perform symmetrization from inside out and pass to the worst-case martingale difference sequence.

Theorem 7.11. *The following relation holds between the empirical process with dependent data and the tree process:*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{M}_f \leq 2\sup_{\mathbf{z}}\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{T}_f \tag{7.9}$$

where the supremum is taken over all $\mathcal{Z}$-valued binary trees of depth n. Furthermore,

$$\frac{1}{2}\left(\sup_{\mathbf{z}}\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{T}_f - \frac{C}{\sqrt{n}}\right) \leq \sup_{\text{MDS}}\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{M}_f \leq 2\sup_{\mathbf{z}}\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{T}_f \tag{7.10}$$

where $C = \sup_{f\in\mathcal{F}}|f|_\infty$.

Proof. By definition,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{M}_f = \mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}\left\{f(z_t)\mid z_1,\ldots,z_{t-1}\right\} - f(z_t)\right) \tag{7.11}$$

Omitting the normalization factor n, let us do the argument for the last time step

n:

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \left(\mathbb{E}\left\{ f(Z_t) \mid Z_1, \ldots, Z_{t-1} \right\} - f(Z_t) \right) + \left(\mathbb{E}\left\{ f(Z_n) \mid Z_1, \ldots, Z_{n-1} \right\} - f(Z_n) \right) \right\}$$

$$\leq \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \left(\mathbb{E}\left\{ f(Z_t) \mid Z_1, \ldots, Z_{t-1} \right\} - f(Z_t) \right) + \left(f(Z_n') - f(Z_n) \right) \right\}$$

$$= \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \left(\mathbb{E}\left\{ f(Z_t) \mid Z_1, \ldots, Z_{t-1} \right\} - f(Z_t) \right) + \epsilon_n \left(f(Z_n') - f(Z_n) \right) \right\}$$

where it is important to keep in mind that Z_n' and Z_n are (conditionally) independent and distributed identically given $Z_1, \ldots, Z_{n-1}$. We now make a very bold step. We upper bound the last quantity by the supremum over Z_n' and Z_n:

$$\mathbb{E} \sup_{z_n, z_n' \in \mathcal{Z}} \mathbb{E}_{\epsilon_n} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-1} \left(\mathbb{E}\left\{ f(Z_t) \mid Z_1, \ldots, Z_{t-1} \right\} - f(Z_t) \right) + \epsilon_n \left(f(z_n') - f(z_n) \right) \right\}$$

Certainly, this is allowed since the expectation can only be smaller. What have we achieved? When we do the same trick for $n-1$, there will be no random variable with a distribution that depends on Z_{n-1}. That is, by passing to the worst-case z_t's we can always perform symmetrization for the previous time step. For $n-2$, we obtain an upper bound of

$$\mathbb{E} \sup_{z_{n-1}, z_{n-1}'} \mathbb{E}_{\epsilon_{n-1}} \sup_{z_n, z_n'} \mathbb{E}_{\epsilon_n} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n-2} \left(\mathbb{E}\left\{ f(Z_t) \mid Z_1, \ldots, Z_{t-1} \right\} - f(Z_t) \right) \right.$$
$$\left. + \epsilon_{n-1} \left(f(z_{n-1}') - f(z_{n-1}) \right) + \epsilon_n \left(f(z_n') - f(z_n) \right) \right\}$$

Proceeding in this manner, we get an upper bound of

$$\sup_{z_1, z_1'} \mathbb{E}_{\epsilon_1} \sup_{z_2, z_2'} \mathbb{E}_{\epsilon_2} \ldots \sup_{z_n, z_n'} \mathbb{E}_{\epsilon_n} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t \left(f(z_t') - f(z_t) \right) \right\} \leq 2 \sup_{z_1} \mathbb{E}_{\epsilon_1} \sup_{z_2} \mathbb{E}_{\epsilon_2} \ldots \sup_{z_n} \mathbb{E}_{\epsilon_n} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t f(z_t) \right\}$$

$$(7.12)$$

"Abandon all hope, ye who enter here", you might say, as we seem to have lost all the randomness of z_t's and replaced them with some worst-case deterministic choices. Let us examine the nth operator $\mathbb{E}_{z_n, z_n'} \mathbb{E}_{\epsilon_n}$, which we replaced with $\sup_{z_n, z_n'} \mathbb{E}_{\epsilon_n}$. Had we instead used $\mathbb{E}_{\epsilon_n} \sup_{z_n, z_n'}$, the resulting value would indeed be too large. However, if the z_t's are chosen before the sign ϵ_n is drawn, we still have the hope that the ϵ_n's are carrying enough "randomness". Indeed, the dyadic

(Walsh-Paley) martingales are "representative" of all the martingales, as this proof shows.

Now, we claim that the right-hand side of (7.12) is nothing but a tree process for the worst-case tree. Indeed, the first supremum is achieved at some $z_1^* \in \mathcal{Z}$. The second supremum is achieved at $z_2^*(+1)$ if $\epsilon_1 = +1$ and at some potentially different value $z_2^*(-1)$ if $\epsilon_1 = -1$. Proceeding in this manner, it is not difficult to see that

$$\sup_{z_1} \mathbb{E}_{\epsilon_1} \sup_{z_2} \mathbb{E}_{\epsilon_2} \ldots \sup_{z_n} \mathbb{E}_{\epsilon_n} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t f(z_t) \right\} = \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_1, \ldots, \epsilon_n} \sup_{f \in \mathcal{F}} \left\{ \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon_1, \ldots, \epsilon_{t-1})) \right\}$$

$$(7.13)$$

$$= n \sup_{\mathbf{z}} \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{T}_f$$

$\square$

> Conclusion of Theorem 7.11: the expected supremum of the worst-case empirical process with dependent data is within a factor of 2 from the expected supremum of the worst-case tree process. That is, when it comes to studying the supremum, the dyadic martingales are representative of all the martingales.

7.5 Rademacher Averages

The expected suprema of the Rademacher and tree processes are so important in our developments that we will give them special names.

Definition 7.12. The expected supremum of a Rademacher process

$$\mathcal{R}^{iid}(\mathcal{F}) \triangleq \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{S}_f = \mathbb{E} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(Z_t) \right\} \tag{7.14}$$

is variably called *Rademacher averages* or *Rademacher complexity* of a class $\mathcal{F}$. Define *conditional Rademacher averages* as

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}; Z_1, \ldots, Z_n) \triangleq \mathbb{E} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(Z_t) \; \middle| \; Z_1, \ldots, Z_n \right\}, \tag{7.15}$$

where the expectation is only with respect to the i.i.d. Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$.

If conditioning on data is understood, we shall omit the word "conditional".

Definition 7.13. The expected supremum of a worst-case tree process

$$\mathcal{R}^{seq}(\mathcal{F}) \triangleq \sup_{\mathbf{z}} \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{T}_f = \sup_{\mathbf{z}} \mathbb{E} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon_{1:t-1})) \right\} \tag{7.16}$$

is called *sequential Rademacher averages* or *sequential Rademacher complexity* of a class $\mathcal{F}$. Define *conditional sequential Rademacher averages* on a given tree $\mathbf{z}$ as

$$\widehat{\mathcal{R}}^{seq}(\mathcal{F}; \mathbf{z}) \triangleq \mathbb{E} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{z}_t(\epsilon_{1:t-1})) \right\} . \tag{7.17}$$

✎ **Exercise 7.1** (⋆)**.** Show that

$$\mathcal{R}^{iid}(\mathcal{F}) \le \sup_{z_1, \dots, z_n} \widehat{\mathcal{R}}^{iid}(\mathcal{F}; z_1, \dots, z_n) \le \mathcal{R}^{seq}(\mathcal{F}) \tag{7.18}$$

✎ **Exercise 7.2** (⋆)**.** Let us make the size n apparent by writing the subscript on $\mathcal{R}_n^{seq}(\mathcal{F})$. Show that $n\mathcal{R}_n^{seq}(\mathcal{F})$ is nondecreasing in n.

✎ **Exercise 7.3** (⋆)**.** Show that $\mathcal{R}_{2n}^{seq}(\mathcal{F}) \le \mathcal{R}_n^{seq}(\mathcal{F})$.

The following properties of Rademacher averages greatly expand the scope of problems that can be studied with the tools introduced in this course. These and further results can be found in Bartlett and Mendelson [5]:

Lemma 7.14. *For any $z_1, \dots, z_n$, conditional Rademacher averages satisfy*

1. *If $\mathcal{F} \subseteq \mathcal{G}$, then $\widehat{\mathcal{R}}^{iid}(\mathcal{F}; z_1, \dots, z_n) \le \widehat{\mathcal{R}}^{iid}(\mathcal{G}; z_1, \dots, z_n)$*

2. *$\widehat{\mathcal{R}}^{iid}(\mathcal{F}; z_1, \dots, z_n) = \widehat{\mathcal{R}}^{iid}(conv(\mathcal{F}); z_1, \dots, z_n)$*

3. *For any $c \in \mathbb{R}$, $\widehat{\mathcal{R}}^{iid}(c\mathcal{F}; z_1, \dots, z_n) = |c|\widehat{\mathcal{R}}^{iid}(\mathcal{F}; z_1, \dots, z_n)$*

For any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n, the above three properties also hold for conditional sequential Rademacher complexity.

✎ **Exercise 7.4** (⋆)**.** Prove Lemma 7.14.

7.6 Skolemization

This is probably a good point to make precise the exchange of expectations and suprema that has taken place several times by now: a) when the interleaved expectations and suprema were collapsed to an expectation over a joint distribution in (7.3), and b) when the sequence of suprema and expectations over random signs became a tree in (7.13). Both of these exchanges follow the same simple (yet very useful) logic. Consider a quantity $\mathbb{E}_a \sup_{b \in \mathcal{B}} \psi(a, b)$ where a is a random variable with values in $\mathcal{A}$. We now claim that

$$\mathbb{E}_a \sup_{b \in \mathcal{B}} \psi(a, b) = \sup_{\gamma} \mathbb{E}_a \psi(a, \gamma(a))$$

where the supremum ranges over all functions $\gamma : \mathcal{A} \to \mathcal{B}$. A simple proof of this statement is left as an exercise.

Whenever faced with a long sequence of expectations and suprema, the trick described above (which we call *skolemization*) is very handy. In particular, the trick gives rise to the joint distribution in Theorem 7.9 and in Eq. (7.3). It also gives rise to the concept of a tree in Eq. (7.13), which is nothing but a sequence of skolemized mappings.

7.7 ... Back to Learning

Putting together Theorem 7.8 and Theorem 7.10, as well as Theorem 7.9 and Theorem 7.11, we get the following two corollaries for learning.

Corollary 7.15. *Let $\widehat{y}$ be the ERM algorithm. Then*

$$\mathbb{E}L(\widehat{y}) - \inf_{f \in \mathcal{F}} L(f) \le \mathbb{E} \sup_{g \in \boldsymbol{\ell}(\mathcal{F})} \mathbb{G}_g \le 2\mathbb{E} \sup_{g \in \boldsymbol{\ell}(\mathcal{F})} \mathbb{S}_g, \tag{7.19}$$

and hence

$$\mathcal{V}^{iid}(\mathcal{F}, n) \le 2 \sup_{P} \mathbb{E} \sup_{g \in \boldsymbol{\ell}(\mathcal{F})} \mathbb{S}_g \tag{7.20}$$

Corollary 7.16. *The value of the sequential prediction problem is upper bounded as*

$$\mathcal{V}^{seq}(\mathcal{F}, n) \le 2 \sup_{\mathbf{z}} \mathbb{E} \sup_{g \in \boldsymbol{\ell}(\mathcal{F})} \mathbb{T}_g \tag{7.21}$$

81

Example: Learning Thresholds

To illustrate the behavior of stochastic processes in conjunction with learning, consider the simplest classification problem possible – thresholds on a unit interval. To this end, let $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$, and

$$\mathcal{F} = \left\{ f_\theta(x) = \mathbf{I}\{x \le \theta\} \; : \; \theta \in [0,1] \right\}.$$

The loss function is the indicator of a mistake: $\ell(f,(x,y)) = \mathbf{I}\{f(x) \ne y\} = |f(x) - y|$.

We now consider the problem of learning thresholds in various scenarios.

8.1 Statistical Learning

Suppose $P_{X \times Y}$ is an arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{D} = \mathcal{Y}^{\mathcal{X}}$. Given the data $\{(X_t, Y_t)\}_{t=1}^n$, we can construct an empirical minimizer $\hat{y} = f_{\hat{\theta}}$ via

$$\hat{\theta} \in \arg\min_{\theta \in [0,1]} \frac{1}{n} \sum_{t=1}^n \left| f_\theta(X_t) - Y_t \right|,$$

a threshold location that incurs the smallest number of mistakes.
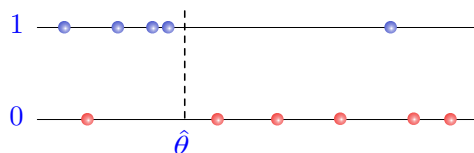
With the tools from the previous section we can almost immediately answer the question: does the excess risk $\mathbb{E}L(\hat{y}) - \inf_{f \in \mathcal{F}} L(f)$ decay as n increases? Along

the way, we will illustrate a couple of techniques that will be useful in the next lecture. First, observe that by Corollary 7.15,

$$\mathbb{E}L(\hat{\boldsymbol{y}}) - \inf_{f \in \mathcal{F}} L(f) \leq \mathbb{E} \sup_{g \in \ell(\mathcal{F})} \mathbb{G}_g \leq 2\mathbb{E} \sup_{g \in \ell(\mathcal{F})} \mathbb{S}_g = 2\mathbb{E} \sup_{\theta \in [0,1]} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t \left| f_\theta(\mathrm{X}_t) - \mathrm{Y}_t \right| \right] \quad (8.1)$$

It is easy to verify that $|a - b| = a(1 - 2b) + b$ for $a, b \in \{0, 1\}$. We can then simplify the supremum of the Rademacher process as

$$\mathbb{E} \sup_{\theta \in [0,1]} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t \left| f_\theta(\mathrm{X}_t) - \mathrm{Y}_t \right| \right] = \mathbb{E} \sup_{\theta \in [0,1]} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t(1 - 2\mathrm{Y}_t) f_\theta(\mathrm{X}_t) + \epsilon_t \mathrm{Y}_t \right] \quad (8.2)$$

$$= \mathbb{E} \sup_{\theta \in [0,1]} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t f_\theta(\mathrm{X}_t) \right] \quad (8.3)$$

The last equality follows by noticing that $\epsilon_t \mathrm{Y}_t$ is zero-mean. Furthermore, conditionally on $\mathrm{Y}_1, \dots, \mathrm{Y}_n$, the distribution of $(1 - 2\mathrm{Y}_t)\epsilon_t$ is again Rademacher and can, therefore, be replaced by ϵ_t.

The last quantity is the supremum of the Rademacher process, but without the Y component and without the loss function:

$$\mathbb{E} \sup_{\theta \in [0,1]} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t | f_\theta(\mathrm{X}_t) - \mathrm{Y}_t | \right] = \mathbb{E} \sup_{\theta \in [0,1]} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t f_\theta(\mathrm{X}_t) \right] \quad (8.4)$$

This step is a precursor of a general technique called a *contraction principle*. Since the class $\mathcal{F}$ of step functions $f_\theta(x) = \mathbf{I}\{x \leq \theta\}$ is somewhat easier to deal with than the loss class $\ell(\mathcal{F})$, it is convenient to "erase" the loss function.

Now, let us pass back to the supremum of the empirical process via Theorem 7.10:

$$\mathbb{E} \sup_{\theta \in [0,1]} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t f_\theta(\mathrm{X}_t) \right] \leq 2\mathbb{E} \sup_{\theta \in [0,1]} \left| \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E} f_\theta(\mathrm{X}) - f_\theta(\mathrm{X}_t) \right) \right| + \frac{1}{\sqrt{n}}$$

Combining all the steps together,

$$\mathbb{E}L(\hat{\boldsymbol{y}}) - \inf_{f \in \mathcal{F}} L(f) \leq 4\mathbb{E} \sup_{\theta \in [0,1]} \left| \mathbb{E}\mathbf{I}\{\mathrm{X} \leq \theta\} - \frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{\mathrm{X}_t \leq \theta\} \right| + \frac{2}{\sqrt{n}} \quad (8.5)$$

This quantity might be familiar. Let $F(\theta) = P(\mathrm{X} \leq \theta)$ be the cumulative distribution function for a random variable X with a marginal distribution P_X. Let $\hat{F}_n(\theta)$ be the

empirical distribution function. By the Law of Large Numbers, $\hat{F}_n(\theta)$ converges to $F(\theta)$ for any given θ. A stronger statement

$$\sup_\theta |\hat{F}_n(\theta) - F(\theta)| \to 0 \quad \text{almost surely}$$

was shown by Glivenko and Cantelli in 1933, and it gives us the desired result: the upper bound in (8.5) converges to zero. Kolmogorov showed that, in fact, $\sup_\theta |\hat{F}_n(\theta) - F(\theta)|$ converges to zero at the rate $n^{-1/2}$, the same rate as that for a single θ. In some sense, we get $\sup_\theta$ for free! This remarkable fact serves as the motivation for the next few lectures.

We conclude that the excess loss $\mathbb{E}L(\hat{y}) - \inf_{f \in \mathcal{F}} L(f)$ of the empirically-best threshold decays at the rate of $n^{-1/2}$ for all distributions $P_{X \times Y}$. Thus, $\mathcal{V}^{iid}(\mathcal{F}, n) = O(n^{-1/2})$.

✏️ **Exercise 8.1** ($\star\star$)**.** Prove that this rate is not improvable if one is allowed to choose any distribution $P_{X \times Y}$.

8.2 Separable (Realizable) Case

Let us describe a non-distribution-free case. One (very strong) assumption we could make is to suppose that $Y_t = f_{\theta^*}(X_t)$ for some threshold $\theta^* \in [0, 1]$. Recall that this assumption lands us right into the PAC framework discussed in the introductory lecture. Such a problem is called "separable" (or, realizable), as the examples labeled with 0 and 1 are on the opposite sides of the threshold. It is not hard to see that the expected loss

$$\mathbb{E}L(\hat{y}) = \mathbb{E}\mathbf{I}\left\{f_{\hat{\theta}}(X) \neq f_{\theta^*}(X)\right\}$$

of the empirically best threshold $\hat{\theta}$ should behave as n^{-1} rather than the previously shown rate of $n^{-1/2}$. To gain intuition, suppose P_X is uniform on $[0, 1]$. Then
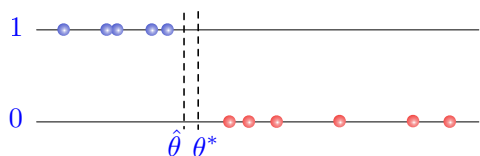

Figure 8.1: Separable case: the ERM solution $\hat{\theta}$ quickly converges to θ^*.

$$\mathbf{I}\left\{f_{\hat{\theta}}(\mathrm{X}) \neq f_{\theta}(\mathrm{X})\right\} = \mathbb{E}|\hat{\theta} - \theta|.$$

Since data are separable, $\hat{\theta}$ can be chosen as the middle of the interval between the right-most value with label 1 and the left-most value with label 0. The distance $|\hat{\theta} - \theta|$ is then at most half the distance between two such extreme values. With n points, the distance should be roughly of the order n^{-1}, up to logarithmic factors which can be removed with a bit more work.

✎ **Exercise 8.2** (⋆)**.** Prove the $O(n^{-1}\log n)$ rate.

8.3   Noise Conditions

Less restrictive assumptions that lead to rates between n^{-1} and $n^{-1/2}$ are assumptions on the noise around the decision boundary. Let $\eta(x) = f_P(x) = \mathbb{E}[\mathrm{Y}|\mathrm{X} = x]$ be the regression function. It is well known (see e.g. [21]) that the best classifier with respect to the indicator loss is $f^* = \mathbf{I}\left\{\eta(x) \geq 1/2\right\}$. Of course, the classifier cannot be computed as P is unknown. Suppose, however, that $f^* = f_{\theta^*} \in \mathcal{F}$. That is, the label 1 is more probable than 0 for $x \leq \theta^*$ and less probable for any $x > \theta^*$. Various assumptions about the behavior of f_P around θ^* translate into intermediate rates, as discussed above. Such conditions are the so-called *Tsybakov's noise condition* and *Massart's noise condition.* Of course, these extend beyond the setting of learning thresholds.
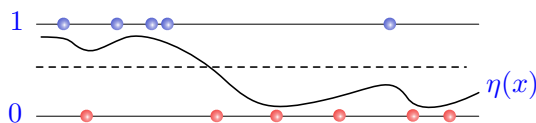
Figure 8.2: In this example, the Bayes classifier $f^* = \mathbf{I}\left\{\eta(x) \geq 1/2\right\}$ is a threshold function. The behavior of $\eta(x)$ around $1/2$ determines the difficulty of the prediction problem.

Importantly, we are not making assumptions about the behavior of the distribution other than through the behavior of f_P at the decision boundary. For the purposes of prediction, it is not important to know the global properties of the underlying distribution, in contrast to the typical assumptions made in statistics.

8.4   Prediction of Individual Sequences

We now consider the same problem in the sequential prediction framework, with no assumption on the sequence $\{(x_t, y_t)\}_{t=1}^{n} \in (\mathcal{X} \times \mathcal{Y})^n$. According to the improper learning protocol, at round t we observe x_t, predict $\widehat{\boldsymbol{y}}_t \in \{0, 1\}$ and observe the label $y_t \in \{0, 1\}$ chosen (simultaneously with $\widehat{\boldsymbol{y}}_t$) by Nature. The question is whether regret

$$\frac{1}{n} \sum_{t=1} \mathbf{I}\{\widehat{\boldsymbol{y}}_t \neq y_t\} - \inf_{\theta \in [0,1]} \frac{1}{n} \sum_{t=1} \mathbf{I}\{f_\theta(x_t) \neq y_t\}$$

can be made small. By Corollary 7.16, the value of the sequential prediction problem is upper bounded as

$$\mathcal{V}^{seq}(\mathcal{F}, n) \leq 2 \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_1, \dots, \epsilon_n} \sup_{\theta \in [0,1]} \mathbb{T}_{\ell(f_\theta)} \tag{8.6}$$

where the supremum is over all $(\mathcal{X} \times \mathcal{Y})$-valued trees $\mathbf{z}$ of depth n. Let us write a $(\mathcal{X} \times \mathcal{Y})$-valued tree equivalently as $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$ and $\mathbf{y}$ are $\mathcal{X}$ and $\mathcal{Y}$-valued. That is, the $[0, 1]$-valued tree in our example has values in the space of covariates, and the $\mathbf{y}$ tree gives labels $\{0, 1\}$ on the corresponding nodes.

We now construct a particular pair $(\mathbf{x}, \mathbf{y})$ which will make the supremum in (8.6) large. Take $\mathbf{y}$ as the tree with all zeros. Define $\mathbf{x}_1 = \frac{1}{2}$, $\mathbf{x}_2(-1) = \frac{1}{4}$, $\mathbf{x}_2(1) = \frac{3}{4}$, $\mathbf{x}_3(-1, -1) = \frac{1}{8}$, $\mathbf{x}_3(-1, 1) = \frac{3}{8}$, $\mathbf{x}_3(1, -1) = \frac{5}{8}$, $\mathbf{x}_3(1, 1) = \frac{7}{8}$ and so forth. The first three levels of the tree are illustrated in figure (8.3).
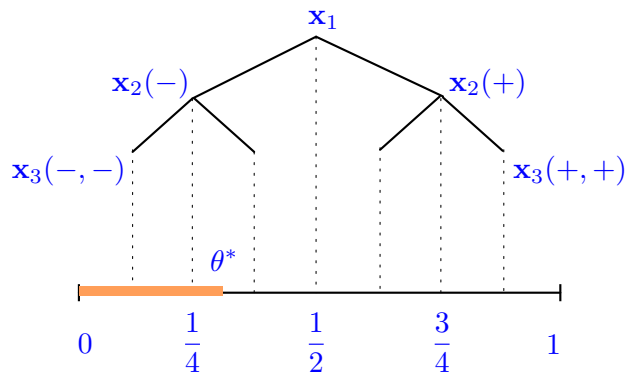

Figure 8.3: Construction of a bad binary tree.

In principle, we can construct an infinite dyadic tree in this manner. Such a tree will play an important role later.

Say, $n = 3$ and the tree $\mathbf{x}$ is exactly as just described. Let $\mathbf{y} = 0$ and take a path, say $(-1, +1, -1)$. For this path,

$$
\begin{aligned}
\sup_{\theta \in [0,1]} \mathbb{T}_{\ell(f_\theta)} &= \sup_{\theta \in [0,1]} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f_\theta(\mathbf{x}_t(\epsilon_{1:t-1})) \\
&= \sup_{\theta \in [0,1]} \frac{1}{3} \Big(-f_\theta(\mathbf{x}_1) + f_\theta(\mathbf{x}_2(-1)) - f_\theta(\mathbf{x}_3(-1, +1)) \Big) \\
&= \sup_{\theta \in [0,1]} \frac{1}{3} \Big(-\mathbf{I}\{\mathbf{x}_1 \leq \theta\} + \mathbf{I}\{\mathbf{x}_2(-1) \leq \theta\} - \mathbf{I}\{\mathbf{x}_3(-1, +1) \leq \theta\} \Big)
\end{aligned}
$$

Observe that there is a θ^* (shown in Figure 8.3) such that the first and third indicators are zero while the second is one. In other words, there is a threshold that annihilates all the indicators with a negative sign in front of them, and keeps the positive ones. This θ^* gives an average of $1/3$ for the supremum, and it is clearly the maximum for the path $(-1, 1, -1)$. Now take the "opposite" path $(1, -1, 1)$. The threshold $1 - \theta^*$ yields $2/3$ for the value of the supremum on this path. It is easy to see that all the paths can be paired in this manner to squeeze the value of $1/2$ on average. We thus have

$$
\mathbb{E}_{\epsilon_1, \ldots, \epsilon_n} \sup_{\theta \in [0,1]} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f_\theta(\mathbf{x}_t(\epsilon_{1:t-1})) = \frac{1}{2}
$$

and conclude that with the constructed tree $\mathbf{x}$ and $\mathbf{y} = 0$, the expected supremum of the tree process does not go to zero.

Theorem 8.1. *For $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$, $\mathcal{F}$ the class of thresholds on $[0,1]$ and the indicator loss,*

$$
\sup_{(\mathbf{x}, \mathbf{y})} \mathbb{E} \left\{ \sup_{\theta \in [0,1]} \mathbb{T}_{\ell(f_\theta)} \right\} \geq \frac{1}{2}
$$

with $\mathbf{x}, \mathbf{y}$ ranging, respectively, over all $\mathcal{X}$ and $\mathcal{Y}$-valued trees.

Why do we care that an upper bound in (8.6) on $\mathcal{V}^{seq}(\mathcal{F}, n)$ does not go to zero with n? Maybe the tree process is an overkill, and it is still possible to play the sequential prediction game, ensuring smallness of $\mathcal{V}^{seq}(\mathcal{F}, n)$? We now show that this is not the case, and the prediction problem is itself hopeless. What is interesting, Nature can use precisely the same tree $\mathbf{x}$ constructed above to ensure the learner incurs n mistakes.

We outline the strategy for Nature. At iteration $t = 1$, $\mathbf{x}_1$ is presented as side information and the learner makes the prediction $\widehat{\mathbf{y}}_1 \in \{0, 1\}$. In the meantime, the Nature flips a fair coin $y_1' \in \{\pm 1\}$ and presents the outcome $y_1 = (y_1' + 1)/2 \in \{0, 1\}$. With probability $1/2$, the prediction $\widehat{\mathbf{y}}_1 \neq y_1$. In the second round, $\mathbf{x}_2(y_1')$ is presented by Nature to the learner who chooses $\widehat{\mathbf{y}}_2$, and the outcome is compared to a fair coin $y_2 \in \{0, 1\}$.

In general, the choice $\mathbf{x}_t(y_1', \ldots, y_{t-1}')$ specifies the strategy of nature. The fair coin flips $y_1', \ldots, y_n'$ make the move of the player irrelevant, as

$$\mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\left\{\widehat{\mathbf{y}}_t \neq y_t\right\}\right\} = \frac{1}{2}.$$

The comparator term, however, is zero, thanks to the structure given by $\mathbf{x}$. Indeed, any time that $y_t' = -1$, the next move is the left child of $\mathbf{x}_t(y_1', \ldots, y_{t-1}')$, while for $y_t' = 1$ it is the right child. Note that any time an element $\mathbf{x}_t$ is followed by the left child, no other $\mathbf{x}_{t'}$ to the right of $\mathbf{x}_t$ will ever be presented, simply given the structure of the tree. By doing so, we are guaranteed that there will be a "consistent" threshold on $\mathbf{x}_t$. It is easy to convince yourself that this strategy ensures existence of a threshold θ^* such that

$$\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\left\{\mathbf{I}\left\{\mathbf{x}_t(y_1', \ldots, y_{t-1}') \leq \theta^*\right\} \neq y_t\right\} = 0 \,.$$

Theorem 8.2. *The class of thresholds on the unit interval is not learnable in the setting of binary prediction of individual sequences:* $\mathcal{V}^{seq}(\mathcal{F}, n) \geq \frac{1}{2}$.

8.5   Discussion

We have considered the problem of learning thresholds in the distribution-free setting of statistical learning, in the setting of PAC learning (the separable case), in the setting of learning with some assumptions on the Bayes function $\mathbb{E}[Y|X = x]$, and in the setting of prediction of individual sequences. In the first scenario, the study of learnability was possible thanks to the Rademacher and Empirical processes. In the last scenario, we showed that the expected supremum of the associated tree process is not decaying to zero, and for a good reason: the prediction problem is not feasible. We will show that in the supervised learning setting convergence of the expected supremum of the tree process takes place if and only if

the associated prediction problem is feasible. This is quite nice, as we do not even need to "play" the prediction game to know whether there is a strategy for predicting well: we can simply study the supremum of the tree process!

In the next lecture, we develop tools to study suprema of stochastic processes. Such statements are called *maximal inequalities*.

9

Maximal Inequalities

9.1 Finite Class Lemmas

At this point, we are hopefully convinced that suprema of various stochastic processes is an object of interest for learning theory. Let us start this endeavor by considering stochastic processes indexed by a finite set.

Lemma 9.1. *Suppose $\{U_s\}_{s\in\mathcal{S}}$ is a finite collection of random variables, and assume that there exists a $c > 0$ such that for all $s \in \mathcal{S}$, $\mathbb{E}\exp(\lambda U_s) \le e^{c^2\lambda^2/2}$ for all $\lambda > 0$. Then*

$$\mathbb{E}\max_{s\in\mathcal{S}} U_s \le c\sqrt{2\log|\mathcal{S}|}\,.$$

Proof. By Jensen's inequality,

$$\exp\left(\lambda\mathbb{E}\max_{s\in\mathcal{S}} U_s\right) \le \mathbb{E}\exp\left(\max_{s\in\mathcal{S}}\lambda U_s\right) = \mathbb{E}\max_{s\in\mathcal{S}}\exp(\lambda U_s)$$
$$\le \mathbb{E}\sum_{s\in\mathcal{S}}\exp(\lambda U_s) = \sum_{s\in\mathcal{S}}\mathbb{E}\exp(\lambda U_s) \le |\mathcal{S}|e^{c^2\lambda^2/2}\,.$$

Taking logarithms of both sides and dividing by λ,

$$\mathbb{E}\max_{s\in\mathcal{S}} U_s \le \frac{\log|\mathcal{S}|}{\lambda} + \frac{c^2\lambda}{2}\,.$$

The proof is concluded by choosing $\lambda = \sqrt{2c^{-2}\log|\mathcal{S}|}$. □

The moment generating condition in Lemma 9.1 should be recognized as a condition on the tail decay of the variables U_s. Variables satisfying such a condition are called *subgaussian*. Examples include gaussian random variables as well as bounded random variables.

Lemma 9.2 (Hoeffding)**.** *For a zero-mean random variable U bounded almost surely as $a \le U \le b$,*

$$\mathbb{E}\exp(\lambda U) \le \exp\left\{\frac{\lambda^2(b-a)^2}{8}\right\} \tag{9.1}$$

The bound of Lemma 9.1 holds for any finite collection of subgaussian random variables. However, the stochastic processes of interest to us are not arbitrary – they have a particular structure. Specifically, all four processes we've discussed (empirical process, Rademacher process, empirical process with dependent data, and the tree process) are all defined as averages of random quantities. We thus expect to obtain more specific statements for the processes of interest. In particular, the typical deviations of averages given by the Central Limit Theorem are $1/\sqrt{n}$, so we expect that c in Lemma 9.1 will be a function of n as well as of magnitudes of the random variables being averaged. Let us make this more precise.

Let us consider the empirical process with dependent data.

Lemma 9.3. *Let $\mathcal{F}$ be a finite class of $[-1,1]$-valued functions on $\mathcal{Z}$. Then*

$$\mathbb{E}\max_{f\in\mathcal{F}}\mathbb{M}_f \le 2\sqrt{\frac{2\log|\mathcal{F}|}{n}}$$

Proof. We need to check the subgaussianity condition and find the appropriate (smallest) constant c. Denote $d_t = \frac{1}{n}\left(\mathbb{E}\left\{f(\mathsf{z}_t)\mid \mathsf{z}_1,\ldots,\mathsf{z}_{t-1}\right\} - f(\mathsf{z}_t)\right)$ and observe that by (9.1)

$$\mathbb{E}\left\{\exp\{\lambda d_t\}\,|\mathsf{z}^{t-1}\right\} \le \exp\left\{\frac{2\lambda^2}{n^2}\right\}$$

because each $d_t \in [-2,2]$. We then have

$$\mathbb{E}\exp\{\lambda\mathbb{M}_f\} = \mathbb{E}\exp\left\{\lambda\sum_{t=1}^{n} d_t\right\} = \mathbb{E}\left[\prod_{t=1}^{n-1}\exp\{\lambda d_t\}\mathbb{E}\left\{\exp\{\lambda d_n\}\,|\mathsf{z}^{n-1}\right\}\right]$$

which is upper bounded by

$$\mathbb{E}\left[\prod_{t=1}^{n-1}\exp\{\lambda d_t\}\right] \times \exp\left\{\frac{2\lambda^2}{n^2}\right\}$$

Repeating the process, we arrive at

$$\mathbb{E}\exp\{\lambda\mathbb{M}_f\} \le \exp\left\{\frac{2\lambda^2}{n}\right\}$$

Appealing to Lemma 9.1 with $c = \frac{2}{\sqrt{n}}$ yields the statement. $\qquad\square$

Of course, the bound of Lemma 9.3 holds for the i.i.d. empirical process $\mathbb{G}_f$ as well. The proofs for the Rademacher and the tree processes are basically identical. Nevertheless, there is an important point to make regarding the boundedness assumption. Since the empirical processes involve random data z_t, we were forced to make a global assumption on the boundedness of $\mathcal{F}$ over $\mathcal{Z}$. Alternatively, we could follow the proof of Lemma 9.3 and replace the global bound on the magnitude of d_t by a conditional variance bound per step. Such results are known and have their merits. For the reasons that will become apparent in the next few lectures, we prefer to deal with the Rademacher and the tree processes, as the magnitudes of d_t are fixed. Indeed, we assume that $z_1, \ldots, z_n$ in the Rademacher process (or the tree $\mathbf{z}$ in the tree process) are fixed. We can then give upper bounds in terms of these fixed quantities.

Let us consider the unnormalized sum to simplify the notation.

Lemma 9.4. *Let $V \subset \mathbb{R}^n$ be a set of N vectors. Then*

$$\mathbb{E} \max_{v \in V} \sum_{t=1}^{n} \epsilon_t v_t \leq \sqrt{2 \log N \max_{v \in V} \sum_{t=1}^{n} v_i^2}$$

Hence, for a finite class $\mathcal{F}$ of functions $\mathcal{Z} \to \mathbb{R}$ and a fixed set $z^n = \{z_1, \ldots, z_n\} \in \mathcal{Z}^n$, it holds that

$$\mathbb{E} \max_{f \in \mathcal{F}} \mathbb{S}_f \leq r \sqrt{\frac{2 \log |\mathcal{F}|}{n}}$$

where $r^2 = \max_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} f^2(z_t)$.

Proof. For a fixed $v \in V$, define $d_t = \epsilon_t v_t$. We then have $\mathbb{E} \exp\{\lambda d_t\} \leq \exp\{(2v_t)^2 \lambda^2/8\}$. Following the proof of Lemma 9.3,

$$\mathbb{E} \exp\left\{\lambda \sum_{t=1}^{n} \epsilon_t v_t\right\} \leq \exp\left\{\left(\sum_{t=1}^{n} v_t^2\right) \lambda^2/2\right\}$$

and the statement follows. $\qquad\qquad\square$

We finish this section with a more powerful lemma that holds for a tree process over a finite index set. Of course, Lemma 9.4 follows from Lemma 9.5 if the trees in the set V are taken to be constant functions $\mathbf{v}_t(\epsilon_{1:t-1}) = v_t$ for all $\epsilon_{1:t-1}$.

Lemma 9.5. *Let V be a set of N real-valued trees of depth n. Then*

$$\mathbb{E} \max_{\mathbf{v} \in V} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon_{1:t-1}) \leq \sqrt{2 \log N \max_{\mathbf{v} \in V} \max_{\epsilon_{1:n}} \sum_{t=1}^{n} \mathbf{v}_t(\epsilon_{1:t-1})^2}$$

Hence, for a finite class $\mathcal{F}$ of functions $\mathcal{Z} \to \mathbb{R}$ and a fixed $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth n, it holds that

$$\mathbb{E} \max_{f \in \mathcal{F}} \mathbb{T}_f \le r \sqrt{\frac{2 \log |\mathcal{F}|}{n}}$$

where

$$r^2 = \max_{f \in \mathcal{F}} \max_{\epsilon_1, \dots, \epsilon_n} \frac{1}{n} \sum_{t=1}^{n} f^2(\mathbf{z}_t(\epsilon_1, \dots, \epsilon_{t-1})).$$

The proof of this lemma is a bit more involved than the previous proofs, as we ask for the upper bound to scale with the largest ℓ_2-norm along *any path in the trees.* To preserve the path structure, we need to peel off the terms in the moment generating function one by one, starting from the last term. We remark that Lemma 9.5 is crucial for the further developments.

Proof. Fix $\lambda > 0$ and a $\mathbb{R}$-valued tree $\mathbf{v} \in V$. For $t \in \{0, \dots, n-1\}$ define a function $A^t : \{\pm 1\}^t \to \mathbb{R}$ by

$$A^t(\epsilon_1, \dots, \epsilon_t) = \max_{\epsilon_{t+1}, \dots, \epsilon_n} \exp\left\{\frac{\lambda^2}{2} \sum_{s=t+1}^{n} \mathbf{v}_s(\epsilon_{1:s-1})^2\right\}$$

and $A^n(\epsilon_1, \dots, \epsilon_n) = 1$. We have that for any $t \in \{1, \dots, n\}$

$$\mathbb{E}_{\epsilon_t} \left\{ \exp\left(\lambda \sum_{s=1}^{t} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times A^t(\epsilon_1, \dots, \epsilon_t) \,\Big|\, \epsilon_1, \dots, \epsilon_{t-1} \right\}$$

$$= \exp\left(\lambda \sum_{s=1}^{t-1} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times \left(\frac{1}{2} e^{\lambda \mathbf{v}_t(\epsilon_{1:t-1})} A^t(\epsilon_1, \dots, \epsilon_{t-1}, +1) + \frac{1}{2} e^{-\lambda \mathbf{v}_t(\epsilon_{1:t-1})} A^t(\epsilon_1, \dots, \epsilon_{t-1}, -1)\right)$$

$$\le \exp\left(\lambda \sum_{s=1}^{t-1} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times \max_{\epsilon_t \in \{\pm 1\}} A^t(\epsilon_1, \dots, \epsilon_t) \left(\frac{1}{2} e^{\lambda \mathbf{v}_t(\epsilon_{1:t-1})} + \frac{1}{2} e^{-\lambda \mathbf{v}_t(\epsilon_{1:t-1})}\right)$$

$$\le \exp\left(\lambda \sum_{s=1}^{t-1} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right) \times A^{t-1}(\epsilon_1, \dots, \epsilon_{t-1})$$

where in the last step we used the inequality $(e^a + e^{-a})/2 \le e^{a^2/2}$. Hence,

$$\mathbb{E}_{\epsilon_1, \dots, \epsilon_n} \left\{ \exp\left(\lambda \sum_{s=1}^{n} \epsilon_s \mathbf{v}_s(\epsilon_{1:s-1})\right)\right\} \le A^0 = \max_{\epsilon_1, \dots, \epsilon_n} \exp\left\{\frac{\lambda^2}{2} \sum_{s=1}^{n} \mathbf{v}_s(\epsilon_{1:s-1})^2\right\} .$$

We conclude that

$$\exp\left(\lambda \mathbb{E} \max_{\mathbf{v} \in V} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon_{1:t-1})\right) \le N \max_{\mathbf{v} \in V} \max_{\epsilon_1, \dots, \epsilon_n} \exp\left\{\frac{\lambda^2}{2} \sum_{s=1}^{n} \mathbf{v}_s(\epsilon_{1:s-1})^2\right\}$$

and the rest of the proof follows the proof of Lemma 9.1. $\qquad\square$

10

Example: Linear Classes

We say that $\mathcal{F}$ is a linear function class if each $f(x) = \langle f, x \rangle$ is linear in x. For finite-dimensional problems we may think of f as a vector. For $p \geq 0$, define

$$\mathsf{B}_p^d \triangleq \left\{ a \in \mathbb{R}^d : \|a\|_p \leq 1 \right\},$$

the unit ball in $\mathbb{R}^d$ with respect to the p-norm. It is well-known that for the conjugate pair $1 \leq p, q \leq \infty$ with $p^{-1} + q^{-1} = 1$,

$$\|a\|_p = \sup_{b \in \mathsf{B}_q^d} \langle a, b \rangle .$$

That is, L_p and L_q norms are *dual* to each other. Hölder's inequality then says that

$$\langle a, b \rangle \leq \|a\|_p \cdot \|b\|_q .$$

Example 3 (L_2/L_2 case)**.** Let $\mathcal{F} = \mathcal{X} = \mathsf{B}_2^d$. For any $\mathcal{X}$-valued tree $\mathbf{x}$ of depth n,

$$\widehat{\mathcal{R}}^{seq}(\mathcal{F}; \mathbf{x}) = \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t \langle f, \mathbf{x}_t(\epsilon) \rangle \right\} = \mathbb{E} \sup_{f \in \mathcal{F}} \left\langle f, \frac{1}{n} \sum_{t=1}^n \epsilon_t \mathbf{x}_t(\epsilon) \right\rangle = \mathbb{E} \left\| \frac{1}{n} \sum_{t=1}^n \epsilon_t \mathbf{x}_t(\epsilon) \right\|_2$$

$$(10.1)$$

One can view the sequential Rademacher complexity as the expected length of a random walk of the martingale with increments $\{\epsilon_t \mathbf{x}_t(\epsilon_{1:t-1})\}$, normalized by n. Recall that $\mathbf{x}$ is $\mathcal{X}$-valued, so the increments are in the Euclidean unit ball. So, how far can such a random walk be expected to walk away for any tree $\mathbf{x}$? An easy calculation shows that

$$\mathbb{E} \left\| \sum_{t=1}^n \epsilon_t \mathbf{x}_t(\epsilon) \right\|_2 \leq \left(\mathbb{E} \left\| \sum_{t=1}^n \epsilon_t \mathbf{x}_t(\epsilon) \right\|_2^2 \right)^{1/2} = \left(\mathbb{E} \sum_{t,s} \langle \epsilon_t \mathbf{x}_t(\epsilon), \epsilon_s \mathbf{x}_s(\epsilon) \rangle \right)^{1/2} = \left(\sum_{t=1}^n \mathbb{E} \|\mathbf{x}_t(\epsilon)\|_2^2 \right)^{1/2} \leq \sqrt{n}$$

In view of (7.18) we conclude that

$$\mathcal{R}^{iid}(\mathcal{F}) \le \mathcal{R}^{seq}(\mathcal{F}) \le \frac{1}{\sqrt{n}} \tag{10.2}$$

Khinchine-Kahane inequality shows that $\mathcal{R}^{iid}(\mathcal{F}) \ge \frac{1}{\sqrt{2n}}$ for any symmetric distribution P on the surface of $\mathcal{X}$, so $\sup_P \mathcal{R}^{iid}(\mathcal{F})$ is within a constant factor from $\mathcal{R}^{seq}(\mathcal{F})$.

Example 4 (L_1/L_∞ case)**.** Suppose now that $\mathcal{X} = \mathsf{B}_\infty^d$, while $\mathcal{F} = \mathsf{B}_1^d$. Observe that

$$\mathcal{F} = \text{conv}\left(\{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_d\}\right).$$

That is, the ℓ_1 ball is a convex hull of $2d$ vertices. Thus,

$$\mathcal{R}^{iid}(\mathcal{F}) \le \mathcal{R}^{seq}(\mathcal{F}) = \mathcal{R}^{seq}(\{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_d\}) \le \sqrt{\frac{2\log(2d)}{n}} \tag{10.3}$$

by Lemma 9.4.

Example 5 (Δ_d/L_∞ case)**.** Suppose now that $\mathcal{X} = \mathsf{B}_\infty^d$ and

$$\mathcal{F} = \Delta_d = \left\{ f \in \mathbb{R}^d : \sum_{i=1}^d f_i = 1, \ f_i \ge 0 \ \forall i \right\}$$

is the d-simplex. Similarly to the previous example, $\mathcal{F} = \text{conv}\left(\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}\right)$, and

$$\mathcal{R}^{iid}(\mathcal{F}) \le \mathcal{R}^{seq}(\mathcal{F}) = \mathcal{R}^{seq}(\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}) \le \sqrt{\frac{2\log(d)}{n}} \tag{10.4}$$

by Lemma 9.4.

Example 6 (General case)**.** $\mathcal{X}$ be a unit ball in a separable Banach space $(\mathcal{B}, \|\cdot\|)$. Consider the dual space—the space of continuous linear functionals on $\mathcal{B}$. Let $\|\cdot\|_*$ be the dual norm, defined for an element $f \in \mathcal{B}^*$ of the dual space by

$$\|f\|_* = \sup_{x \in \mathcal{X}} \langle f, x \rangle . \tag{10.5}$$

Let Ψ^* be a σ-*strongly convex* function with respect to $\|\cdot\|_*$ on $\mathcal{F}$. That is,

$$\forall f, g \in \mathcal{F}, \quad \Psi^*(f) \ge \Psi^*(g) + \langle f - g, \nabla\Psi^*(g) \rangle + \frac{\sigma}{2}\|f - g\|_*^2 \tag{10.6}$$

95

Defining the *convex conjugate* Ψ of Ψ^* as

$$\Psi(x) = \sup_{f \in \mathcal{F}} \langle f, x \rangle - \Psi^*(f),$$

it is possible to verify the opposite property (smoothness) for the conjugate function:

$$\forall x, y \in \mathcal{X}, \quad \Psi(x) \leq \Psi(y) + \langle \nabla \Psi(y), x - y \rangle + \frac{1}{2\sigma} \|x - y\|^2. \tag{10.7}$$

Let $M^2 = \sup_{f \in \mathcal{F}} \Psi^*(f)$. Using the definition of conjugacy, for any $\lambda > 0$,

$$\widehat{\mathcal{R}}^{seq}(\mathcal{F}; \mathbf{x}) = \frac{1}{\lambda} \mathbb{E} \sup_{f \in \mathcal{F}} \left\langle f, \frac{\lambda}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon) \right\rangle \leq \frac{1}{\lambda} \left(\sup_{f \in \mathcal{F}} \Psi^*(f) + \mathbb{E} \Psi \left(\frac{\lambda}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t(\epsilon) \right) \right) \tag{10.8}$$

The first term is upper bounded by M^2, while for the second term we use (10.7):

$$\mathbb{E} \Psi(Z_n) \leq \mathbb{E} \left(\Psi(Z_{n-1}) + \left\langle \nabla \Psi(Z_{n-1}), \frac{\lambda}{n} \epsilon_n \mathbf{x}_n(\epsilon) \right\rangle + \frac{1}{2\sigma} \left\| \frac{\lambda}{n} \mathbf{x}_n(\epsilon) \right\|^2 \right) \tag{10.9}$$

with $Z_k = \frac{\lambda}{n} \sum_{t=1}^{k} \epsilon_t \mathbf{x}_t(\epsilon)$. The first-order term disappears under the expectation, and the second-order term is bounded by $\lambda^2/(2\sigma n^2)$ since the $\mathbf{x}$ tree is $\mathcal{X}$-valued. Peeling off all the terms in the sum in the similar manner, we arrive at

$$\widehat{\mathcal{R}}^{seq}(\mathcal{F}; \mathbf{x}) \leq \frac{M^2}{\lambda} + \frac{\lambda}{2\sigma n} = M \sqrt{\frac{2}{\sigma n}} \tag{10.10}$$

for $\lambda = M\sqrt{2\sigma n}$. Of course, the radius of the ball $\mathcal{X}$ is 1 and does not appear in the bound; otherwise, the bound would scale linearly with it.

<div style="text-align: right; font-size: 3em; font-style: italic;">11</div>

Statistical Learning: Classification

We are now armed with a powerful tool: an upper bound on the expected supremum of $\mathbb{G}_f, \mathbb{S}_f, \mathbb{M}_f$, and $\mathbb{T}_f$ indexed by a finite set $\mathcal{F}$. How about an infinite class $\mathcal{F}$? In this lecture we address this question for classification problems within the scope of Statistical Learning.

11.1 From Finite to Infinite Classes: First Attempt

The first idea that comes to mind is to approximate the class by a finite "representative" set. The set will be called a "cover". Intuitively, the larger the set, the better is the approximation, but the worse is the log-size-of-finite-set bound.

As the first attempt, let us implement the idea of a cover to upper bound $\mathbb{G}_f$ for the simple case of thresholds in one dimension, studied in Section 8.1. Effectively, we are aiming at an upper bound on

$$\mathbb{E} \sup_{\theta \in [0,1]} \left[\mathbb{E}\mathbf{I}\{X \le \theta\} - \frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{X_t \le \theta\} \right] \tag{11.1}$$

(see Eq. (8.5)) which was graciously provided to us by Kolmogorov. Recall that there is an unknown underlying probability distribution $P_{X \times Y}$ on which we place no restriction. Let $\Theta_N = \{\theta_1, \ldots, \theta_N\}$ be equally-spaced on the interval $[0, 1]$ and let $c(\theta) \in \Theta_N$ be the element of the representative set closest to θ. With the notation

$$\mathbb{G}_\theta = \mathbb{E}\mathbf{I}\{X \le \theta\} - \frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{X_t \le \theta\}$$

<div style="text-align: center;">97</div>

we can then write

$$\mathbb{E} \sup_{\theta \in [0,1]} [\mathbb{G}_\theta] = \mathbb{E} \sup_{\theta \in [0,1]} \left[\mathbb{G}_{c(\theta)} + \mathbb{G}_\theta - \mathbb{G}_{c(\theta)} \right] \leq \mathbb{E} \sup_{\theta \in [0,1]} \left[\mathbb{G}_{c(\theta)} \right] + \mathbb{E} \sup_{\theta \in [0,1]} \left[\mathbb{G}_\theta - \mathbb{G}_{c(\theta)} \right]$$

(11.2)

The first term is the expected supremum over the finite set Θ_N:

$$\mathbb{E} \sup_{\theta \in [0,1]} \left[\mathbb{G}_{c(\theta)} \right] = \mathbb{E} \max_{\theta_i \in \Theta_N} \left[\mathbb{G}_{\theta_i} \right]$$

but how do we control the second term? Since θ and $c(\theta)$ are close, we are tempted to say that it is small. However, we made no assumption on P_X, so it is very well possible that all of its mass is concentrated on some interval $[\theta_i, \theta_{i+1}]$, in which case the above expression is equal to the one in (11.1). Hence, we have achieved nothing by passing to the finite subset! Clearly, the discretization needs to depend on the unknown distribution P_X.

To rescue the situation, we may try to place the elements $\theta_1, \ldots, \theta_N$ such that $\mathbb{E}\mathbf{I}\{\theta_i \leq X \leq \theta_{i+1}\}$ is the same for all the intervals $[\theta_i, \theta_{i+1}]$. This path might get us somewhere close to the desired result for thresholds in one dimension, but it is rather unsatisfying: we need to reason about the unknown distribution P_X. For more general situations, such an analysis will be loose and impractical. A better way to do it is by working with the Rademacher process directly. Since the supremum of this process is within a factor 2 from the supremum of the empirical process (Theorem 7.10), we are not losing much.

11.2   From Finite to Infinite Classes: Second Attempt

For the example of thresholds, let us stay with the Rademacher averages of $\mathcal{F}$ rather than pass to the empirical process (11.1). We now reason conditionally on $X_1, \ldots, X_n$ and provide an upper bound on conditional Rademacher averages

$$\mathbb{E} \left\{ \sup_{\theta \in [0,1]} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{I}\{X_t \leq \theta\} \right] \; \middle| \; X_1, \ldots, X_n \right\}$$

(11.3)

An upper bound on the latter quantity that is independent of $X_1, \ldots, X_n$ would be quite interesting. It would effectively say that the unknown distribution P_X was edged out of the picture and replaced with random signs.

Conditional Rademacher averages in (11.3) can be equivalently written as

$$\mathbb{E}_\epsilon \sup_{a \in \mathcal{F}|_{\mathrm{X}^n}} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t a_t \right] \tag{11.4}$$

where $\mathcal{F}|_{\mathrm{X}^n} = \left\{ (f_\theta(\mathrm{X}_1), \ldots, f_\theta(\mathrm{X}_n)) \right\} \subseteq \{0,1\}^n$, the projection of $\mathcal{F}$ onto X^n. How big is this projection? Since $(\mathrm{X}_1, \ldots, \mathrm{X}_n)$ is held fixed, by varying θ we can realize vectors of the form $(1, \ldots, 1, 0, \ldots, 0)$, and there are $n+1$ of them. Clearly, the Euclidean length of the vectors $a \in \mathcal{F}|_{\mathrm{X}^n}$ is at most $\sqrt{n}$, so by Lemma 9.4,

$$\mathbb{E} \max_{f \in \mathcal{F}} \mathbb{S}_f \le \sqrt{\frac{2 \log(n+1)}{n}}$$

11.3   The Growth Function and the VC Dimension

The size of the projection $\mathcal{F}|_{\mathrm{X}^n}$ played an important role in obtaining an upper bound on Rademacher averages for the case of classification with thresholds. The same ideas carry over to a general classification setting.

Definition 11.1. For a binary valued function class $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$, the *growth function* is defined as

$$\Pi_{\mathcal{F}}(n) = \max \left\{ \mathrm{card}\left(\mathcal{F}|_{x_1, \ldots, x_n}\right) : x_1, \ldots, x_n \in \mathcal{X} \right\} \tag{11.5}$$

We have the following proposition that follows immediately from our previous arguments.

Proposition 11.2. *For classification with some domain $\mathcal{X}$, label set $\mathcal{Y} = \{0,1\}$, class of function $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ and loss function $\boldsymbol{\ell}(f, (x,y)) = \mathbf{I}\{f(x) \ne y\}$, it holds that for any distribution $P_{X \times Y}$,*

$$\mathbb{E} \sup_{g \in \boldsymbol{\ell}(\mathcal{F})} \mathbb{G}_g \le 2 \mathbb{E} \sup_f \mathbb{S}_f \le 2 \sqrt{\frac{2 \log \Pi_{\mathcal{F}}(n)}{n}}$$

The growth function measures expressiveness of $\mathcal{F}$. In particular, if $\mathcal{F}$ can produce all possible signs (that is, $\Pi_{\mathcal{F}}(n) = 2^n$), the bound becomes useless.

Definition 11.3. We say that $\mathcal{F}$ *shatters* some set $x_1, \ldots, x_n$ (a term due to J. Michael Steele) if $\mathcal{F}|_{x^n} = \{0,1\}^n$. That is,

$$\forall (b_1, \ldots, b_n) \in \{0,1\}^n, \ \exists f \in \mathcal{F} \ \text{ s.t. } \ f(x_t) = b_t \ \forall t \in \{1, \ldots, n\}$$

It is possible to show that if a set $x_1, \ldots, x_n$ is shattered for some n, then not only is the bound of Proposition 11.2 vacuous, but the learning problem itself (for the given n) is impossible.

✎ **Exercise 11.1** ($\star \star \star$). Prove the above statement.

We see that the growth function is quite important as a measure of complexity of $\mathcal{F}$ when it comes to distribution-free learning. But what is the behavior of the growth function? The situation turns out to be quite interesting. Define the following combinatorial parameter of $\mathcal{F}$:

Definition 11.4. The Vapnik-Chervonenkis (VC) dimension of the class $\mathcal{F} \subseteq \{0, 1\}^{\mathcal{X}}$ is defined as

$$\text{vc}(\mathcal{F}) \triangleq \max\left\{t \ : \ \Pi_{\mathcal{F}}(t) = 2^t\right\}$$

or $\text{vc}(\mathcal{F}) = \infty$ if max does not exist. Further, for $x_1, \ldots, x_n \in \mathcal{X}$, define

$$\text{vc}(\mathcal{F}, x_1, \ldots, x_n) \triangleq \text{vc}\left(\mathcal{F}|_{x_1, \ldots, x_n}\right) \triangleq \max\left\{t \ : \ \exists i_1, \ldots, i_t \in \{1, \ldots, n\} \text{ s.t. } \text{card}\left(\mathcal{F}|_{x_{i_1}, \ldots, x_{i_t}}\right) = 2^t\right\}$$

Vapnik-Chervonenkis dimension is the largest t such that $\mathcal{F}$ can produce all possible sequences of bits $\{0, 1\}$ on some set of examples. Clearly, the bound of Proposition 11.2 is useless for $n \leq \text{vc}(\mathcal{F})$. What about $n > \text{vc}(\mathcal{F})$? The following beautiful and surprising result was proved around the same time by Vapnik and Chervonenkis, Sauer, and Shelah.[1]

Lemma 11.5 (Vapnik-Chervonenkis, Sauer, Shelah). *It holds that*

$$\Pi_{\mathcal{F}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$$

whenever $vc(\mathcal{F}) = d < \infty$.

There are several ways to prove this lemma, and we give one that will be useful later in the course. Importantly, the sum that upper bounds the growth function has a $O(n^d)$ behavior. In fact,

$$\sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d, \tag{11.6}$$

---

and we leave the proof as an easy exercise. It is quite remarkable that the growth function is 2^t for $t \le d$ and polynomial afterwards, a non-obvious fact that makes the Vapnik-Chervonenkis results appealing.

As an immediate corollary of Proposition 11.2 we have

Corollary 11.6. *Under the setting of Proposition 11.2, if $vc(\mathcal{F}) = d$,*

$$\mathbb{E} \sup_{g \in \ell(\mathcal{F})} \mathbb{G}_g \le 2\mathbb{E} \sup_f \mathbb{S}_f \le 2\sqrt{\frac{2d \log(en/d)}{n}}$$

Example 7. The class of thresholds

$$\mathcal{F} = \{f_\theta(x) = \mathbf{I}\{x \le \theta\} \,:\, \theta \in [0,1]\}$$

on $\mathcal{X} = [0,1]$ has $vc(\mathcal{F}) = 1$. Indeed, one cannot find $x_1, x_2 \in \mathcal{X}$, $x_1 < x_2$, such that some threshold gives a label 0 to x_1 and 1 to x_2. Thus, no set of two points is shattered by $\mathcal{F}$.

Example 8. The class of step-up and step-down thresholds

$$\mathcal{F} = \{f_\theta(x) = \mathbf{I}\{x \le \theta\} \,:\, \theta \in [0,1]\} \cup \{g_\theta(x) = \mathbf{I}\{x \ge \theta\} \,:\, \theta \in [0,1]\}$$

on $\mathcal{X} = [0,1]$ has $vc(\mathcal{F}) = 2$.

The following analogue holds in d dimensions:

Example 9. Define the class of linear thresholds in $\mathbb{R}^d$ by

$$\mathcal{F} = \left\{ f_\theta(x) = \mathbf{I}\{\langle \theta, x \rangle \ge 0\} : \theta \in \mathbb{R}^d \right\}$$

Then $vc(\mathcal{F}) = d + 1$, which justifies our use of the same letter d for both dimensionality of the space and the combinatorial dimension of $\mathcal{F}$.

Example 10. Let the set $\mathcal{H}$ of functions $\mathcal{X} \to \mathbb{R}$ be a subset of a d-dimensional vector space. Then the class

$$\mathcal{F} = \left\{ \mathbf{I}\{h(x) \ge 0\} : h \in \mathcal{H} \right\}$$

has VC dimension at most d.

As the next example shows, it is not correct to think of the VC dimension as the number of parameters.

Example 11. The VC dimension of the class

$$\mathcal{F} = \left\{ f_\alpha(x) = \mathbf{I}\{\sin(\alpha x) \geq 0\} : \alpha \in \mathbb{R} \right\}$$

is infinite despite the fact that $\mathcal{F}$ is parametrized by a single number.

We now turn to the proof of the Vapnik-Chervonenkis-Sauer-Shelah lemma:

Proof of Lemma 11.5. Define a function

$$g(d, n) = \sum_{i=0}^{d} \binom{n}{i}$$

for $d, n \geq 0$. We would like to prove $\Pi_\mathcal{F}(n) \leq g(d, n)$ for $d = \text{vc}(\mathcal{F})$. By the way of induction, assume that the statement holds for $(d-1, n-1)$ and $(d-1, n)$. Fix any $x_1, \ldots, x_n \in \mathcal{X}$, and suppose $\text{vc}(\mathcal{F}, x_1, \ldots, x_n) = d$. Define

$$F = \left\{ r \in \mathcal{F}|_{x_2,\ldots,x_n} : (0, r), (1, r) \in \mathcal{F}|_{x_1,\ldots,x_n} \right\}$$

the set of projections of $\mathcal{F}$ on $(x_2, \ldots, x_n)$ such that both labels are realized on x_1. Let $F' = \mathcal{F}|_{x_2,\ldots,x_n}$, the set of all projections of $\mathcal{F}$ on $(x_2, \ldots, x_n)$. Observe that $F \subset F'$. We claim that

$$\text{card}\left(\mathcal{F}|_{x_1,\ldots,x_n}\right) = \text{card}(F) + \text{card}(F'). \tag{11.7}$$

To see this, suppose $r \in F$. Then both $(0, r)$ and $(1, r)$ are counted on the left-hand side of (11.7), and this is matched by counting r twice (in F and F'). If $r \in F' \setminus F$, then only $(0, r)$ or $(1, r)$ appears on the left-hand side, and the same is true for the right-hand side.

We now claim that $\text{vc}(F)$ is at most $d-1$, for otherwise $\text{vc}\left(\mathcal{F}|_{x_1,\ldots,x_n}\right) = d+1$ (indeed, both 0 and 1 can be appended to the full projection of size 2^d to create a projection of size 2^{d+1}). Then

$$\text{card}\left(\mathcal{F}|_{x_1,\ldots,x_n}\right) = \text{card}(F) + \text{card}(F') \leq g(d-1, n-1) + g(d, n-1) \tag{11.8}$$

by the induction hypothesis. Since this holds for any $x_1, \ldots, x_n$, we also have

$$\Pi_\mathcal{F}(n) \leq g(d-1, n-1) + g(d, n-1).$$

The induction step follows from the identity

$$g(d, n-1) + g(d-1, n-1) = g(d, n).$$

The base of the induction is easy to verify. □

The proof of Lemma 11.5 might appear magical and unintuitive. Let us point (in a very informal way) to a few important pieces that make it work. In fact, the technique is rather general and will be used once again in a few lectures when we prove an analogue of the combinatorial lemma for sequential prediction. The abstract idea is that the behavior of $\mathcal{F}$ on a sample can be separated into two pieces (F and F'). Let us now forget about the way that these pieces are defined, and rather look into how their respective sizes can be used to unwind a recursion. Imagine that a split into F and F' can be done so that the smaller piece F is always of size 1. The recursion would then unfold as $g(d, n) \leq g(d, n-1) + 1 \leq g(d, n-2) + 2 \leq \ldots \leq n$. Of course, such a recursion is rather trivial. Now, imagine another type of recursion: both pieces are of the same size. Then we would have $g(d, n) \leq 2g(d, n-1) \leq 4g(d, n-2) \leq \ldots \leq 2^n$. Once again, such a recursion is rather trivial, and the bound is not useful. In some sense, these two cases are extremes of how a recursion might play out. But note that we did not use d in the two recursions. Imagine now that each time we perform the split into F and F', a "counter" d is decreased for the smaller piece. That is, $g(d, n) \leq g(d-1, n-1) + g(d, n-1)$. This situation is in-between the bounds of n and 2^n, and, in fact, it is of the order n^d as the proof shows.

The above observation is a rather general recipe: for a problem at hand, find a parameter ("counter") so that the size of the smaller set being split off has to have a smaller value for this parameter. In this case, the recursion gives a polynomial rather than exponential size.

Statistical Learning: Real-Valued Functions

In the real-valued supervised learning problem with i.i.d. data, we consider a bounded set $\mathcal{Y} = [-1, 1]$. Fix a set $\mathcal{F}$ of functions $\mathcal{X} \to \mathcal{Y}$, for some input space $\mathcal{X}$. In the distribution-free scenario, we have i.i.d. data $\{(X_t, Y_t)\}_{t=1}^{n}$ from an unknown $P_{X \times Y}$, on which we place no assumption.

Recall that the cardinality of the set

$$\mathcal{F}|_{x_1, \dots, x_n} = \left\{ (f(x_1), \dots, f(x_n)) : f \in \mathcal{F} \right\}$$

played an important role in the analysis of classification problems. For real-valued functions, however, the cardinality of this set is of little use since it is, in general, uncountable. However, two functions f and g with almost identical values

$$f(x_t) \approx g(x_t), \quad t \in \{1, \dots, n\}$$

on the sample are essentially the same for our purposes, and we need to find a way to measure complexity of $\mathcal{F}|_{x_1, \dots, x_n}$ without paying attention to small discrepancies between functions. This idea is captured by the notion of *covering numbers*.

12.1   Covering Numbers

Recall our shorthand $x^n = \{x_1, \dots, x_n\}$.

Definition 12.1. A set $V \subset \mathbb{R}^n$ is an α-*cover* on $x_1, \dots, x_n$ with respect to ℓ_2 norm if

$$\forall f \in \mathcal{F}, \ \exists v \in V \ \ \text{s.t.} \ \ \left(\frac{1}{n} \sum_{t=1}^{n} (f(x_t) - v_t)^2 \right)^{1/2} \leq \alpha. \tag{12.1}$$

An α-*covering number* is defined as

$$\mathcal{N}_2(\mathcal{F}, \alpha, x^n) = \min\left\{\mathrm{card}\,(V) \;:\; V \text{ is an } \alpha\text{-cover}\right\}.$$
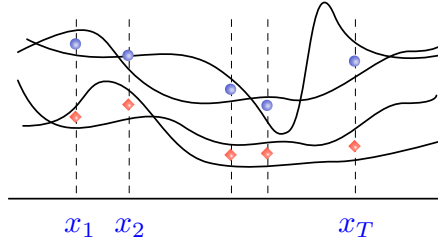

Figure 12.1: Two sets of levels provide an α-cover for the four functions. Only the values of functions on $x_1, \ldots, x_n$ are relevant.

As far as the cover is concerned, the values of functions outside $x_1, \ldots, x_n$ are immaterial. Hence, we may equivalently talk of a cover for the set $\mathcal{F}|_{x^n} \subseteq [-1, 1]^n$. We may then write $\mathcal{N}_2(\mathcal{F}|_{x^n}, \alpha)$ for the α-covering number.

It is important to note that here and below we refer to ℓ_p norms, but the definitions in fact use an extra normalization factor $1/n$. We may equivalently write the α-closeness requirement (12.1) as $\|\mathsf{f} - v\|_2 \le \sqrt{n}\alpha$ where $\mathsf{f} = (f(x_1), \ldots, f(x_t)) \in \mathcal{F}|_{x_1, \ldots, x_n}$.

If the set V is realized as $V = \left\{(g(x_1), \ldots, g(x_n)) : g \in \mathcal{G}\right\}$ for some collection $\mathcal{G} \subseteq \mathcal{F}$, we say that the cover is *proper*. A proper cover picks a subset $\mathcal{G}$ of $\mathcal{F}$ so that any $f \in \mathcal{F}$ is close to some $g \in \mathcal{G}$ on $x_1, \ldots, x_n$ in the ℓ_2 sense. The distinction between proper and improper covers is minor, and one can show that a proper covering number at the α level is sandwiched between improper covering numbers at α and $\alpha/2$ levels.

Of course, we can define ℓ_p-covers as

Definition 12.2. A set $V \subset \mathbb{R}^n$ is an α-*cover* on $x_1, \ldots, x_n$ with respect to ℓ_p norm, for $p \in [1, \infty)$, if

$$\forall f \in \mathcal{F}, \, \exists v \in V \;\; \text{s.t.} \;\; \left(\frac{1}{n} \sum_{t=1}^{n} |f(x_t) - v_t|^p\right)^{1/p} \le \alpha,$$

and with respect to ℓ_∞ norm if

$$\forall f \in \mathcal{F}, \, \exists v \in V \;\; \text{s.t.} \;\; |f(x_t) - v_t| \le \alpha \;\; \forall t \in \{1, \ldots, n\}.$$

The α-covering numbers $\mathcal{N}_p(\mathcal{F}, \alpha, x^n)$ and $\mathcal{N}_\infty(\mathcal{F}, \alpha, x^n)$ are defined as before.

The above notions of a cover can be seen as instances of a more general idea of an α-cover for a metric space (T, d). A subset $T' \subset T$ is an α-cover of T if for any $t \in T$ there exists a $s \in T'$ such that $d(t, s) \leq \alpha$. Above definitions correspond to endowing $\mathcal{F}$ with an empirical metric $L_p(\mu_n)$ where μ_n is the empirical distribution on the data $x_1, \ldots, x_n$, thus explaining the extra $1/n$ normalization factor.
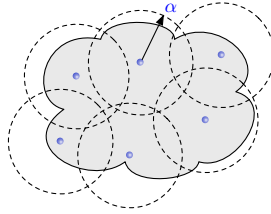

Figure 12.2: An α-cover for a metric space (n, d).

✎ **Exercise 12.1** ($\star$). Prove that for any $1 \leq p \leq q \leq \infty$, $\mathcal{N}_p(\mathcal{F}, \alpha, x^n) \leq \mathcal{N}_q(\mathcal{F}, \alpha, x^n)$.

Covering of $\mathcal{F}$ allows us to squint our eyes and view the set $\mathcal{F}|_{x^n}$ as a finite set at a granularity α. We can then apply the maximal inequalities for stochastic processes indexed by finite sets. Of course, by doing so we lose α of precision. Clearly, smaller α means larger α-cover, so there is tension between fine granularity and small cover. This is quantified in the following proposition.

Proposition 12.3. *For any $x^n = \{x_1, \ldots, x_n\}$, conditional Rademacher averages of a function class $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ satisfy*

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ \alpha + \sqrt{\frac{2 \log \mathcal{N}_1(\mathcal{F}, \alpha, x^n)}{n}} \right\}$$

Proof. Fix $x_1, \ldots, x_n$ and $\alpha > 0$. Let V be a minimal α-cover of $\mathcal{F}$ on x^n with respect to ℓ_1-norm. For $f \in \mathcal{F}$, denote by $v[f] \in V$ any element that "α-covers" f in the sense given by the definition. Denote by $v[f]_t$ the tth coordinate of the vector.

106

Write the conditional Rademacher averages as

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) = \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(x_t)\right\}$$

$$= \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(f(x_t) - v[f]_t\right) + \epsilon_t v[f]_t\right\}$$

$$\leq \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(f(x_t) - v[f]_t\right)\right\} + \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t v[f]_t\right\}$$

$$\leq \sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}|f(x_t) - v[f]_t| + \mathbb{E}\left\{\max_{v\in V}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t v_t\right\}$$

By definition of α-cover, the first term is upper bounded by α. The second term is precisely controlled by the maximal inequality of Lemma 9.4, which gives an upper bound of

$$\sqrt{\frac{2\log|V|}{n}},$$

due to the fact that $\|v\|_2 \leq \sqrt{n}$ for any $v \in V$. Note that if any coordinate of v is outside the interval $[-1,1]$, it can be truncated.

Since $\alpha > 0$ is arbitrary, the statement follows. $\square$

Example 12. For certain classes $\mathcal{F}$, one can show that $\mathcal{F}_{x_1,\ldots,x_n}$ is a subset of $[-1,1]^n$ of dimension lower than n. This happens, for instance, if

$$\mathcal{F} = \left\{f(x) = \langle f, x\rangle : f \in B_p^d\right\} \text{ and } \mathcal{X} = B_q^d,$$

where B_p^d and B_q^d are unit balls in $\mathbb{R}^d$, with $1/p + 1/q = 1$. For instance, for $p = \infty$, the function class is identified with $[-1,1]^d$ and it is easy to check that

$$\mathcal{N}_\infty(\mathcal{F}, \alpha, x^n) \leq \left(\frac{2}{\alpha}\right)^d$$

by a discretization argument. The bound of Proposition 12.3 then yields

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) \leq \inf_{\alpha\geq 0}\left\{\alpha + \sqrt{\frac{2d\log(2/\alpha)}{n}}\right\} \tag{12.2}$$

Up to constant factors, the best choice for α is $n^{-1/2}$, which yields an upper bound of the order $O\left(\sqrt{\frac{d\log n}{n}}\right)$.

Since $B_p^d \subseteq B_\infty^d$ for any $p > 0$, the above bound also holds for other $\mathcal{F}$, though it might be loose. We also remark that an ℓ_∞ cover is quite a stringent requirement, as Proposition 12.3 only requires the ℓ_1 cover. Finally, we note that the discretization argument yields more than an ℓ_∞ cover on x^n: it yields a pointwise cover *for all $x \in \mathcal{X}$*. In general such pointwise bounds are not possible as soon as we consider infinite-dimensional $\mathcal{F}$.

12.2 Chaining Technique and the Dudley Entropy Integral

An application of Proposition 12.3 in Example 12 yields an $O\left(\sqrt{\frac{d \log n}{n}}\right)$ upper bound, containing an extraneous $\log n$ factor. How do we know that it is extraneous? Well, because we can get a bound without this factor. While we might not care about a logarithmic factor, the situation is actually more serious: we can provide an example where Proposition 12.3 gives the wrong rate. Somehow, squinting our eyes and looking at $\mathcal{F}$ at a single level of granularity does not give the right picture. What if we looked at $\mathcal{F}$ at different coarseness levels and integrated them to build a full-resolution panorama? The next theorem does exactly that.

Theorem 12.4. *For any $x^n = \{x_1, \ldots, x_n\}$, conditional Rademacher averages of a function class $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ satisfy*

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) \leq \widehat{\mathcal{D}}^{iid}(\mathcal{F})$$

where

$$\widehat{\mathcal{D}}^{iid}(\mathcal{F}) \triangleq \inf_{\alpha \geq 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1 \sqrt{\log \mathcal{N}_2(\mathcal{F}, \delta, x^n)}\,d\delta\right\} \tag{12.3}$$

Proof. Fix $x_1, \ldots, x_n$ and let $F = \mathcal{F}|_{x^n}$. Let $\alpha_j = 2^{-j}$ for $j = 1, 2, \ldots$. Let V_j be a minimal α_j-cover of size $\mathcal{N}_2(\mathcal{F}, \alpha_j, x^n)$. For any $\mathsf{f} \in F$, let $v[\mathsf{f}]^j \in V_j$ be an element α_j-close to f, as promised by the definition of an α_j-cover. Each f is therefore associated with a chain of approximations $v[\mathsf{f}]^1, v[\mathsf{f}]^2, \ldots$ at increasingly fine levels of granularity. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ denote the vector of Rademacher random variables. Then we may succinctly write conditional Rademacher averages as

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) = \mathbb{E}\left\{\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^n \epsilon_i f(x_i)\right\} = \frac{1}{n}\mathbb{E}\sup_{\mathsf{f} \in F}\langle \epsilon, \mathsf{f}\rangle$$

where $\langle a, b \rangle$ is the inner product. We now rewrite the (unnormalized) averages as in Proposition 12.3, but with a chain of N telescoping differences (N to be determined later):

$$\mathbb{E} \sup_{f \in F} \langle \epsilon, f \rangle = \mathbb{E} \sup_{f \in F} \left\{ \left\langle \epsilon, f - v[f]^N \right\rangle + \sum_{j=1}^{N} \left\langle \epsilon, v[f]^j - v[f]^{j-1} \right\rangle \right\}$$

$$\leq \mathbb{E} \sup_{f \in F} \left\langle \epsilon, f - v[f]^N \right\rangle + \sum_{j=1}^{N} \mathbb{E} \sup_{f \in F} \left\langle \epsilon, v[f]^j - v[f]^{j-1} \right\rangle$$

with $v[f]^0 = \mathbf{0}$. The first term is upper bounded above by Cauchy-Shwartz inequality as

$$\mathbb{E} \sup_{f \in F} \left\langle \epsilon, f - v[f]^N \right\rangle \leq \mathbb{E} \sup_{f \in F} \| \epsilon \| \cdot \| f - v[f]^N \| = \sqrt{n} \sup_{f \in F} \left(\sum_{t=1}^{n} \left(f_t - v[f]_t^N \right)^2 \right)^{1/2} = n \alpha_N$$

$$(12.4)$$

by definition. Turning to the second term, observe that, for a given j, the difference $v[f]^j - v[f]^{j-1}$ takes on at most $\text{card}(V_j) \times \text{card}(V_{j-1})$ possible values. This allows us to invoke the maximal inequality of Lemma 9.4:

$$\mathbb{E} \sup_{f \in F} \left\langle \epsilon, v[f]^j - v[f]^{j-1} \right\rangle = r \sqrt{2 \log(\text{card}(V_j) \times \text{card}(V_{j-1}))}$$

where $r = \sup_{f \in F} \| v[f]^j - v[f]^{j-1} \|$. For any $f \in \mathcal{F}$,

$$\left\| v[f]^j - v[f]^{j-1} \right\| \leq \left\| v[f]^j - f \right\| + \left\| v[f]^{j-1} - f \right\| \leq \sqrt{n} \alpha_j + \sqrt{n} \alpha_{j-1} = 3\sqrt{n} \alpha_j .$$

Since size of the minimal cover does not decrease for finer resolution, it holds that

$$\text{card}(V_j) \times \text{card}(V_{j-1}) \leq \text{card}(V_j)^2 .$$

We conclude that conditional Rademacher averages are upper bounded as

$$\frac{1}{n} \mathbb{E} \sup_{f \in F} \langle \epsilon, f \rangle \leq \alpha_N + \frac{6}{\sqrt{n}} \sum_{j=1}^{N} \alpha_j \sqrt{\log \text{card}(V_j)}$$

Using the simple identity $2(\alpha_j - \alpha_{j+1}) = \alpha_j$, as well as the fact that each V_j is a minimal cover, we may rewrite the upper bound as

$$\alpha_N + \frac{12}{\sqrt{n}} \sum_{j=1}^{N} (\alpha_j - \alpha_{j+1}) \sqrt{\log \mathcal{N}_2(\mathcal{F}, \alpha_j, x^n)} \leq \alpha_N + \frac{12}{\sqrt{n}} \int_{\alpha_{N+1}}^{1} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \delta, x^n)} d\delta.$$

Now, fix any $\alpha > 0$ and let $N = \sup\{j : \alpha_j \geq 2\alpha\}$. Then $\alpha_{N+1} < 2\alpha$ and $4\alpha > \alpha_N \geq 2\alpha$. Hence, $\alpha_{N+1} = \alpha_N/2 > \alpha$ and

$$\frac{1}{n}\mathbb{E}\sup_{f\in F} \leq \inf_{\alpha>0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1 \sqrt{\log\mathcal{N}_2(\mathcal{F},\delta,x^n)}d\delta\right\}$$

The result also holds for $\alpha = 0$, and this completes the proof. $\qquad\square$

12.3 Example: Nondecreasing Functions

Before discussing the virtues of Theorem 12.4, let us mention that $\mathcal{N}(\mathcal{F},\alpha,x^n)$ is always finite for a class $\mathcal{F}$ of bounded functions. Indeed, one can create a cover (let's say in ℓ_∞) by discretizing the interval $[-1,1]$ to the level of α for the values of functions on each of $x_1,\ldots,x_n$. Then, let V be a set of vectors obtained by taking any of the $2/\alpha$ possible values on x_1, any of the $2/\alpha$ possible values on x_2, and so on. Clearly, the size of such a cover is $\left(\frac{2}{\alpha}\right)^n$. This is illustrated in Figure 12.3. We conclude that for any class $\mathcal{F}$ of bounded functions, the covering numbers on
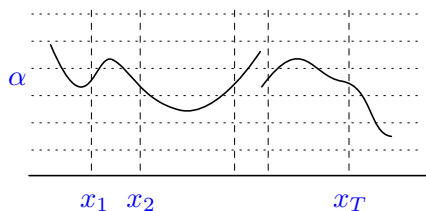
Figure 12.3: Without any assumptions on the function, the size of an α-cover is $\Omega\left((1/\alpha)^n\right)$

$x_1,\ldots,x_n$ are finite, yet they yield vacuous upper bounds in both Proposition 12.3 and in Theorem 12.4. Without any assumptions on $\mathcal{F}$, learning is not possible, in which case we should not expect to obtain any meaningful upper bounds.

Consider a class of functions $\mathcal{F}$ on $\mathcal{X} = \mathbb{R}$. Let us place only one restriction on the functions: they are non-decreasing. What happens to the covering numbers? Do we still require $\Omega\left((1/\alpha)^n\right)$ vectors to provide an α-cover for $\mathcal{F}$ on $x_1,\ldots,x_n$? Is the associated prediction problem (say, with absolute loss) learnable? To answer these questions, let us construct an ℓ_∞ cover V of $\mathcal{F}$ on x^n. Without loss of generality, suppose $x_1,\ldots,x_n$ are ordered on the real line. Then the elements of V should

12.3 Example: Nondecreasing Functions

Figure 12.4: The class of non-decreasing functions on $\mathbb{R}$ has a manageable covering number.

be n-dimensional vectors with non-decreasing coordinates. How many such vectors $v \in V$ are required to provide α-approximation to any non-decreasing function? For any $\bar{y}$ in the set of $(2/\alpha)$ discretized y-values, we need to specify the coordinate $i \in \{1,\ldots,n\}$ such that the i-th coordinate of v is below $\bar{y}$ 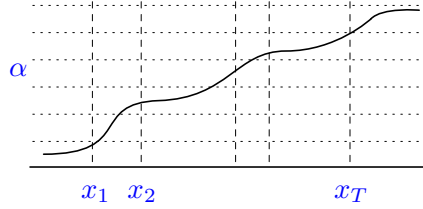but $i+1$th is greater than $\bar{y}$. This gives all the possibilities. A simple counting argument gives $|V| \leq n^{2/\alpha}$, and, therefore,

$$\mathcal{N}_\infty(\mathcal{F}, \alpha, x^n) \leq n^{2/\alpha}. \tag{12.5}$$

for any $x_1,\ldots,x_n$. Observe that the seemingly innocuous assumption of monotonicity drastically decreased the covering number from $(2/\alpha)^n$ to $n^{2/\alpha}$.

Let us now apply the upper bound of Proposition 12.3 to the class of nondecreasing functions on $\mathbb{R}$:

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ \alpha + \sqrt{\frac{2(2/\alpha)\log n}{n}} \right\}$$

Optimizing the bound leads to $\alpha = \left(\frac{4\log n}{n}\right)^{1/3}$, and the final bound of

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) \leq 2\left(\frac{4\log n}{n}\right)^{1/3}.$$

Let us now compare this to the bound given by Theorem 12.4 and see if our efforts paid off. We have

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) \leq \widehat{\mathcal{D}}^{iid}(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{(2/\delta)\log n}\, d\delta \right\}$$

Simplifying the integral

$$\int_\alpha^1 \sqrt{2/\delta}\, d\delta = 2\sqrt{2}\left[\sqrt{\delta}\right]_\alpha^1 \leq 2\sqrt{2}$$

and taking $\alpha = 0$, we obtain

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) \leq 24\sqrt{\frac{2\log n}{n}}, \tag{12.6}$$

a qualitatively different rate from that given by Proposition 12.3.

As a final remark, we mention that the domain $\mathcal{X} = \mathbb{R}$ of the functions in the above example is unbounded and there is no hope to have a finite pointwise cover of $\mathcal{F}$ over $\mathcal{X}$. Early results in statistical learning theory relied on the availability of such covering numbers, which greatly limited the applicability of the results. The symmetrization technique allows us to treat data as fixed and to only consider covering numbers of $\mathcal{F}$ on the data x^n.

✎ **Exercise 12.2** ($\star\star$)**.** Show that the bound of Theorem 12.4 removes the extraneous logarithmic factor in the upper bound (12.2) of Example 12.

12.4   Improved Bounds for Classification

Classification scenario (that is, $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$) can also benefit from the upper bound of Theorem 12.4. First, notice that a projection $\mathcal{F}|_{x_1,\dots,x_n}$ is nothing but a cover of $\mathcal{F}$ at scale $\alpha = 0$. Indeed, for any function $f \in \mathcal{F}$, trivially, there exists an element $v \in V = \mathcal{F}|_{x_1,\dots,x_n}$ such that $f(x_t) = v_t$ for all t. Henceforth, we shall refer to such a cover as a *zero-cover* or *exact cover*, and the covering number by $\mathcal{N}_0(\mathcal{F}, x_1,\dots,x_n)$. For such a cover at scale $\alpha = 0$, we need not specify the ℓ_p norm. Alternatively, we can denote the zero cover by $\mathcal{N}_\infty(\mathcal{F}, x^n, \alpha)$ for any $0 \leq \alpha < 1/2$.

The Vapnik-Chervonenkis-Sauer-Shelah lemma (Lemma 11.5) states that for any $x_1,\dots,x_n$,

$$\mathcal{N}_0(\mathcal{F}, x_1,\dots,x_n) \leq \left(\frac{en}{d}\right)^d \tag{12.7}$$

for $d = \text{vc}(\mathcal{F})$.

Recall that Proposition 11.2 yielded an upper bound with an extra $\log n$ factor. In fact, it is not possible to avoid this factor if we are dealing with the exact cover.

However, what if we obtain ℓ_2 covering numbers for $\mathcal{F}$ and apply Theorem 12.4? Recall from the last section (see the last exercise) that this indeed removes the extraneous factor.

Refining the proof of Dudley [22], Mendelson [38] gives a nice probabilistic argument that for any $x_1, \ldots, x_n$, the ℓ_p-covering numbers of the class $\mathcal{F}$ of $\{0, 1\}$-valued functions satisfy

$$\mathcal{N}_p(\mathcal{F}, \alpha, x^n) \leq \left((2pe^2) \log \frac{2e^2}{\alpha} \right)^d \left(\frac{1}{\alpha} \right)^{pd}. \tag{12.8}$$

Taking $p = 2$, for classification problems with $\mathrm{vc}(\mathcal{F}) = d$, Theorem 12.4 yields an $O\left(\sqrt{\frac{d}{n}} \right)$ upper bound without the superfluous logarithmic factors.

12.5 Combinatorial Parameters

Back to the setting of real-valued prediction, a natural question to ask is whether there is a combinatorial parameter of $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ that controls the behavior of covering numbers. Historically, the first such parameter was *pseudo-dimension*, due to D. Pollard:

Definition 12.5. The *pseudo-dimension* of $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ is defined as the Vapnik-Chervonenkis dimension of the class

$$\mathcal{G} = \left\{ g(x, y) = \mathbf{I}\left\{ f(x) - y \geq 0 \right\} : f \in \mathcal{F} \right\}.$$

Equivalently, pseudo-dimension is the largest d such that there exist $(x_1, \ldots, x_n) \in \mathcal{X}^n$ and $(y_1, \ldots, y_n) \in \mathbb{R}^n$ with the following property:

$$\forall (b_1, \ldots, b_n) \in \{0, 1\}^n, \ \exists f \in \mathcal{F} \ \text{ s.t. } \ f(x_t) \geq y_t \ \Leftrightarrow \ b_t = 1$$

Pseudo-dimension of $\mathcal{F}$ is a measure of its expressiveness. To show that $\mathcal{F}$ has pseudo-dimension at least d one needs to provide d input variables $x_1, \ldots, x_n$ and d real-valued "levels" y_t such that for any bit sequence $b_1, \ldots, b_n$, there is a function that passes above and below these levels at $x_1, \ldots, x_n$ as dictated by the bit sequence.

It turns out that finiteness of pseudo-dimension is sufficient to guarantee an upper bound on the covering numbers, but not necessary. For instance, consider the class of non-decreasing functions on the real line, as discussed earlier. Let us

prove that the pseudo-dimension is infinite. Indeed, for d as large as we want, let $x_t = t$ and $y_t = t/d$, for each $t \in \{1, \ldots, d\}$. Then, for any bit sequence $(b_1, \ldots, b_d) \in \{0, 1\}$, there is a non-decreasing function that passes above the level y_t if $b_t = 1$ and under if $b_t = 0$. This is clearly due to the fact that there is enough wiggle room between $y_t = t/d$ and $y_{t+1} = (t+1)/d$ for a non-decreasing function to pass either above or below the threshold.

While the pseudo-dimension is infinite for the class of non-decreasing functions, we saw earlier that the covering numbers grow at most as $n^{2/\alpha}$, yielding a rate of $\sqrt{(\log n)/n}$. Thus, the pseudo-dimension is not a satisfactory parameter for capturing the problem complexity. Intuitively, at least in the above example, the problem stems from the fact that there is no notion of *scale* in the definition of pseudo-dimension. For instance, if the requirement was to pass α-above and α-below the levels y_t, the above construction of a shattered set with $y_t = t/d$ would break as soon as $\alpha = 1/d$. This intuition indeed leads to the "right" combinatorial notion, introduced by Kearns and Schapire [31]:

Definition 12.6. We say that $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ α-shatters a set $(x_1, \ldots, x_n)$ if there exist $(y_1, \ldots, y_n) \in \mathbb{R}^n$ (called a *witness to shattering*) with the following property:

$$\forall (b_1, \ldots, b_n) \in \{0, 1\}^n, \ \exists f \in \mathcal{F} \ \text{ s.t. } \ f(x_t) > y_t + \frac{\alpha}{2} \ \text{ if } \ b_t = 1 \ \text{ and } \ f(x_t) < y_t - \frac{\alpha}{2} \ \text{ if } \ b_t = 0$$

The *fat-shattering dimension* of $\mathcal{F}$ at scale α, denoted by $\mathrm{fat}(\mathcal{F}, \alpha)$, is the size of the largest α-shattered set.
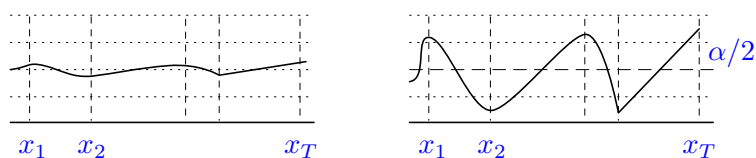
Figure 12.5: Left: Pseudo-dimension can be infinite for functions that barely wiggle. Right: In contrast, fat-shattering dimension requires the functions to pass above and below levels y_t by a margin $\alpha/2$. Here, y_t's are constant.

Example 13. For the class of non-decreasing functions on the real line, the fat-shattering dimension satisfies $\mathrm{fat}(\mathcal{F}, \alpha) \leq 4/\alpha$. Indeed, suppose $x_1 < x_2 \ldots < x_d$ are

α-shattered by $\mathcal{F}$. Suppose $y_1 \leq \ldots \leq y_n$ is a witness to shattering, and take the alternating sequence $(0, 1, 0, 1, \ldots)$ of bits. For simplicity, suppose d is even. Then, there must be a non-decreasing function that passes below every odd-indexed and above every even-indexed y_t by a margin α. The jump of the function is then at least α between every odd and even index, and there can be at most $2/\alpha$ such jumps for a non-decreasing $[-1, 1]$-valued function. Thus,

$$\text{fat}(\mathcal{F}, \alpha) \leq 4/\alpha \tag{12.9}$$

The first part of the following theorem is due to Mendelson and Vershynin [39], and the second to [48] (the proofs are difficult and go well beyond this course):

Theorem 12.7. *For $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ and any $0 < \alpha < 1$,*

$$\mathcal{N}_2(\mathcal{F}, \alpha, x^n) \leq \left(\frac{2}{\alpha}\right)^{K \cdot \text{fat}(\mathcal{F}, c\alpha)} \tag{12.10}$$

where K, c are positive absolute constants. Furthermore, for any $\delta \in (0, 1)$,

$$\log \mathcal{N}_\infty(\mathcal{F}, \alpha, x^n) \leq C d \log\left(\frac{n}{d\alpha}\right) \log^\delta(n/d), \tag{12.11}$$

where $d = \text{fat}(\mathcal{F}, c\delta\alpha)$ for some positive constant c.

The first result can be seen as a generalization of (12.8) to real-valued functions, and the second – as a generalization of the Vapnik-Chervonenkis-Sauer-Shelah lemma. To the best of our knowledge, it is an open question whether the extra term involving $\log^\delta(n/d)$ can be removed, and it is believed that a more natural bound $\log \mathcal{N}_\infty(\mathcal{F}, \alpha, x^n) \leq C d \log\left(\frac{n}{d\alpha}\right)$ with $d = \text{fat}(\mathcal{F}, c\alpha)$ should be possible.

♛ Prove this fact and earn an A+ in the course.

A few more remarks about Theorem 12.7. First, upper bounds on $\mathcal{N}_p$ for $1 \leq p < \infty$ are of the same form as (12.10), with constants c, K depending only on p (see [38, 39]). Second, note that the ℓ_∞ covering numbers depend on n, a feature already observed in the upper bound of the Vapnik-Chervonenkis-Sauer-Shelah lemma (see Eq.(12.7)). In fact, this dependence is inevitable for $\mathcal{N}_\infty$. On the other hand, ℓ_p covering numbers ($1 \leq p < \infty$) are independent of n. Thankfully, the bound of Theorem 12.4 is in terms of an ℓ_2 covering number, so we need not concern ourselves with covering in the ℓ_∞ norm. Finally, we remark that for sequential

problems, the analogue of (12.11) is relatively easy to prove without the extra logarithmic factor, which is somewhat surprising. We will later point out the place where the difficulty arises in the i.i.d. but not in the sequential proof.

Example 14. We turn again to the example of non-decreasing functions on the real line. Recall that we calculated ℓ_∞ covering numbers in (12.5) and used this upper bound in conjunction with Theorem 12.4 to get an $O\left(\sqrt{\frac{\log n}{n}}\right)$ upper bound in Eq. (12.6). We have also calculated an upper bound $\mathrm{fat}(\mathcal{F}, \alpha) \le 4/\alpha$ on the fat-shattering dimension in (12.9). Using this bound together with Theorem 12.7, we get a direct estimate

$$\mathcal{N}_2(\mathcal{F}, \alpha, x^n) \le \left(\frac{2}{\alpha}\right)^{c/\alpha}, \quad \text{for some constant } c > 0$$

on the ℓ_2 covering numbers without a detour through the inferior ℓ_∞ covering numbers. Theorem 12.4 then yields for any $x_1, \ldots, x_n$

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) \le \inf_{\alpha \ge 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{\frac{c}{\delta} \log\left(\frac{2}{\delta}\right)} d\delta \right\}$$

Taking $\alpha = 1/\sqrt{n}$, we observe that the integral above is bounded by the integral of a polynomial of power larger than δ^{-1}: say, $\delta^{-3/4}$. The latter integral over $[\alpha, 1]$ is bounded by a constant, yielding an upper bound

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) \le \frac{C}{\sqrt{n}}$$

for some absolute constant C. We conclude that a more careful analysis with a fat-shattering dimension and the Dudley Entropy Integral removes the superfluous logarithmic factor. In other situations, this approach can also lead to "correct" rates.

Of course, putting together Theorem 12.4 and the bound (12.10) of Theorem 12.7 leads to a general upper bound

Corollary 12.8. *For $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$, for any $x_1, \ldots, x_n \in \mathcal{X}$,*

$$\widehat{\mathcal{D}}^{iid}(\mathcal{F}) \le \inf_{\alpha \ge 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{K \mathrm{fat}(\mathcal{F}, c\delta) \log\left(\frac{2}{\delta}\right)} d\delta \right\}$$

for some absolute constants c, K.

12.6 Contraction

So far, we have developed many tools for upper bounding the Rademacher averages of a function class. In addition to the properties outlined in Lemma 7.14, the following lemma about Rademacher averages is useful:

Lemma 12.9. *If $\phi : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz (that is, $\phi(a) - \phi(b) \le L|a - b|$ for all $a, b \in \mathbb{R}$), then*

$$\widehat{\mathcal{R}}^{iid}(\phi \circ \mathcal{F}) \le L\widehat{\mathcal{R}}^{iid}(\mathcal{F})$$

Proof. For brevity, let us prove the result for $L = 1$. We have

$$\widehat{\mathcal{R}}^{iid}(\phi \circ \mathcal{F}) = \mathbb{E}_{\epsilon_1,\dots,\epsilon_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \phi(f(x_t))$$

$$= \frac{1}{2}\left\{ \mathbb{E}_{\epsilon_1,\dots,\epsilon_{n-1}} \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n-1} \epsilon_t \phi(f(x_t)) + \phi(f(x_n)) \right] + \mathbb{E}_{\epsilon_1,\dots,\epsilon_{n-1}} \sup_{g \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n-1} \epsilon_t \phi(g(x_t)) - \phi(g(x_n)) \right] \right\}$$

$$= \frac{1}{2}\mathbb{E}_{\epsilon_1,\dots,\epsilon_{n-1}} \sup_{f,g \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n-1} \epsilon_t \Big(\phi(f(x_t)) + \phi(g(x_t))\Big) + \phi(f(x_n)) - \phi(g(x_n)) \right]$$

$$= \frac{1}{2}\mathbb{E}_{\epsilon_1,\dots,\epsilon_{n-1}} \sup_{f,g \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n-1} \epsilon_t \Big(\phi(f(x_t)) + \phi(g(x_t))\Big) + \big|f(x_n)) - g(x_n)\big| \right]$$

In the first step, we expanded the expectation over ϵ_n as a sum of two terms; in the second, we combined two suprema into one, and in the third used the Lipschitz property. Now, fix $\epsilon_{1:n-1}$ and suppose the supremum is achieved at some (f^*, g^*). If $f^*(x_n) \ge g^*(x_n)$, we can remove the absolute values in the last term. Otherwise, the pair (g^*, f^*) gives the same value as the last expression, and we can remove the absolute values once again. Thus, the last expression is upper bounded by

$$\frac{1}{2}\mathbb{E}_{\epsilon_1,\dots,\epsilon_{n-1}} \sup_{f,g \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n-1} \epsilon_t \Big(\phi(f(x_t)) + \phi(g(x_t))\Big) + (f(x_n)) - g(x_n)) \right]$$

$$\le \frac{1}{2}\left\{ \mathbb{E}_{\epsilon_1,\dots,\epsilon_{n-1}} \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n-1} \epsilon_t \phi(f(x_t)) + f(x_n) \right] + \mathbb{E}_{\epsilon_1,\dots,\epsilon_{n-1}} \mathbb{E}_{\epsilon_1,\dots,\epsilon_{n-1}} \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n-1} \epsilon_t \phi(f(x_t)) - f(x_n) \right] \right\}$$

$$= \mathbb{E}_{\epsilon_1,\dots,\epsilon_n} \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n-1} \epsilon_t \phi(f(x_t)) + \epsilon_n f(x_n) \right]$$

Continuing in this fashion, we "remove" ϕ for every term. $\qquad\square$

Lemma 12.9 will be called "Contraction Principle". This result is indeed the missing link that allows us to relate the supremum of the Rademacher processes over $\ell(\mathcal{F})$ to the supremum of the Rademacher process over $\mathcal{F}$ itself whenever ℓ is a Lipschitz function. In the case of the indicator loss (which is *not* Lipschitz), we already observed in (8.4) that

$$\mathbb{E} \sup_{g \in \ell(\mathcal{F})} \mathbb{S}_g = \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{S}_f .$$

Now, in view of Lemma 12.9, we also conclude that

$$\mathbb{E} \sup_{g \in \ell(\mathcal{F})} \mathbb{S}_g \leq L \cdot \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{S}_f$$

whenever the loss function is L-Lipschitz.

12.7 Discussion

We now have a full arsenal of tools for proving upper bound on the supremum of the Rademacher process and, hence, on the excess risk

$$\mathbb{E} \boldsymbol{L}(\hat{\boldsymbol{y}}) - \inf_{f \in \mathcal{F}} \boldsymbol{L}(f)$$

of an empirical risk minimizer $\hat{\boldsymbol{y}}$. Any *distribution-independent* upper bound also yields an upper bound on the minimax value $\mathcal{V}^{iid}(\mathcal{F})$, but the tools we developed in fact also give us data-dependent upper bounds. In a few lectures, we will show certain concentration arguments that allow us to *estimate* the supremum of the Rademacher process, opening a door to fully data-driven (rather than worst-case) results.

Since we have covered so many techniques, let us give a short recipe of how one can proceed with a new problem. To obtain an upper bound on the performance of empirical risk minimization, first pass to the supremum of the empirical process and then to the supremum of the Rademacher process, as in Theorem 7.10. Then use the properties of the Rademacher averages to simplify the class over which the Rademacher averages are taken, using Lemma 7.14 and Lemma 12.9. This can involve stripping away any Lipschitz compositions (e.g. loss function), and passing from convex hulls of functions to a set of vertices. Then attempt to build a covering, preferably in the ℓ_2 sense. If a fat-shattering or VC-dimension can be calculated, use it to immediately deduce the size of the cover, e.g. with the help

of Theorem 12.7. Use it in conjunction with the Dudley entropy integral bound of Theorem 12.4 to obtain the final bound. This development for an L-Lipschitz loss can be summarized by the following series of upper bounds:

$$\mathbb{E}L(\widehat{y}) - \inf_{f \in \mathcal{F}} L(f) \le \mathbb{E}\sup_{f \in \mathcal{F}} \underbrace{\{L(f) - \widehat{L}(f)\}}_{\mathbb{G}_{\ell(f)}} \le 2\mathbb{E}\underbrace{\left\{\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \mathbb{S}_{\ell(f)}\right\}}_{\widehat{\mathcal{R}}^{iid}(\ell(\mathcal{F}))} \le 2L\mathbb{E}\underbrace{\left\{\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \mathbb{S}_f\right\}}_{\widehat{\mathcal{R}}^{iid}(\mathcal{F})}$$

$$\le 2L\mathbb{E}\inf_{\alpha \ge 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1 \sqrt{\log \mathcal{N}_2(\mathcal{F}, \delta, x^n)}\, d\delta\right\} \le 2L\inf_{\alpha \ge 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1 \sqrt{K\text{fat}(\mathcal{F}, c\delta)\log\left(\frac{2}{\delta}\right)}\, d\delta\right\}$$

As the next section shows, the sequence of upper bounds is (in a certain sense) tight.

12.8 Supplementary Material: Back to the Rademacher

In Theorem 7.10, we showed that the supremum of an empirical process is within a factor of 2 from the supremum of the Rademacher process, both as an upper and a lower bound. Hence, not much is lost by working with the latter. We then upper bounded the supremum of the Rademacher process by the Dudley entropy integral $\widehat{\mathcal{D}}^{iid}(\mathcal{F})$. We now show that this upper bound is also quite tight, as it is within a logarithmic factor from the worst-case statistical Rademacher averages.

Definition 12.10. Worst-case statistical Rademacher averages are defined as

$$\overline{\mathcal{R}}^{iid}(\mathcal{F}, n) \triangleq \sup_{x_1, \dots, x_n} \widehat{\mathcal{R}}^{iid}(\mathcal{F}, x_1, \dots, x_n).$$

Lemma 12.11. *For any n and $\alpha > 2\overline{\mathcal{R}}^{iid}(\mathcal{F}, n)$,*

$$\text{fat}(\mathcal{F}, \alpha) \le \frac{8n\overline{\mathcal{R}}^{iid}(\mathcal{F}, n)^2}{\alpha^2} \tag{12.12}$$

Proof. Let $d = \text{fat}(\mathcal{F}, \alpha)$ and $x_1^*, \dots, x_d^*$ be the set of points α-shattered by $\mathcal{F}$ with the witness $y_1^*, \dots, y_d^*$. First, suppose by the way of contradiction that $d \ge n$. Then, the worst-case statistical Rademacher based on the first $(x_1^*, \dots, x_n^*)$ out of the shattered set gives

$$\overline{\mathcal{R}}^{iid}(\mathcal{F}, n) \ge \widehat{\mathcal{R}}^{iid}(\mathcal{F}, x_1, \dots, x_n) = \mathbb{E}_{\epsilon_1, \dots, \epsilon_n} \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^n \epsilon_t f(x_t^*) = \mathbb{E}\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^n \epsilon_t (f(x_t^*) - y_t^*) \ge \alpha/2,$$

contradicting our assumption that $\alpha > 2\overline{\mathcal{R}}^{iid}(\mathcal{F}, n)$. Hence, necessarily $d < n$.

For simplicity of exposition, assume that $n = kd$ for some integer $k > 0$ (otherwise, can take n' to be the nearest multiple of d, with $n \le n' \le 2n$). We can then take the set $(x_1,\ldots,x_n)$ as $(x_1^*,\ldots,x_1^*,\ldots,x_d^*,\ldots,x_d^*)$, where each element x_i^* is repeated k times, and perform the same concatenation for the witnesses $y_1,\ldots,y_n$. We then lower bound

$$\overline{\mathcal{R}}^{iid}(\mathcal{F}, n) \ge \widehat{\mathcal{R}}^{iid}(\mathcal{F}, x_1,\ldots,x_n)$$

$$= \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t (f(x_t) - y_t)$$

$$= \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{d} (f(x_i^*) - y_i^*) \left(\sum_{j=1}^{k} \epsilon_{(i-1)k+j} \right)$$

For a given sequence $(\epsilon_1,\ldots,\epsilon_n)$, let σ_i be the majority vote of the signs $\epsilon_{(i-1)k+1},\ldots,\epsilon_{ik}$ on the ith block. By the definition of α-shattering, there exists a function $f \in \mathcal{F}$ that passes above/below y_i^* on x_i^* by a margin $\alpha/2$ according to the sign σ_i, for each i. This yields

$$\overline{\mathcal{R}}^{iid}(\mathcal{F}, n) \ge \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{d} (\alpha/2) \left| \sum_{j=1}^{k} \epsilon_{(i-1)k+j} \right| \ge (\alpha/2) \sqrt{\frac{1}{2k}} = \sqrt{\frac{\alpha^2 \mathrm{fat}(\mathcal{F}, \alpha)}{8n}}$$

$\square$

With the upper bound on $\mathrm{fat}(\mathcal{F}, \alpha)$ given by Lemma 12.11, we can further upper bound the integral in Corollary 12.8 to obtain

Corollary 12.12. *For $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$, for any $x_1,\ldots,x_n$,*

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) \le \widehat{\mathcal{D}}^{iid}(\mathcal{F}) \le \overline{\mathcal{R}}^{iid}(\mathcal{F}, n) \cdot O(\log^{3/2} n) .$$

Proof. Taking $\alpha = 2\overline{\mathcal{R}}^{iid}(\mathcal{F}, n)/c$, for the absolute constant c given in (12.10). Then

$$\widehat{\mathcal{R}}^{iid}(\mathcal{F}) \le 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{K \mathrm{fat}(\mathcal{F}, c\delta) \log\left(\frac{2}{\delta}\right)} d\delta \le 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{K \frac{8n\overline{\mathcal{R}}^{iid}(\mathcal{F}, n)^2}{c^2 \delta^2} \log\left(\frac{2}{\delta}\right)} d\delta$$

Simplifying the above expression, we obtain an upper bound

$$\overline{\mathcal{R}}^{iid}(\mathcal{F}, n) \left(8/c + 12 \int_\alpha^1 \sqrt{\frac{8K}{c^2 \delta^2} \log\left(\frac{2}{\delta}\right)} d\delta \right) = \overline{\mathcal{R}}^{iid}(\mathcal{F}, n) \cdot O(\log^{3/2} n)$$

$\square$

12.9   Supplementary Material: Lower Bound on the Minimax Value

We now show a lower bound on the value of statistical learning with absolute loss (Eq. (5.9)) in terms of the worst-case statistical Rademacher averages $\overline{\mathcal{R}}^{iid}$ defined in the previous section, with the caveat that we only consider proper learning methods (see page 53).

Lemma 12.13 ([50])**.** *Consider the i.i.d. setting of supervised learning with absolute loss, and let $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$. Let us only consider estimators $\hat{\boldsymbol{y}}$ that take values in $\mathcal{F}$ (that is, proper learning). Then*

$$\mathcal{V}^{iid,ab}(\mathcal{F}, n) \geq \overline{\mathcal{R}}^{iid}(\mathcal{F}, 2n) - \frac{1}{2}\overline{\mathcal{R}}^{iid}(\mathcal{F}, n) \tag{12.13}$$

Proof. Recall that $|y - f(x)| = 1 - yf(x)$ for $y \in \{\pm 1\}$ and $f(x) \in [-1,1]$. Hence, we can write

$$\mathbb{E}\left|\hat{\boldsymbol{y}}(\text{X}) - \text{Y}\right| - \inf_{f \in \mathcal{F}} \mathbb{E}\left|f(\text{X}) - \text{Y}\right| = \mathbb{E}\left(1 - \text{Y}\hat{\boldsymbol{y}}(\text{X})\right) - \inf_{f \in \mathcal{F}} \mathbb{E}\left(1 - \text{Y}f(\text{X})\right)$$

$$= \sup_{f \in \mathcal{F}} \mathbb{E}\left(\text{Y}f(\text{X})\right) - \mathbb{E}\left(\text{Y}\hat{\boldsymbol{y}}(\text{X})\right)$$

where the expectation in the second term is over the draw of data (of size n) as well as another i.i.d. copy (X, Y). Take any $x_1, \ldots, x_{2n} \in \mathcal{X}$ and let P_X be the uniform distribution on these $2n$ points. For any $\epsilon = (\epsilon_1, \ldots, \epsilon_{2n}) \in \{\pm 1\}^{2n}$, let the distribution P_ϵ be such that the marginal distribution of the variable X is P_X while the conditional of Y|X $= x_i$ is deterministically ϵ_i, for all $i \in [2n]$. This defines a family of distributions P_ϵ indexed by $\epsilon \in \{\pm 1\}^{2n}$. Recall that, by definition, an estimator $\hat{\boldsymbol{y}}$ is a mapping n samples drawn i.i.d. from this discrete distribution into $\mathcal{F}$. Then

$$\mathcal{V}^{iid,ab}(\mathcal{F}, n) = \inf_{\hat{\boldsymbol{y}}} \sup_{P} \left\{ \mathbb{E}\left|\hat{\boldsymbol{y}}(\text{X}) - \text{Y}\right| - \inf_{f \in \mathcal{F}} \mathbb{E}\left|f(\text{X}) - \text{Y}\right| \right\}$$

$$\geq \inf_{\hat{\boldsymbol{y}}} \sup_{x_1, \ldots, x_{2n} \in \mathcal{X}} \mathbb{E}_\epsilon \left\{ \sup_{f \in \mathcal{F}} \mathbb{E}\left(\text{Y}f(\text{X})\right) - \mathbb{E}\left(Y\hat{\boldsymbol{y}}(\text{X})\right) \right\}$$

$$\geq \left\{ \sup_{x_1, \ldots, x_{2n} \in \mathcal{X}} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \mathbb{E}\left(\text{Y}f(\text{X})\right) \right\} - \left\{ \sup_{\hat{\boldsymbol{y}}} \sup_{x_1, \ldots, x_{2n} \in \mathcal{X}} \mathbb{E}_\epsilon \mathbb{E}\left(Y\hat{\boldsymbol{y}}(\text{X})\right) \right\}$$

The first term above is precisely

$$\sup_{x_1,\dots,x_{2n}\in\mathcal{X}} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \mathbb{E}\left(\mathrm{Y} f(\mathrm{X})\right) = \sup_{x_1,\dots,x_{2n}\in\mathcal{X}} \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}}\left\{\frac{1}{2n}\sum_{t=1}^{2n}\epsilon_t f(x_t)\right\} = \overline{\mathcal{R}}^{iid}(\mathcal{F},2n).$$

For the second term, note that a uniform draw of n data points from $2n$ with replacement is equivalent to drawing indices $i_1,\dots,i_n$ uniformly at random from $\{1,\dots,2n\}$, and let J stand for the set of unique indices, $|J|\le n$. We can then write the second term as

$$\sup_{\widehat{y}}\ \sup_{x_1,\dots,x_{2n}\in\mathcal{X}} \mathbb{E}_\epsilon\mathbb{E}\left(Y\widehat{y}(\mathrm{X})\right) = \sup_{\widehat{y}}\ \sup_{x_1,\dots,x_{2n}\in\mathcal{X}} \mathbb{E}_{i_1,\dots,i_n}\mathbb{E}_\epsilon\left\{\frac{1}{2n}\sum_{t=1}^{2n}\epsilon_t\widehat{y}(x_t)\right\}$$

$$= \sup_{\widehat{y}}\ \sup_{x_1,\dots,x_{2n}\in\mathcal{X}} \mathbb{E}_{i_1,\dots,i_n}\mathbb{E}_\epsilon\left\{\frac{1}{2n}\sum_{t\in J}\epsilon_t\widehat{y}(x_t)\right\}$$

where the last equality is due to the fact that

$$\mathbb{E}_\epsilon\left\{\frac{1}{2n}\sum_{t\notin J}\epsilon_t\widehat{y}(x_t)\right\} = 0.$$

However, we can upper bound

$$\sup_{\widehat{y}}\ \sup_{x_1,\dots,x_{2n}\in\mathcal{X}} \mathbb{E}_{i_1,\dots,i_n}\mathbb{E}_\epsilon\left\{\frac{1}{2n}\sum_{t\in J}\epsilon_t\widehat{y}(x_t)\right\} \le \sup_{\widehat{y}}\mathbb{E}_{i_1,\dots,i_n}\ \sup_{x_1,\dots,x_{|J|}\in\mathcal{X}}\mathbb{E}_\epsilon\left\{\frac{1}{2n}\sum_{t\in J}\epsilon_t\widehat{y}(x_t)\right\}$$

$$\le \sup_{\widehat{y}}\ \sup_{x_1,\dots,x_n\in\mathcal{X}}\mathbb{E}_\epsilon\left\{\frac{1}{2n}\sum_{t=1}^{n}\epsilon_t\widehat{y}(x_t)\right\}$$

$$\le \sup_{x_1,\dots,x_n\in\mathcal{X}}\mathbb{E}_\epsilon\sup_{f\in\mathcal{F}}\left\{\frac{1}{2n}\sum_{t=1}^{n}\epsilon_t f(x_t)\right\}$$

which is $\frac{1}{2}\overline{\mathcal{R}}^{iid}(\mathcal{F},n)$. This concludes the proof. $\qquad\square$

13
Sequential Prediction: Classification

We now turn to the setting of sequential prediction, with the goal of mimicking many of the results of the previous few lectures. Within sequential prediction, "n-tuples of data" are replaced with "trees", the empirical process with the martingale process, and the Rademacher process – with the tree process. We already observed certain similarities in the way that the suprema over finite classes are bounded both in the i.i.d. and sequential cases. We also saw that the simple example of thresholds on an interval is "easy" for i.i.d. learning, but "hard" for sequential prediction, constituting an apparent divide between the two settings. How much of the i.i.d.-style analysis can be pushed through for the sequential prediction, and what are the key differences? This is the subject of the next few lectures.[1]

In particular, we first focus on the *supervised* scenario. Let us consider the improper learning version: at each round t, the learner observes $x_t \in \mathcal{X}$, makes a prediction $\widehat{y}_t$, and observes the outcome y_t which is chosen by Nature simultaneously with the learner's choice. The sequence of x_t's and y_t's is chosen by Nature, possibly adaptively based on the moves $\widehat{y}_1, \ldots, \widehat{y}_{t-1}$ of the player. The classification problem involves prediction of a binary label $y_t \in \{0, 1\}$, and the goal of the learner is to have small regret

$$\frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{\widehat{y}_t \neq y_t\} - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{f(x_t) \neq y_t\}$$

against some class $\mathcal{F}$ of binary-valued functions. The indicator loss can be equivalently written as the absolute value of the difference. According to Theorems 7.9

123

and 7.11, the value of the sequential prediction problem is upper bounded by

$$\mathcal{V}^{seq}(\mathcal{F}, n) \leq 2 \sup_{\mathbf{x}, \mathbf{y}} \mathbb{E} \sup_{g \in \ell(\mathcal{F})} \mathbb{T}_g = 2 \sup_{\mathbf{x}, \mathbf{y}} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{I}\left\{ f(\mathbf{x}_t(\epsilon_{1:t-1})) \neq \mathbf{y}_t(\epsilon_{1:t-1}) \right\}$$

(13.1)

Just as we did in Eq. (8.4), we pass to the tree process on $\mathcal{F}$ itself, rather than the loss class. This step is a version of the contraction principle as stated in Lemma 12.9 for the i.i.d. case.

Lemma 13.1. *For* $\mathcal{Y} = \{0, 1\}$, $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, *and* $\ell(f, (x, y)) = \mathbf{I}\left\{ f(x) \neq y \right\}$,

$$\sup_{\mathbf{x}, \mathbf{y}} \mathbb{E} \sup_{g \in \ell(\mathcal{F})} \mathbb{T}_g = \sup_{\mathbf{x}} \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{T}_f$$

(13.2)

With the definition of sequential Rademacher complexity, the statement can be written more succinctly as

$$\mathcal{R}^{seq}(\ell(\mathcal{F})) = \mathcal{R}^{seq}(\mathcal{F}).$$

A proof of this lemma requires a few intermediate steps, and postponing it until the end of the lecture seems like a good idea.

13.1   From Finite to Infinite Classes: First Attempt

Lemma 9.5 gives us a handle on the supremum of the tree process for a finite class $\mathcal{F}$. Mimicking the development for the i.i.d. case, we would like to now encompass infinite classes as well. The threshold example is not interesting since the tree process does not converge to zero, according to Theorem 8.1. Consider a different example instead. The example is rather trivial, but will serve us for the purposes of demonstration. Let $\mathcal{X} = [0, 1]$ and

$$\mathcal{F} = \{f_a : a \in [0, 1], \ f_a(x) = 0 \ \forall x \neq a, \ f_a(a) = 1\}$$

(13.3)

the class of functions that are zero everywhere except on one designated point. This is an infinite class and the question is whether the expected supremum of the tree process decays to zero with increasing n for any $\mathcal{X}$-valued tree **x**. Now, recall our development for the i.i.d. case with the class of thresholds. We argued that if we condition on the data, the effective number of possible values that the

13.1 From Finite to Infinite Classes: First Attempt

functions take is finite even if the class is infinite. This led us to the idea of the growth function.

Following the analogy, let us define $\mathcal{F}|_{\mathbf{x}}$ as the set of all $\{0,1\}$-valued trees

$$\mathcal{F}|_{\mathbf{x}} = \{f \circ \mathbf{x} : f \in \mathcal{F}\}$$

where $f \circ \mathbf{x}$ is defined as a tree $(f \circ \mathbf{x}_1, f \circ \mathbf{x}_2, \ldots, f \circ \mathbf{x}_n)$. Since $\mathbf{x}_t$ is a function from $\{\pm 1\}^{t-1}$ to $\mathcal{X}$, the t-th level $f \circ \mathbf{x}_t$ of the $f \circ \mathbf{x}$ tree is a function from $\{\pm 1\}^{t-1}$ to $\{0,1\}$. Thinking of $\mathcal{F}|_{\mathbf{x}}$ as the projection (or, an "imprint") of $\mathcal{F}$ on $\mathbf{x}$, we can write

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1})) = \mathbb{E} \max_{\mathbf{v} \in \mathcal{F}|_{\mathbf{x}}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon_{1:t-1}) \tag{13.4}$$

because the set $\mathcal{F}|_{\mathbf{x}}$ is clearly finite.

Given $\mathbf{x}$, how can we describe $\mathcal{F}|_{\mathbf{x}}$? For any $f_a \in \mathcal{F}$, $f_a \circ \mathbf{x}$ is a tree that is zero everywhere except for those (if any) nodes in the tree which are equal a. Observe that for two functions $f_a, f_b \in \mathcal{F}$, $f_a \circ \mathbf{x} = f_b \circ \mathbf{x}$ if and only if $a, b \in [0,1]$ both do not appear in the tree $\mathbf{x}$:

$$f_a \circ \mathbf{x} = f_b \circ \mathbf{x} \quad \Leftrightarrow \quad a, b \notin \text{Img}(\mathbf{x}) \text{ or } a = b.$$

Suppose $\mathbf{x}$ is such that $\mathbf{x}_n : \{\pm 1\}^{n-1} \mapsto [0,1]$ takes on 2^{n-1} distinct values for each path. Then

$$\text{card}(\mathcal{F}|_{\mathbf{x}}) = 2^{n-1}.$$

Since this cardinality is exponential in n, Lemma 9.5 gives a vacuous bound.

It appears that there is no hope for passing from a finite to an infinite class by studying the size of the projection of $\mathcal{F}$ on $\mathbf{x}$, as it will be exponential in n for any nontrivial class and a "diverse enough" $\mathbf{x}$. But maybe this is for a good reason, and the problem is not learnable, just like in the case of thresholds? This turns out to be not the case:

✎ **Exercise 13.1** (⋆⋆)**.** Provide a strategy for the learner that will suffer $O(\sqrt{(\log n)/n})$ regret for the binary prediction problem with the class defined in (13.3). Can you get a $O(1/\sqrt{n})$ algorithm? (This should be doable.)

Of course, one may hypothesize that the problem might be learnable but the bound of Theorem 7.9 is loose. This is again not the case, and one can prove that the expected supremum of the tree process indeed converges to zero for any $\mathbf{x}$, but $\mathcal{F}|_{\mathbf{x}}$ is a wrong quantity to consider.

13.2   From Finite to Infinite Classes: Second Attempt

As we have already noticed, the cardinality of the set $\mathcal{F}|_{\mathbf{x}}$ is too large. However, this does not reflect the fact that every $f_a \circ \mathbf{x}$, in the case of the function class defined in (13.3), is quite "simple": the values are zero, except when $\mathbf{x}_t = a$.

First, for illustration purposes, consider the case when the tree $\mathbf{x}$ contains unique elements along any path (left-most tree in Figure 13.1). That is, for any $(\epsilon_1, \dots, \epsilon_n)$, if $\mathbf{x}_t(\epsilon_{1:t-1}) = \mathbf{x}_s(\epsilon_{1:s-1})$ then $s = t$. In this case, $f \circ \mathbf{x}$ is a tree with at most a single value 1 along any path. Consider the set $V = \{\mathbf{v}^{(0)}, \mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)}\}$ of $n+1$ binary-
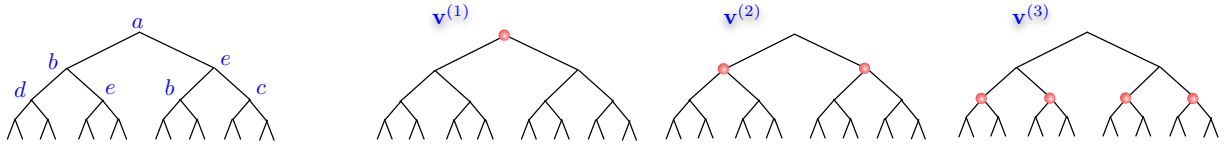
Figure 13.1: Left: an $\mathcal{X}$-valued tree $\mathbf{x}$ with unique elements within any path. Right: first three of the $n+1$ covering trees. Circled nodes are defined to be 1 while the rest are 0.

valued trees defined as follows. The tree $\mathbf{v}^{(0)}$ is identically 0-valued. For $j \geq 1$, the tree $\mathbf{v}^{(j)}$ has zeros everywhere except for the j-th level. That is, for any $j \in \{1, \dots, n\}$, $\mathbf{v}_t^{(j)}(\epsilon_{1:t-1}) = 1$ whenever $t = j$ and zero otherwise. These trees are depicted in Figure 13.1. Given the uniqueness of values of $\mathbf{x}$ along any path, we have the following important property:

$$\forall f \in \mathcal{F}, \ \forall (\epsilon_1, \dots, \epsilon_n) \in \{\pm 1\}^n, \ \exists \mathbf{v} \in V \ \text{s.t.} \ f(\mathbf{x}_t(\epsilon_{1:t-1})) = \mathbf{v}_t(\epsilon_{1:t-1}) \ \forall t \in \{1, \dots, n\}$$
(13.5)

That is, for any $f \in \mathcal{F}$ and any path, there exists a "covering tree" $\mathbf{v}$ in V such that f on $\mathbf{x}$ agrees with $\mathbf{v}$ on the given path. For instance, consider a function $f_d \in \mathcal{F}$ which takes values zero everywhere except on $x = d$. Consider the left-most path $\epsilon = (-1, -1, -1, \dots)$. Then, for the $\mathbf{x}$ given in Figure 13.1, the function takes on a value 1 at the third node, since $\mathbf{x}_3(-1, -1) = d$, and zero everywhere else. But then the covering tree $\mathbf{v}^{(3)}$ in Figure 13.1 provides exactly these values on the path $(-1, -1, -1, \dots)$. It is not hard to convince oneself that the property (13.5) holds

true. What is crucial is that for such a tree $\mathbf{x}$,

$$\mathcal{R}^{seq}(\mathcal{F}, \mathbf{x}) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1})) = \mathbb{E} \max_{\mathbf{v} \in V} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \mathbf{v}_t(\epsilon_{1:t-1}) \leq \sqrt{\frac{2\log(n+1)}{n}}$$

(13.6)

by Lemma 9.5 since $\text{card}(V) = n+1$. While this achieves our goal, the argument crucially relied on the fact that $\mathbf{x}$ contains unique elements along any path. What if $\mathbf{x}$ has several identical elements a along some path? In this case, the function f_a will take on the value 1 on multiple nodes on this path, and the set V defined above no longer does the job. A straightforward attempt to create a set V with all possible subsets of rows (taking on the value 1) will fail, as there is an exponential number of such possibilities.

It is indeed remarkable that there exists a set of $\{0, 1\}$-valued trees V of cardinality at most $n+1$ that satisfies property (13.5) without the assumption of uniqueness of elements along the paths. Here is how such a set can be constructed inductively. Suppose we have at our disposal two sets V^ℓ and V^r of covering trees of depth $n-1$
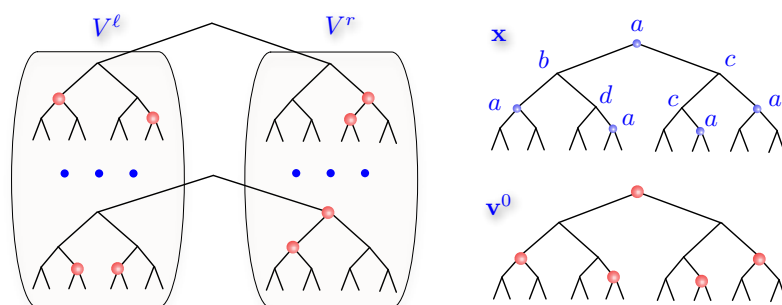
Figure 13.2: V is constructed inductively by pairing up trees in V^ℓ and V^r, plus an additional tree $\mathbf{v}^0$ (bottom right) which takes on value 1 only when $\mathbf{x}$ (top right) at the corresponding node takes on the value $\mathbf{x}_1$.

for the left and right subtrees of $\mathbf{x}$ at the root. Suppose that, inductively, on the two subtrees the sets V^ℓ and V^r satisfy property (13.5). For a $\mathbf{v}^\ell \in V^\ell$ and $\mathbf{v}^r \in V^r$ define a joined tree $\mathbf{v}$ as having 0 at the root and $\mathbf{v}^\ell$ and $\mathbf{v}^r$ as the two subtrees at the root (see Figure 13.2). In this fashion, take pairs from both sets such that each element of V^ℓ and V^r occurs in at least one pair. This construction gives a set V of size $\max\{\text{card}(V^\ell), \text{card}(V^r)\}$. We add to V one more tree $\mathbf{v}^0$ defined as zero

everywhere except on those nodes where **x** has the same value as the root $\mathbf{x}_1$ (in Figure 13.2, $\mathbf{x}_1 = a$ and $\mathbf{v}^0$ is constructed accordingly). That is,

$$\mathbf{v}_t^0(\epsilon_{1:t-1}) = 1 \text{ if } \mathbf{x}_t(\epsilon_{1:t-1}) = \mathbf{x}_1 \text{ and } \mathbf{v}_t^0(\epsilon_{1:t-1}) = 0 \text{ otherwise}$$

We claim that the set V satisfies (13.5) and its size is larger than that of V^{ℓ} and V^r by at most 1. Indeed, for $f_{\mathbf{x}_1} \in \mathcal{F}$ the tree $\mathbf{v}^0$ matches the values along any path. For other $f \in \mathcal{F}$, the value at the root is zero, and (13.5) holds by induction on both subtrees.

The size of V increases by at most one when passing from depth $n-1$ to n. For the base of the induction, we require 2 trees for $n = 1$: one with the value 0 and one with the value 1 at the single vertex. We conclude that the size of the set satisfying (13.5) is at most $n+1$. As a consequence, the bound of (13.6) holds for any **x**. We conclude that for the example under the consideration with the function class $\mathcal{F}$ defined in (13.3),

$$\mathcal{V}^{seq}(\mathcal{F}, n) \leq 2 \sup_{\mathbf{x}} \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{T}_f \leq 2\sqrt{\frac{2\log(n+1)}{n}} \tag{13.7}$$

13.3 The Zero Cover and the Littlestone's Dimension

To summarize the development so far, we have seen that the size of the projection $\mathcal{F}|_{\mathbf{x}}$ is not the right quantity to consider. However, if we can find a set V of binary-valued trees such that property (13.5) holds, the supremum over the class $\mathcal{F}$ is equal to the maximum over the trees in V. Let us make this statement more formally.

Definition 13.2. A set V of $\{0, 1\}$-valued trees of depth n is called a *zero-cover* of $\mathcal{F}$ on a given $\mathcal{X}$-valued tree **x** if

$$\forall f \in \mathcal{F}, \ \forall(\epsilon_1, \ldots, \epsilon_n) \in \{\pm 1\}^n, \ \exists \mathbf{v} \in V \text{ s.t. } f(\mathbf{x}_t(\epsilon_{1:t-1})) = \mathbf{v}_t(\epsilon_{1:t-1}) \ \forall t \in \{1, \ldots, n\}$$

A *zero-covering number* is defined as

$$\mathcal{N}_0(\mathcal{F}, \mathbf{x}) = \min\Big\{\text{card}(V) \ : \ V \text{ is an zero-cover}\Big\}.$$

Recall that in Section 12.4 we defined an i.i.d. zero cover on tuples of points and observed that it is the same as the projection $\mathcal{F}|_{x^n}$. For trees, however, the zero

cover is different from the projection $\mathcal{F}|_{\mathbf{x}}$. This is a key departure from the non-tree definitions. Observe that the difference is really in the order of quantifiers: the covering tree $\mathbf{v}$ can be chosen according to the given path $(\epsilon_1,\dots,\epsilon_n)$. It is not required that there exists a tree $\mathbf{v}$ equal to $f \circ \mathbf{x}$ on *every* path, but rather that we can find some $\mathbf{v}$ for any path. It is indeed because of this correct order of quantifiers that we are able to control the supremum of the tree process even though $\mathcal{F}|_{\mathbf{x}}$ is an exponentially large set.

We have the following analogue of Proposition 11.2:

Proposition 13.3. *Let $\mathbf{x}$ be an $\mathcal{X}$-valued tree of depth n, and suppose $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$. Then*

$$\mathcal{R}^{seq}(\mathcal{F},\mathbf{x}) \triangleq \mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{x}_t(\epsilon_{1:t-1})) \le \sqrt{\frac{2\log\mathcal{N}_0(\mathcal{F},\mathbf{x})}{n}}$$

In a somewhat surprising turn, we can define a combinatorial dimension analogous to the Vapnik-Chervonenkis dimension, which controls the size of the zero-cover in exactly the same way that the VC dimension controls the size of the zero cover (Lemma 11.5).

Definition 13.4. We say that $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ *shatters* a tree $\mathbf{x}$ of depth n if

$$\forall(\epsilon_1,\dots,\epsilon_n)\in\{\pm 1\}^n, \ \exists f\in\mathcal{F} \ \text{ s.t. } \ f(\mathbf{x}_t(\epsilon_{1:t-1})) = \tilde{\epsilon}_t \ \forall t\in\{1,\dots,n\}$$

where $\tilde{\epsilon}_t = (\epsilon_t + 1)/2 \in \{0,1\}$

Definition 13.5. The *Littlestone's dimension* $\mathrm{ldim}(\mathcal{F})$ of the class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ is defined as the depth of the largest complete binary $\mathcal{X}$-valued tree $\mathbf{x}$ that is shattered by $\mathcal{F}$, and infinity if the maximum does not exist. Further, for a given $\mathbf{x}$, define $\mathrm{ldim}(\mathcal{F},\mathbf{x})$ as the depth of the largest complete binary tree with values in $\mathrm{Img}(\mathbf{x})$.

It is easy to see that the Littlestone's dimension becomes the VC dimension if the trees $\mathbf{x}$ are only allowed to have constant mappings $\mathbf{x}_t(\epsilon_{1:t-1}) = x_t$ for all $\epsilon_{1:t-1}$. Hence, we have the rather easy lemma:

Lemma 13.6. *For any $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$,*

$$vc(\mathcal{F}) \le ldim(\mathcal{F})$$

Example 15. The Littlestone's dimension of the class of thresholds $\mathcal{F} = \{f_\theta(x) = \mathbf{I}\{x \le \theta\} : \theta \in [0,1]\}$ is infinite. Indeed, a $[0,1]$-valued tree $\mathbf{x}$ of arbitrary depth can be constructed as in Figure 8.3, and it is shattered by $\mathcal{F}$. In contrast, $vc(\mathcal{F}) = 1$.

Example 16. The Littlestone's dimension of the class defined in (13.3) is 1.

We are now ready to state an analogue of the Vapnik-Chervonenkis-Sauer-Shelah Lemma:

Theorem 13.7. *For any* $\mathbf{x}$,

$$\mathcal{N}_0(\mathcal{F},\mathbf{x}) \leq \sum_{i=0}^{d} \binom{n}{i}$$

whenever $ldim(\mathcal{F}) = d < \infty$.

With the help of Eq. 11.6, we have the following analogue of Corollary 11.6:

Corollary 13.8. *Under the setting of Proposition 13.3, if* $ldim(\mathcal{F}) = d$,

$$\mathcal{R}^{seq}(\mathcal{F}) \leq 2\sqrt{\frac{2d\log(en/d)}{n}}$$

Proof of Theorem 13.7. As in the proof of Lemma 11.5, define the function $g(d,n) = \sum_{i=0}^{d}\binom{n}{i}$. The theorem claims that the size of a minimal zero-cover is at most $g(d,n)$. The proof proceeds by induction on $n + d$.

Base: For $d = 1$ and $n = 1$, there is only one node in the tree, i.e. the tree is defined by the constant $\mathbf{x}_1 \in \mathcal{X}$. Functions in $\mathcal{F}$ can take at most two values on $\mathbf{x}_1$, so $\mathcal{N}_0(\mathcal{F},\mathbf{x}) \leq 2$. Using the convention $\binom{n}{0} = 1$, we indeed verify that $g(1,1) = 2$. The same calculation gives the base case for $n = 1$ and any $d \in \mathbb{N}$. Furthermore, for any $n \in \mathbb{N}$ if $d = 0$, then there is no point which is 1-shattered by $\mathcal{F}$. This means that all functions in $\mathcal{F}$ are identical, proving that there is a zero-cover of size $1 = g(0,n)$.

Induction step: Suppose by the way of induction that the statement holds for $(d,n-1)$ and $(d-1,n-1)$. Consider any tree $\mathbf{x}$ of depth n with $ldim(\mathcal{F},\mathbf{x}) = d$. Define the *partition* $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$ with $\mathcal{F}_i = \{f \in \mathcal{F} : f(\mathbf{x}_1) = i\}$ for $i \in \{0,1\}$. We first argue that it cannot be the case that $ldim(\mathcal{F}_0,\mathbf{x}) = ldim(\mathcal{F}_1,\mathbf{x}) = d$. For, otherwise there exist two trees $\mathbf{z}$ and $\mathbf{w}$ of depth d, shattered by $\mathcal{F}_0$ and $\mathcal{F}_1$, respectively, and with $Img(\mathbf{z}), Img(\mathbf{w}) \subseteq Img(\mathbf{x})$. Since functions within $\mathcal{F}_0$ take on the same value on $\mathbf{x}_1$, we conclude that $\mathbf{x}_1 \notin Img(\mathbf{z})$ (this follows immediately from the definition of shattering). Similarly, $\mathbf{x}_1 \notin Img(\mathbf{w})$. We now *join* the two shattered $\mathbf{z}$ and $\mathbf{w}$ trees with $\mathbf{x}_1$ at the root and observe that $\mathcal{F}_0 \cup \mathcal{F}_1$ shatters this resulting tree of depth $d + 1$, which is a contradiction. We conclude that $ldim(\mathcal{F}_i,\mathbf{x}) = d$ for at most one $i \in \{0,1\}$.
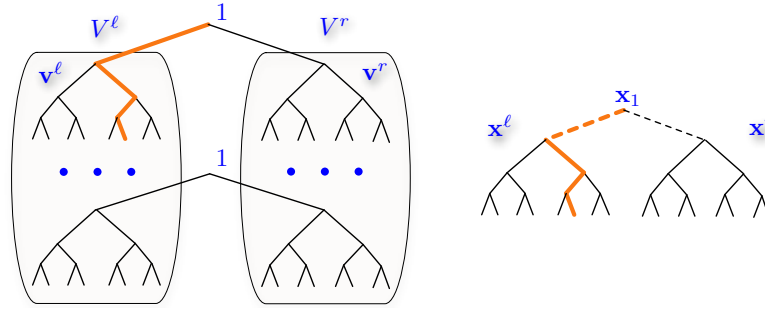
Figure 13.3: The joined trees provide a zero cover for $\mathcal{F}_1$. For any path, the induction hypothesis provides a zero-cover on the subtree, as illustrated.

Without loss of generality, assume $\text{ldim}(\mathcal{F}_0, \mathbf{x}) \le d$ and $\text{ldim}(\mathcal{F}_1, \mathbf{x}) \le d - 1$. By induction, there are zero-covers V^ℓ and V^r of $\mathcal{F}_1$ on the subtrees $\mathbf{x}^\ell$ and $\mathbf{x}^r$, respectively, both of size at most $g(d-1, n-1)$. Out of these covers we can create a zero-cover V for $\mathcal{F}_1$ on $\mathbf{x}$ by pairing the covering trees in V^ℓ and V^r. Formally, consider a set of pairs $(\mathbf{v}^\ell, \mathbf{v}^r)$ of trees, with $\mathbf{v}^\ell \in V^\ell$, $\mathbf{v}^r \in V^r$ and such that each tree in V^ℓ and V^r appears in at least one of the pairs. Clearly, this can be done using at most $g(d-1, n-1)$ pairs, and such a pairing is not unique. We join the subtrees in every pair $(\mathbf{v}^\ell, \mathbf{v}^r)$ with a constant 1 as the root, thus creating a set V of trees, $\text{card}(V) \le g(d-1, n-1)$. We claim that V is a zero cover for $\mathcal{F}_1$ on $\mathbf{x}$ (see Figure 13.3). Note that all the functions in $\mathcal{F}_1$ take on the same value 1 on $\mathbf{x}_1$ and by construction $\mathbf{v}_1 = 1$ for any $\mathbf{v} \in V$. Now, consider any $f \in \mathcal{F}_1$ and $\epsilon \in \{\pm 1\}^n$. Without loss of generality, assume $\epsilon_1 = -1$. By assumption, there is a $\mathbf{v}^\ell \in V^\ell$ such that $\mathbf{v}_t^\ell(\epsilon_{2:t}) = f(\mathbf{x}_{t+1}(\epsilon_{1:t}))$ for any $t \in \{1, \dots, n-1\}$. By construction $\mathbf{v}^\ell$ appears as a left subtree of at least one tree in V, which, therefore, matches the values of f for $\epsilon_{1:n}$. The same argument holds for $\epsilon_1 = +1$ by finding an appropriate subtree in V^r. We conclude that V is a zero cover of $\mathcal{F}_1$ on $\mathbf{x}$. A similar argument yields a zero cover for $\mathcal{F}_0$ on $\mathbf{x}$ of size at most $g(d, n-1)$ by induction. Thus, the size of the resulting zero cover of $\mathcal{F}$ on $\mathbf{x}$ is at most

$$g(d, n-1) + g(d-1, n-1) = g(d, n),$$

completing the induction step and yielding the statement of the theorem. $\qquad \square$

13.4 Removing the Indicator Loss, or Fun Rotations with Trees

Finiteness of the Littlestone's dimension is necessary and sufficient for the expected supremum of the tree process for the worst-case tree to converge to zero. However, for the study of sequential prediction in the supervised setting, it remains to connect the tree process indexed by $\mathcal{F}$ to that indexed by $\ell(\mathcal{F})$ for the indicator loss. This is the contents of Lemma 13.1, and we now provide the proof.

We start by proving two supporting lemmas, which can be found in [46]. To motivate the first lemma, take a fixed vector $s \in \{\pm 1\}^n$, and let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher random variables. Then, surely, the vector $(\epsilon_1 s_1, \ldots, \epsilon_n s_n)$ is a vector of i.i.d. Rademacher random variables. Surprisingly, Lemma 13.9 below says that such a statement can be pushed much further: s_t does not need to be fixed but can be chosen according to $(\epsilon_1, \ldots, \epsilon_{t-1})$. This exactly means that for any $\{\pm 1\}$-valued tree $\mathbf{s}$, the sequence $(\epsilon_1 \mathbf{s}_1, \epsilon_2 \mathbf{s}_2(\epsilon_1), \ldots, \epsilon_n \mathbf{s}_n(\epsilon_{1:n-1}))$ is still i.i.d. Rademacher. The result can also be interpreted as a bijection between the set of sign sequences $(\epsilon_1, \ldots, \epsilon_n)$ and the set of sign sequences $(\epsilon_1 \mathbf{s}_1, \ldots, \epsilon_n \mathbf{s}_n(\epsilon_{1:n-1}))$.

For vectors $a, b \in \{\pm 1\}^n$, let the operation $a \star b$ denote element-wise multiplication. Further, for a $\{\pm 1\}$-valued tree $\mathbf{s}$, let $\mathbf{s}_{1:n}(\epsilon) = (\mathbf{s}_1(\epsilon), \ldots, \mathbf{s}_n(\epsilon))$ denote the vector of elements on the path $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. Let us recall the convention that we write $\mathbf{x}_t(\epsilon)$ even though $\mathbf{x}_t$ only depends on $\epsilon_{1:t-1}$.

Lemma 13.9. *Let $\mathbf{s}$ be any $\{\pm 1\}$-valued tree and let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ be a vector of i.i.d. Rademacher variables. Then the elements of the vector $\epsilon \star \mathbf{s}_{1:n}(\epsilon)$ are also i.i.d. Rademacher.*

Proof. We just need to show that the set of sign patterns $P(\mathbf{s}) = \{\epsilon \star \mathbf{s}_{1:n}(\epsilon) : \epsilon \in \{\pm 1\}^n\}$ is the same as all sign patterns, i.e. $P(\mathbf{s}) = \{\pm 1\}^n$. This is easy to do by induction on n. Lemma is obvious for $n = 1$. Assume it for $n = k$ and let us prove it for $n = k + 1$. Fix a tree $\mathbf{s}$ of depth $k + 1$. Let $\mathbf{s}^L$ and $\mathbf{s}^R$ be the left and right subtrees (of depth k), i.e.

$$\mathbf{s}_t^L(\epsilon) = \mathbf{s}_{t+1}(-1, \epsilon), \quad \mathbf{s}_t^R(\epsilon) = \mathbf{s}_{t+1}(+1, \epsilon)$$

Now, by definition of $P(\mathbf{s})$, we have

$$P(\mathbf{s}) = \{(-\mathbf{s}_1, \mathbf{b}) : \mathbf{b} \in P(\mathbf{s}^L)\} \cup \{(+\mathbf{s}_1, \mathbf{b}) : \mathbf{b} \in P(\mathbf{s}^R)\}$$

where $\mathbf{s}_1$ is simply the root of $\mathbf{s}$. Invoking induction hypothesis this is equal to

$$P(\mathbf{s}) = \{(-\mathbf{s}_1, \mathbf{b}) : \mathbf{b} \in \{\pm 1\}^k\} \cup \{(+\mathbf{s}_1, \mathbf{b}) : \mathbf{b} \in \{\pm 1\}^k\}$$

and thus $P(\mathbf{s}) = \{\pm 1\}^{k+1}$ no matter what $\mathbf{s}_1 \in \{+1, -1\}$ is. $\qquad \square$

Lemma 13.9 can now be used to prove the following result.

Lemma 13.10. *For any $\mathcal{X}$-valued tree $\mathbf{x}$ and $\{\pm 1\}$-valued tree $\mathbf{s}$, there exists another instance tree $\mathbf{x}'$ such that*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \mathbf{s}_t(\epsilon) f(\mathbf{x}_t(\epsilon))\right] = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}'_t(\epsilon))\right] .$$

Specifically, $\mathbf{x}'$ is defined by $\mathbf{x}_t(\epsilon) = \mathbf{x}'_t(\epsilon \star \mathbf{s}_{1:n}(\epsilon))$ for $t = 1, \ldots, n$.

Proof. First, $\mathbf{x}'$ is well-defined because $\left\{\epsilon \star \mathbf{s}_{1:n}(\epsilon) : \epsilon \in \{\pm 1\}^n\right\} = \{\pm 1\}^n$ by Lemma 13.9. Next, for any f,

$$\sum_{t=1}^{n} \epsilon_t \mathbf{s}_t(\epsilon) f(\mathbf{x}_t(\epsilon)) = \sum_{t=1}^{n} \epsilon_t \mathbf{s}_t(\epsilon) f(\mathbf{x}'_t(\epsilon \star \mathbf{s}_{1:n}(\epsilon))) .$$

Taking sup over f's followed by expectation over ϵ's on both sides gives us

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \mathbf{s}_t(\epsilon) f(\mathbf{x}_t(\epsilon))\right] = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} [\epsilon \star \mathbf{s}_{1:n}(\epsilon)]_t f(\mathbf{x}'_t(\epsilon \star \mathbf{s}_{1:n}(\epsilon)))\right]$$

The proof is complete now by appealing to Lemma 13.9 that asserts that the distribution of $\epsilon \star \mathbf{s}_{1:n}(\epsilon)$ is also i.i.d. Rademacher no matter what the signed tree $\mathbf{s}$ is. $\qquad \square$

One can interpret the statement of Lemma 13.10 as an equality due to the "rotation" of the tree $\mathbf{x}$ provided by the tree $\mathbf{s}$ of signs. For instance, suppose for the illustration purposes that $\mathbf{s}$ is a tree of all -1's. Hence,

$$\epsilon_t \mathbf{s}_t(\epsilon) f(\mathbf{x}_t(\epsilon)) = -\epsilon_t f(\mathbf{x}_t(\epsilon)) = -\epsilon_t f(\mathbf{x}'_t(-\epsilon))$$

where $\mathbf{x}'$ is the mirror reflection of $\mathbf{x}$ defined as $\mathbf{x}'_t(\epsilon_{1:t-1}) = \mathbf{x}_t(-\epsilon_{1:t-1})$. It is then not surprising that Lemma 13.10 holds for this particular choice of $\mathbf{s}$. As another example, take the tree $\mathbf{s}$ with -1 at the root and $+1$ everywhere else. Then the resulting tree $\mathbf{x}'$ has the left and right subtrees of $\mathbf{x}$ reversed. What is interesting, the result of Lemma 13.10 holds for *any* $\{\pm 1\}$-valued tree $\mathbf{s}$, and this is exactly what we need to "erase" the indicator loss.

Proof of Lemma 13.1. Fix any $\mathcal{X}$-valued tree $\mathbf{x}$ and any $\{0,1\}$-valued tree $\mathbf{y}$, both of depth n. Using $\mathbf{I}\{a \neq b\} = a(1-2b) + b$ for $a, b \in \{0,1\}$,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\mathbf{I}\{f(\mathbf{x}_t(\epsilon)) \neq \mathbf{y}_t(\epsilon)\}\right\} = \mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(1-2\mathbf{y}_t(\epsilon))f(\mathbf{x}_t(\epsilon))\right\} + \mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\mathbf{y}_t(\epsilon)\right\}$$

(13.8)

and the second term disappears under the expectation, thanks to the fact that $\mathbf{y}_t(\epsilon)$ and ϵ_t are independent. Using Lemma 13.10 with the $\{\pm 1\}$-valued tree $\mathbf{s} = 1 - 2\mathbf{y}$, the expression in (13.8) is equal to

$$\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{x}'_t(\epsilon))$$

for another tree $\mathbf{x}'$ specified in Lemma 13.10. We conclude that

$$\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{T}_{\ell(f)} \leq \sup_{\mathbf{x}'}\mathbb{E}\sup_{f\in\mathcal{F}}\mathbb{T}_f \ .$$

Since this holds for all $\mathbf{x}, \mathbf{y}$, the statement follows. $\qquad\square$

13.5 The End of the Story

We finish this section with a lower bound on the value $\mathcal{V}^{seq}(\mathcal{F}, n)$ for prediction of binary sequences with indicator loss. Together with Lemma 13.1, it implies that the value is characterized by the expected supremum of the tree process for the worst-case tree, up to a constant 2.

Theorem 13.11. *Let* $\mathcal{Y} = \{0,1\}$, $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, *and* $\ell(f, (x, y)) = \mathbf{I}\{f(x) \neq y\}$. *Then the value of sequential prediction defined in Eq. 5.18 is*

$$\mathcal{R}^{seq}(\mathcal{F}) \leq \mathcal{V}^{seq}(\mathcal{F}, n) \leq 2\mathcal{R}^{seq}(\mathcal{F})$$

In particular,
$$\mathcal{V}^{seq}(\mathcal{F}, n) \to 0 \quad \textit{if and only if} \quad \mathrm{ldim}(\mathcal{F}) < \infty$$

Proof. The upper follows from Corollary 7.16 and Lemma 13.1. We now prove the lower bound. Fix any $\mathcal{X}$-valued tree $\mathbf{x}$ of depth n. Consider the following strategy for Nature, already employed in Section 8.4. At round one, $\mathbf{x}_1$ is presented to the learner, who predicts $\widehat{\mathbf{y}}_1$ while Nature flips a coin $y'_1 \in \{-1, 1\}$ and presents

$y_1 = (y_1' + 1)/2 \in \{0,1\}$ to the learner. At round t, Nature presents $\mathbf{x}_t(y_1', \ldots, y_{t-1}')$, learner chooses $\widehat{\mathbf{y}}_t$, and Nature flips a coin y_t' and its $\{0,1\}$ version is presented to the learner. It is easy to see that the expected loss of the player is

$$\mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\{\widehat{\mathbf{y}}_t \neq y_t\}\right\} = \frac{1}{2}.$$

Thus, expected regret is

$$\mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\{\widehat{\mathbf{y}}_t \neq y_t\} - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\{f(\mathbf{x}_t(y_{1:t-1}')) \neq y_t'\}\right\} = \frac{1}{2}\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\left(1 - 2\mathbf{I}\{f(\mathbf{x}_t(y_{1:t-1}')) \neq y_t\}\right)\right\}$$

$$= \frac{1}{2}\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}y_t'(2f(\mathbf{x}_t(y_{1:t-1}')) - 1)\right\}$$

The proof of the lower bound is completed by observing that -1 disappears under the expectation and $\mathbf{x}$ was chosen arbitrarily.

The fact that the lower bound is not converging to zero if the Littlestone's dimension is infinite follows analogously to Theorem 8.2 by the definition of shattering. $\qquad\square$

Sequential Prediction: Real-Valued Functions

14.1 Covering Numbers

The development for sequential prediction with classes of real-valued functions parallels that for statistical learning. Recall that the notion of a projection $\mathcal{F}|_{x_1,\dots,x_n}$ coincided with the notion of a zero cover for tuples $x_1,\dots,x_n$ of data. In the previous lecture we also argued that for trees, the size of the projection $\mathcal{F}|_{\mathbf{x}}$ is not the same as the appropriately defined zero cover and that the latter is a better notion for controlling the supremum of the tree process. It is then natural to extend the definition of the zero cover (Definition 13.2) to the case of real-valued functions on trees. This can be done as follows.

Definition 14.1. A set V of $\mathbb{R}$-valued trees is a α-*cover* (with respect to ℓ_p) of $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ on an $\mathcal{X}$-valued tree $\mathbf{x}$ if

$$\forall f \in \mathcal{F}, \ \forall \epsilon \in \{\pm 1\}^n, \ \exists \mathbf{v} \in V \ \ \text{s.t.} \ \ \left(\frac{1}{n} \sum_{t=1}^{n} |f(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon)|^p \right)^{1/p} \leq \alpha . \qquad (14.1)$$

A α-*covering number* is defined as

$$\mathcal{N}_p(\mathcal{F},\mathbf{x},\alpha) = \min\left\{ \mathrm{card}(V) \ : \ V \text{ is an } \alpha\text{-cover} \right\}.$$

A set V of $\mathbb{R}$-valued trees is a α-*cover* (with respect to ℓ_∞) of $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ on an $\mathcal{X}$-valued tree $\mathbf{x}$ if

$$\forall f \in \mathcal{F}, \ \forall \epsilon \in \{\pm 1\}^n, \ \exists \mathbf{v} \in V \ \ \text{s.t.} \ \ |f(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon)| \leq \alpha \ \ \forall t \in \{1,\dots,n\}, \qquad (14.2)$$

and $\mathcal{N}_\infty(\mathcal{F},\alpha,\mathbf{x})$ is the minimal size of such a α-cover.

Once again, the order of quantifiers is crucial. For any function, the tree $\mathbf{v}$ providing the cover can depend on the particular path ϵ on which we seek to approximate the values of the function.

As in the i.i.d. case, we can now squint our eyes and view the class $\mathcal{F}$ at the fixed granularity α, similarly to Proposition 12.3:

Proposition 14.2. *For any* $\mathbf{x}$*, sequential Rademacher averages of a function class* $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ *satisfy*

$$\mathcal{R}^{seq}(\mathcal{F};\mathbf{x}) \le \inf_{\alpha \ge 0}\left\{\alpha + \sqrt{\frac{2\log\mathcal{N}_1(\mathcal{F},\alpha,\mathbf{x})}{n}}\right\}$$

Proof. Let V be a minimal α-cover of $\mathcal{F}$ on $\mathbf{x}$ with respect to ℓ_1-norm. For $f \in \mathcal{F}$ and $\epsilon \in \{\pm 1\}^n$, denote by $\mathbf{v}[f,\epsilon] \in V$ any element that "α-covers" f in the sense given by the definition. Observe that, in contrast to the proof of Proposition 12.3, the element $\mathbf{v}$ depends not only on f but also on the path ϵ, and this is crucial. Since $\mathbf{v}$ is a tree, $\mathbf{v}[f,\epsilon]_t$ is a function $\{\pm 1\}^{t-1} \to \mathbb{R}$. So, the value $\mathbf{v}[f,\epsilon]_t(\epsilon)$, while intimidating to parse, is the value at the t-th level on the path ϵ of that tree in V which "provides the cover" to the function f on the path ϵ.

We have

$$\mathcal{R}^{seq}(\mathcal{F};\mathbf{x}) = \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t f(\mathbf{x}_t(\epsilon))\right\}$$

$$= \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(f(\mathbf{x}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t(\epsilon)\right) + \epsilon_t\mathbf{v}[f,\epsilon]_t(\epsilon)\right\}$$

$$\le \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(f(\mathbf{x}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t(\epsilon)\right)\right\} + \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\mathbf{v}[f,\epsilon]_t(\epsilon)\right\}$$

$$\le \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\left|f(\mathbf{x}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t(\epsilon)\right|\right\} + \mathbb{E}\left\{\max_{\mathbf{v}\in V}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\mathbf{v}_t(\epsilon)\right\}$$

By definition of α-cover, the first term is upper bounded by α. The second term is precisely controlled by the maximal inequality of Lemma 9.5, which gives an upper bound of

$$\sqrt{\frac{2\log|V|}{n}},$$

due to the fact that $\sum_{t=1}^{n}\mathbf{v}_t(\epsilon)^2 \le n$ for any $\mathbf{v} \in V$ and $\epsilon \in \{\pm 1\}^n$. Note that if any value of the tree $\mathbf{v}$ is outside the interval $[-1,1]$, it can be truncated.

Since $\alpha > 0$ is arbitrary, the statement follows. $\qquad\square$

14.2   Chaining with Trees

As in the i.i.d. case, Proposition 14.2 does not always give the correct rates because it only considers covering numbers at one resolution. Thankfully, an analogue of Theorem 12.4 holds. The proof is more messy than for the i.i.d. case, as we need to deal with trees and paths rather than a single n-tuple of data. We provide the proof for completeness, but suggest skipping the proof unless one cannot overcome the curiosity.

Theorem 14.3. *For any $\mathcal{X}$-valued tree $\mathbf{x}$ of depth n, the sequential Rademacher averages of a function class $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ over $\mathbf{x}$ satisfy*

$$\mathcal{R}^{seq}(\mathcal{F};\mathbf{x}) \leq \mathcal{D}^{seq}(\mathcal{F};\mathbf{x})$$

where

$$\mathcal{D}^{seq}(\mathcal{F};\mathbf{x}) \triangleq \inf_{\alpha}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_{\alpha}^{1}\sqrt{\log \mathcal{N}_2(\mathcal{F},\mathbf{x},\delta)}\,d\delta\right\}$$

Proof. Define $\alpha_0 = 1$ and $\alpha_j = 2^{-j}$. For a fixed tree $\mathbf{x}$ of depth n, let V_j be an ℓ_2-cover at scale α_j. For any path $\epsilon \in \{\pm 1\}^n$ and any $f \in \mathcal{F}$, let $\mathbf{v}[f,\epsilon]^j \in V_j$ the element of the cover such that

$$\left(\frac{1}{n}\sum_{t=1}^{n}|f(\mathbf{x}_t(\epsilon)) - \mathbf{v}[f,\epsilon]_t^j(\epsilon)|^2\right)^{1/2} \leq \alpha_j$$

As before, $\mathbf{v}[f,\epsilon]_t^j$ denotes the t-th mapping of the tree $\mathbf{v}[f,\epsilon]^j$. For brevity, let us define the n-dimensional vectors

$$\mathrm{f}(\epsilon) = (f(\mathbf{x}_1(\epsilon)),\ldots,f(\mathbf{x}_n(\epsilon))) \quad \text{and} \quad v[f,\epsilon]^j = (\mathbf{v}[f,\epsilon]_1^j(\epsilon),\ldots,\mathbf{v}[f,\epsilon]_n^j(\epsilon))$$

With this notation,

$$\mathrm{f}(\epsilon) = \mathrm{f}(\epsilon) - v[f,\epsilon]^N + \sum_{j=1}^{N}(v[f,\epsilon]^j - v[f,\epsilon]^{j-1})$$

where $v[f,\epsilon]^0 = \mathbf{0}$. Further,

$$\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\mathbf{x}_t(\epsilon)) = \mathbb{E}\sup_{f\in\mathcal{F}}\langle\epsilon, \mathrm{f}(\epsilon)\rangle \tag{14.3}$$

which can be decomposed as in the proof of Theorem 12.4:

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\left\langle\epsilon,f(\epsilon)-v[f,\epsilon]^N\right\rangle+\sum_{j=1}^{N}\left\langle\epsilon,v[f,\epsilon]^j-v[f,\epsilon]^{j-1}\right\rangle\right\}$$

$$\leq\mathbb{E}\sup_{f\in\mathcal{F}}\left\langle\epsilon,f(\epsilon)-v[f,\epsilon]^N\right\rangle+\sum_{j=1}^{N}\mathbb{E}\sup_{f\in\mathcal{F}}\left\langle\epsilon,v[f,\epsilon]^j-v[f,\epsilon]^{j-1}\right\rangle \tag{14.4}$$

Unlike Theorem 12.4, the vectors $v[f,\epsilon]$ and $f(\epsilon)$ depend on ϵ. The first term above can be bounded via the Cauchy-Schwarz inequality exactly as in Eq. (12.4):

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left\langle\epsilon,f(\epsilon)-v[f,\epsilon]^N\right\rangle\leq n\alpha_N.$$

The second term in (14.4) is bounded by considering successive refinements of the cover. The argument, however, is more delicate than in Theorem 12.4, as the trees $\mathbf{v}[f,\epsilon]^j$, $\mathbf{v}[f,\epsilon]^{j-1}$ depend on the particular path. Consider all possible pairs of $\mathbf{v}^s\in V_j$ and $\mathbf{v}^r\in V_{j-1}$, for $1\leq s\leq\text{card}(V_j)$, $1\leq r\leq\text{card}(V_{j-1})$, where we assumed an arbitrary enumeration of elements. For each pair $(\mathbf{v}^s,\mathbf{v}^r)$, define a real-valued tree $\mathbf{w}^{(s,r)}$ by

$$\mathbf{w}_t^{(s,r)}(\epsilon)=\begin{cases}\mathbf{v}_t^s(\epsilon)-\mathbf{v}_t^r(\epsilon) & \text{if there exists }f\in\mathcal{F}\text{ s.t. }\mathbf{v}^s=\mathbf{v}[f,\epsilon]^j,\mathbf{v}^r=\mathbf{v}[f,\epsilon]^{j-1}\\ 0 & \text{otherwise.}\end{cases}$$

for all $t\in[n]$ and $\epsilon\in\{\pm1\}^n$. It is crucial that $\mathbf{w}^{(s,r)}$ can be non-zero only on those paths ϵ for which $\mathbf{v}^s$ and $\mathbf{v}^r$ are indeed the members of the covers (at successive resolutions) close to $f(\mathbf{x}(\epsilon))$ (in the ℓ_2 sense) *for some $f\in\mathcal{F}$*. It is easy to see that $\mathbf{w}^{(s,r)}$ is well-defined. Let the set of trees W_j be defined as

$$W_j=\left\{\mathbf{w}^{(s,r)}:1\leq s\leq\text{card}(V_j),1\leq r\leq\text{card}(V_{j-1})\right\}$$

Now, the second term in (14.4) is upper bounded as

$$\sum_{j=1}^{N}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t(\mathbf{v}[f,\epsilon]_t^j(\epsilon)-\mathbf{v}[f,\epsilon]_t^{j-1}(\epsilon))\leq\sum_{j=1}^{N}\mathbb{E}\max_{\mathbf{w}\in W_j}\sum_{t=1}^{n}\epsilon_t\mathbf{w}_t(\epsilon)$$

The last inequality holds because for any $j\in[N]$, $\epsilon\in\{\pm1\}^n$ and $f\in\mathcal{F}$ there is some $\mathbf{w}^{(s,r)}\in W_j$ with $\mathbf{v}[f,\epsilon]^j=\mathbf{v}^s$, $\mathbf{v}[f,\epsilon]^{j-1}=\mathbf{v}^r$ and

$$\mathbf{v}_t^s(\epsilon)-\mathbf{v}_t^r(\epsilon)=\mathbf{w}_t^{(s,r)}(\epsilon)\quad\forall t\leq n.$$

Clearly, $\text{card}(W_j) \leq \text{card}(V_j) \cdot \text{card}(V_{j-1})$. To invoke Lemma 9.5, it remains to bound the magnitude of all $\mathbf{w}^{(s,r)} \in W_j$ along all paths. For this purpose, fix $\mathbf{w}^{(s,r)}$ and a path ϵ. If there exists $f \in \mathcal{F}$ for which $\mathbf{v}^s = \mathbf{v}[f, \epsilon]^j$ and $\mathbf{v}^r = \mathbf{v}[f, \epsilon]^{j-1}$, then $\mathbf{w}_t^{(s,r)}(\epsilon) = \mathbf{v}[f, \epsilon]_t^j - \mathbf{v}[f, \epsilon]_t^{j-1}$ for any $t \in [n]$. By triangle inequality

$$\left(\sum_{t=1}^n \mathbf{w}_t^{(s,r)}(\epsilon)^2 \right)^{1/2} \leq \left(\sum_{t=1}^n (\mathbf{v}[f, \epsilon]_t^j(\epsilon) - f(\mathbf{x}_t(\epsilon)))^2 \right)^{1/2} + \left(\sum_{t=1}^n (\mathbf{v}[f, \epsilon]_t^{j-1}(\epsilon) - f(\mathbf{x}_t(\epsilon)))^2 \right)^{1/2} \leq \sqrt{n}(\alpha_j + \alpha_{j-1}),$$

and the latter quantity is equal to $3\sqrt{n}\alpha_j$. If there exists no such $f \in \mathcal{F}$ for the given ϵ and (s, r), then $\mathbf{w}_t^{(s,r)}(\epsilon)$ is zero for all $t \geq t_o$, for some $1 \leq t_o < n$, and thus

$$\sum_{t=1}^n \mathbf{w}_t^{(s,r)}(\epsilon)^2 \leq \sum_{t=1}^n \mathbf{w}_t^{(s,r)}(\epsilon')^2$$

for any other path ϵ' which agrees with ϵ up to t_o. Hence, the bound of $3\sqrt{n}\alpha_j$ holds for all $\epsilon \in \{\pm 1\}^n$ and all $\mathbf{w}^{(s,r)} \in W_j$.

Now, back to (14.4), we put everything together and apply Lemma 9.5:

$$\frac{1}{n} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(\mathbf{x}_t(\epsilon)) \leq \alpha_N + \frac{1}{\sqrt{n}} \sum_{j=1}^N 3\alpha_j \sqrt{2\log(\text{card}(V_j) \cdot \text{card}(V_{j-1}))}$$

and passing to the integral expression is exactly as in the proof of Theorem 12.4.

$\square$

14.3 Combinatorial Parameters

We continue to mirror the development for statistical learning, and define scale-sensitive dimensions relevant to sequential prediction. The following definition is the analogue of Definition 12.6.

Definition 14.4. An $\mathcal{X}$-valued tree $\mathbf{x}$ of depth n is α-*shattered* by a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$, if there exists an $\mathbb{R}$-valued tree $\mathbf{s}$ of depth n such that

$$\forall \epsilon \in \{\pm 1\}^n, \ \exists f \in \mathcal{F} \ \text{ s.t. }, \ \epsilon_t(f(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) > \alpha/2 \ \ \forall t \in \{1, \ldots, n\}$$

The tree $\mathbf{s}$ is called the *witness to shattering*. The *(sequential) fat-shattering dimension* $\text{fat}(\mathcal{F}, \alpha)$ at scale α is the largest n such that $\mathcal{F}$ α-shatters an $\mathcal{X}$-valued tree of depth n. We write $\text{fat}(\mathcal{F}, \alpha, \mathbf{x})$ for the fat-shattering dimension over $\text{Img}(\mathbf{x})$-valued trees.

We will also denote the fat-shattering dimension as $\text{fat}_\alpha(\mathcal{F})$. The definition can be seen as a natural scale-sensitive extension of Definition 13.4 to real-valued functions. We now show that the covering numbers are bounded in terms of the fat-shattering dimension. The following result is an analogue of the ℓ_∞ bound of Theorem 12.7. Somewhat surprisingly, the extra logarithmic term of that bound, which is so difficult to remove, is not present here from the outset.

Theorem 14.5. *Suppose $\mathcal{F}$ is a class of $[-1,1]$-valued functions on $\mathcal{X}$. Then for any $\alpha > 0$, any $n > 0$, and any $\mathcal{X}$-valued tree $\mathbf{x}$ of depth n,*

$$\mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{x}) \leq \left(\frac{2en}{\alpha}\right)^{\text{fat}(\mathcal{F}, \alpha)} \tag{14.5}$$

The proof of this theorem relies on the following extension of Theorem 13.7 to multi-class prediction. Once again, this bound comes about quite naturally when dealing with trees, while the analogous bound for the i.i.d. case is not known. One can compare the bound below to the multi-class bound obtained in [2, Theorem 3.3] and references therein.

Theorem 14.6. *Let $\mathcal{F} \subseteq \{0, \ldots, k\}^{\mathcal{X}}$ be a class of functions with $\text{fat}_2(\mathcal{F}) = d$. Then*

$$\mathcal{N}_\infty(1/2, \mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i} k^i .$$

Proof. For any $d \geq 0$ and $n \geq 0$, define the function $g_k(d, n) = \sum_{i=0}^{d} \binom{n}{i} k^i$, and observe that the function used for binary classification ($k = 1$) in Theorem 13.7 is exactly $g_1(d, n)$. The function g_k satisfies the recurrence

$$g_k(d, n) = g_k(d, n-1) + k g_k(d-1, n-1)$$

for all $d, n \geq 1$. To visualize this recursion, consider a $k \times n$ matrix and ask for ways to choose at most d columns followed by a choice among the k rows for each chosen column. The task can be decomposed into (a) making the d column choices out of the first $n-1$ columns, followed by picking rows (there are $g_k(d, n-1)$ ways to do it) or (b) choosing $d-1$ columns (followed by row choices) out of the first $n-1$ columns and choosing a row for the nth column (there are $k g_k(d-1, n-1)$ ways to do it). This gives the recursive formula.

In what follows, we shall refer to an L_∞ cover at scale $1/2$ simply as a $1/2$-cover. The theorem claims that the size of a minimal $1/2$-cover is at most $g_k(d, n)$. The proof proceeds by induction on $n + d$.

Base: For $d = 1$ and $n = 1$, there is only one node in the tree, i.e. the tree is defined by the constant $\mathbf{x}_1 \in \mathcal{X}$. Functions in $\mathcal{F}$ can take up to $k+1$ values on $\mathbf{x}_1$, i.e. $\mathcal{N}(0, \mathcal{F}, 1) \leq k+1$ (and, thus, also for the 1/2-cover). Using the convention $\binom{n}{0} = 1$, we indeed verify that $g_k(1,1) = 1 + k = k + 1$. The same calculation gives the base case for $n = 1$ and any $d \in \mathbb{N}$. Furthermore, for any $n \in \mathbb{N}$ if $d = 0$, then there is no point which is 2-shattered by $\mathcal{F}$. This means that functions in $\mathcal{F}$ differ by at most 1 on any point of $\mathcal{X}$. Thus, there is a 1/2 cover of size $1 = g_k(0, n)$, verifying this base case.

Induction step: Suppose by the way of induction that the statement holds for $(d, n-1)$ and $(d-1, n-1)$. Consider any tree $\mathbf{x}$ of depth n with $\text{fat}_2(\mathcal{F}, \mathbf{x}) = d$. Define the partition $\mathcal{F} = \mathcal{F}_0 \cup \ldots \cup \mathcal{F}_k$ with $\mathcal{F}_i = \{f \in \mathcal{F} : f(\mathbf{x}_1) = i\}$ for $i \in \{0, \ldots, k\}$, where $\mathbf{x}_1$ is the root of $\mathbf{x}$. Let $n = |\{i : \text{fat}_2(\mathcal{F}_i, \mathbf{x}) = d\}|$.

Suppose first, for the sake of contradiction, that $\text{fat}_2(\mathcal{F}_i, \mathbf{x}) = \text{fat}_2(\mathcal{F}_j, \mathbf{x}) = d$ for $|i - j| \geq 2$. Then there exist two trees $\mathbf{z}$ and $\mathbf{w}$ of depth d which are 2-shattered by $\mathcal{F}_i$ and $\mathcal{F}_j$, respectively, and with $\text{Img}(\mathbf{z}), \text{Img}(\mathbf{w}) \subseteq \text{Img}(\mathbf{x})$. Since functions within each subset $\mathcal{F}_i$ take on the same values on $\mathbf{x}_1$, we conclude that $\mathbf{x}_1 \notin \text{Img}(\mathbf{z}), \mathbf{x}_1 \notin \text{Img}(\mathbf{w})$. This follows immediately from the definition of shattering. We now *join* the two shattered $\mathbf{z}$ and $\mathbf{w}$ trees with $\mathbf{x}_1$ at the root and observe that $\mathcal{F}_i \cup \mathcal{F}_j$ 2-shatters this resulting tree of depth $d+1$, which is a contradiction. Indeed, the witness $\mathbb{R}$-valued tree $\mathbf{s}$ is constructed by joining the two witnesses for the 2-shattered trees $\mathbf{z}$ and $\mathbf{w}$ and by defining the root as $\mathbf{s}_1 = (i + j)/2$. It is easy to see that $\mathbf{s}$ is a witness to the shattering. We conclude that the number of subsets of $\mathcal{F}$ with fat-shattering dimension equal to d cannot be more than two (for otherwise at least two indices will be separated by 2 or more). We have three cases: $m = 0$, $m = 1$, or $m = 2$, and in the last case it must be that the indices of the two subsets differ by 1.

First, consider any $\mathcal{F}_i$ with $\text{fat}_2(\mathcal{F}_i, \mathbf{x}) \leq d - 1$, $i \in \{0, \ldots, k\}$. By induction, there are 1/2-covers V^ℓ and V^r of $\mathcal{F}_i$ on the subtrees $\mathbf{x}^\ell$ and $\mathbf{x}^r$, respectively, both of size at most $g_k(d-1, n-1)$. Just as in the proof of Theorem 13.7, out of these 1/2-covers we can create a 1/2-cover V for $\mathcal{F}_i$ on $\mathbf{x}$ by pairing the 1/2-covers in V^ℓ and V^r with i at the root. The resulting cover of $\mathcal{F}_i$ is of size at most $g_k(d-1, n-1)$. Such a construction holds for any $i \in \{0, \ldots, k\}$ with $\text{fat}_2(\mathcal{F}_i, \mathbf{x}) \leq d - 1$. Therefore, the total size of a 1/2-cover for the union $\cup_{i:\text{fat}_2(\mathcal{F}_i, \mathbf{x}) \leq d-1} \mathcal{F}_i$ is at most $(k+1-m) g_k(d-1, n-1)$.

If $m = 0$, the induction step is proven because $g_k(d-1, n-1) \leq g_k(d, n-1)$ and

so the total size of the constructed cover is at most

$$(k+1)g_k(d-1,n-1) \le g_k(d,n-1) + kg_k(d-1,n-1) = g_k(d,n).$$

Now, consider the case $m = 1$ and let $\mathrm{fat}_2(\mathcal{F}_i, \mathbf{x}) = d$. An argument exactly as above yields a $1/2$-cover for $\mathcal{F}_i$, and this cover is of size at most $g_k(d, n-1)$ by induction. The total $1/2$-cover is therefore of size at most

$$g_k(d,n-1) + kg_k(d-1,n-1) = g_k(d,n).$$

Lastly, for $m = 2$, suppose $\mathrm{fat}_2(\mathcal{F}_i, \mathbf{x}) = \mathrm{fat}_2(\mathcal{F}_j, \mathbf{x}) = d$ for $|i - j| = 1$. Let $\mathcal{F}' = \mathcal{F}_i \cup \mathcal{F}_j$. Note that $\mathrm{fat}_2(\mathcal{F}', \mathbf{x}) = d$. Just as before, the $1/2$-covering for $\mathbf{x}$ can be constructed by considering the $1/2$-covers for the two subtrees. However, when joining any $(\mathbf{v}^\ell, \mathbf{v}^r)$, we take $(i + j)/2$ as the root. It is straightforward to check that the resulting cover is indeed an $1/2$-cover, thanks to the relation $|i - j| = 1$. The size of the constructed cover is, by induction, $g_k(d, n-1)$, and the induction step follows. This concludes the induction proof, yielding the main statement of the theorem.

$\square$

To see that the growth of $g_k(d, n)$ is polynomial for $n \ge d$, observe that

$$\sum_{i=1}^d \binom{n}{i} k^i \le \left(\frac{kn}{d}\right)^d \sum_{i=1}^d \binom{n}{i} \left(\frac{d}{n}\right)^i \le \left(\frac{kn}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \le \left(\frac{ekn}{d}\right)^d$$

and it always holds that this sum is at most $(ekn)^d$. With the help of Theorem 14.6 we can prove Theorem 14.5. The idea is that when we discretize the values of the functions in $\mathcal{F}$ to the level α, we may view the resulting function as taking on values $\{0, \ldots, k\}$. Theorem 14.6 then tells that the size of the cover is upper bounded by $(ekn)^d$, and this bound translates immediately into the bound for the real-valued class $\mathcal{F}$.

Proof of Theorem 14.5. For any $\alpha > 0$ define an α-discretization of the $[-1, 1]$ interval as $B_\alpha = \{-1 + \alpha/2, -1 + 3\alpha/2, \ldots, -1 + (2k+1)\alpha/2, \ldots\}$ for $0 \le k$ and $(2k+1)\alpha \le 4$. Also for any $a \in [-1, 1]$ define $\lfloor a \rfloor_\alpha = \arg\min_{r \in B_\alpha} |r - a|$ with ties being broken by choosing the smaller discretization point. For a function $f : \mathcal{X} \to [-1, 1]$ let the function $\lfloor f \rfloor_\alpha$ be defined pointwise as $\lfloor f(x) \rfloor_\alpha$, and let $\lfloor \mathcal{F} \rfloor_\alpha = \{\lfloor f \rfloor_\alpha : f \in \mathcal{F}\}$. First,
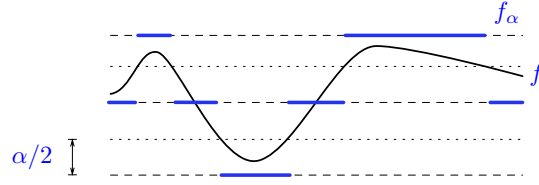
Figure 14.1: A $[-1,1]$-valued function f is discretized to the level α. The resulting discretized function f_α takes on at most $\lfloor 2/\alpha \rfloor + 1$ values. Any cover of the discretized class $\lfloor \mathcal{F} \rfloor_\alpha$ at scale $\alpha/2$ is also a cover of $\mathcal{F}$ at scale α.

we prove that $\mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{x}) \le \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{F} \rfloor_\alpha, \mathbf{x})$. Indeed, suppose the set of trees V is a minimal $\alpha/2$-cover of $\lfloor \mathcal{F} \rfloor_\alpha$ on $\mathbf{x}$. That is,

$$\forall f_\alpha \in \lfloor \mathcal{F} \rfloor_\alpha, \ \forall \epsilon \in \{\pm 1\}^n \ \exists \mathbf{v} \in V \text{ s.t. } |\mathbf{v}_t(\epsilon) - f_\alpha(\mathbf{x}_t(\epsilon))| \le \alpha/2$$

Pick any $f \in \mathcal{F}$ and let $f_\alpha = \lfloor f \rfloor_\alpha$. Then $\|f - f_\alpha\|_\infty \le \alpha/2$. Then for all $\epsilon \in \{\pm 1\}^n$ and any $t \in [n]$

$$\left| f(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon) \right| \le \left| f(\mathbf{x}_t(\epsilon)) - f_\alpha(\mathbf{x}_t(\epsilon)) \right| + \left| f_\alpha(\mathbf{x}_t(\epsilon)) - \mathbf{v}_t(\epsilon) \right| \le \alpha,$$

and so V also provides an L_∞ cover at scale α.

We conclude that $\mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{x}) \le \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{F} \rfloor_\alpha, \mathbf{x}) = \mathcal{N}_\infty(1/2, \mathcal{G}, \mathbf{x})$ where $G = \frac{1}{\alpha} \lfloor \mathcal{F} \rfloor_\alpha$. The functions of $\mathcal{G}$ take on a discrete set of at most $\lfloor 2/\alpha \rfloor + 1$ values. Obviously, by adding a constant to all the functions in $\mathcal{G}$, we can make the set of values to be $\{0, \ldots, \lfloor 2/\alpha \rfloor\}$. We now apply Theorem 14.6 with an upper bound $\sum_{i=0}^d \binom{n}{i} k^i \le (ekn)^d$ which holds for any $n > 0$. This yields $\mathcal{N}_\infty(1/2, \mathcal{G}, \mathbf{x}) \le (2en/\alpha)^{\mathrm{fat}_2(\mathcal{G})}$.

It remains to prove $\mathrm{fat}_2(\mathcal{G}) \le \mathrm{fat}_\alpha(\mathcal{F})$, or, equivalently (by scaling) $\mathrm{fat}_{2\alpha}(\lfloor \mathcal{F} \rfloor_\alpha) \le \mathrm{fat}_\alpha(\mathcal{F})$. To this end, suppose there exists an $\mathbb{R}$-valued tree $\mathbf{x}$ of depth $d = \mathrm{fat}_{2\alpha}(\lfloor \mathcal{F} \rfloor_\alpha)$ such that there is an witness tree $\mathbf{s}$ with

$$\forall \epsilon \in \{\pm 1\}^d, \ \exists f_\alpha \in \lfloor \mathcal{F} \rfloor_\alpha \text{ s.t. } \forall t \in [d], \ \epsilon_t(f_\alpha(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) \ge \alpha$$

Using the fact that for any $f \in \mathcal{F}$ and $f_\alpha = \lfloor f \rfloor_\alpha$ we have $\|f - f_\alpha\|_\infty \le \alpha/2$, it follows that

$$\forall \epsilon \in \{\pm 1\}^d, \ \exists f \in \mathcal{F} \text{ s.t. } \forall t \in [d], \ \epsilon_t(f(\mathbf{x}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) \ge \alpha/2$$

That is, $\mathbf{s}$ is a witness to α-shattering by $\mathcal{F}$. Thus for any $\mathbf{x}$,

$$\mathcal{N}_\infty(\alpha, \mathcal{F}, \mathbf{x}) \le \mathcal{N}_\infty(\alpha/2, \lfloor \mathcal{F} \rfloor_\alpha, \mathbf{x}) \le \left(\frac{2en}{\alpha} \right)^{\mathrm{fat}_{2\alpha}(\lfloor \mathcal{F} \rfloor_\alpha)} \le \left(\frac{2en}{\alpha} \right)^{\mathrm{fat}_\alpha(\mathcal{F})}$$

$\square$

Plugging the estimates on the ℓ_∞ covering numbers into the integral of Theorem 14.3, we obtain the following analogue of Corollary 12.8.

Corollary 14.7. *For $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ and any $\mathcal{X}$-valued tree $\mathbf{x}$ of depth n,*

$$\mathcal{D}^{seq}(\mathcal{F};\mathbf{x}) \le \inf_{\alpha \ge 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{\mathrm{fat}(\mathcal{F},\delta) \log\left(\frac{2en}{\delta}\right)} d\delta \right\}$$

where $\mathrm{fat}(\mathcal{F},\delta)$ *is the sequential fat-shattering dimension.*

Observe the unfortunate $\log n$ factor, which is unavoidable for ℓ_∞ covering numbers. Ideally, we would like to use ℓ_2 covering numbers in conjunction with Theorem 14.3, but for that we need to prove an analogue of (12.10), which appears to be a rather difficult problem.

♚ Prove the analogue of (12.10) for the ℓ_2 tree covering numbers, and get an $A+$ in the course.

14.4 Contraction

Lemma 14.8. *Fix any $\mathcal{X}$-valued tree $\mathbf{x}$ of depth n. If $\phi : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz (that is, $\phi(a) - \phi(b) \le L|a-b|$ for all $a, b \in \mathbb{R}$), then*

$$\mathcal{R}^{seq}(\phi \circ \mathcal{F}) \le L\mathcal{R}^{seq}(\mathcal{F}) \times O(\log^{3/2} n)$$

Note that this upper bound falls short of the analogous result in Lemma 12.9 for two reasons: it is only for $\sup_{\mathbf{x}} \mathcal{R}^{seq}(\mathcal{F};\mathbf{x})$ rather than $\mathcal{R}^{seq}(\mathcal{F};\mathbf{x})$, and there is an extra poly-logarithmic factor. The first shortcoming is not a big issue, but we believe that the extra factor should not be there.

♚ Prove (4) without the extra poly-logarithmic factor, and get an $A+$ in the course.

As the next proposition shows, we can avoid the poly-logarithmic factor for the problem of supervised learning with a *convex* loss function (in place of ϕ) by passing directly from the value of the prediction problem to the sequential

Rademacher complexity of the base class $\mathcal{F}$. This means that we can indeed "erase" the loss function for the majority of the problems of interest (recall that this works for the indicator loss too, as shown in Lemma 13.1). However, it would be nice to prove a general result without the extra factor.

Proposition 14.9. *Let* $\mathcal{Y} \subset \mathbb{R}$, $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, *and assume* $\boldsymbol{\ell}(y_1, y_2)$ *is convex and L-Lipschitz in* y_1 *for any* y_2. *Consider the (improper) supervised sequential prediction problem. Then*

$$\mathcal{V}^{seq}(\mathcal{F}, n) \le 2L\mathcal{R}^{seq}(\mathcal{F})$$

✎ **Exercise 14.1** ($\star\star$)**.** Let $\mathcal{X} \subseteq \mathbb{R}$, $L > 0$, and define a class $\mathcal{F}_L([a,b]) = \left\{ f_\theta : \mathbb{R} \to [0,1] | \theta \in [a,b] \right\}$ where

$$f_\theta(x) = \begin{cases} 1 & \text{if } x \le \theta \\ 1 - L(x-\theta) & \text{if } \theta < x \le \theta + 1/L \\ 0 & \text{otherwise} \end{cases}$$

is a "ramp" function with slope L. Prove that $\mathcal{F}_L([-\infty, \infty])$ has infinite (sequential) fat-shattering dimension for any scale $\alpha < 1/2$. Prove, however, that sequential fat-shattering dimension of $\mathcal{F}_L([a,b])$, for any $a, b \in \mathbb{R}$, is finite. Conclude that the value of a supervised learning problem with ramp functions and Lipschitz loss is converging to zero in both sequential and the statistical learning frameworks.

14.5 Lower Bounds

Matching (up to a constant) lower bounds on the value of the game can be shown in several cases of interest: supervised learning and linear loss. These two cover a wide range of problems, including online convex optimization without additional curvature assumptions on the functions.

For the supervised learning case, the lower bound of sequential Rademacher complexity is already shown in Theorem 13.11 for the binary prediction case, but exactly the same argument works for the real-valued prediction with absolute loss.

We now prove a matching lower bound for the linear loss case. Suppose $\mathcal{F}$ is a convex subset of some separable Banach space, and $\mathcal{Z}$ is a convex subset of the dual space. Let **z** be any $\mathcal{Z}$-valued tree of depth n. We now exhibit a lower bound on the value of the game, achieved by the adversary playing according to this tree. The proof is in the same spirit as the proof of Theorem 13.11 as well as the lower bounds in Section 8.4.

Lemma 14.10. *Suppose $\mathcal{F}$ is a convex subset of a separable Banach space, and $\mathcal{Z}$ is a convex subset of the dual space, symmetric about the origin. For the linear game defined with $\ell(f,z) = \langle f,z \rangle$, we have*

$$\mathcal{R}^{seq}(\mathcal{F}) \leq \mathcal{V}^{seq}(\mathcal{F},n) \leq 2\mathcal{R}^{seq}(\mathcal{F})$$

Proof. Given $\mathbf{z}$, the strategy for the adversary will be the following: at round 1, draw ϵ_1, observe the learner's move f_t and present $\epsilon_1 \mathbf{z}$. On round t, the adversarial move is defined by $\epsilon_t \mathbf{z}_t(\epsilon_{1:t-1})$. This gives the following lower bound on the value of the game:

$$
\begin{aligned}
n \cdot \mathcal{V}^{seq}(\mathcal{F},n) &= \left\langle\!\!\!\left\langle \inf_{q_t} \sup_{z_t} \mathbb{E}_{\widehat{\mathbf{y}}_t \sim q_t} \right\rangle\!\!\!\right\rangle_{t=1}^n \left\{ \sum_{t=1}^n \langle \widehat{\mathbf{y}}_t, z_t \rangle - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \langle f, z_t \rangle \right\} \\
&\geq \left\langle\!\!\!\left\langle \inf_{q_t} \mathbb{E}_{\widehat{\mathbf{y}}_t \sim q_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\!\right\rangle_{t=1}^n \left\{ \sum_{t=1}^n \langle \widehat{\mathbf{y}}_t, \epsilon_t \mathbf{z}_t(\epsilon) \rangle - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \langle f, \epsilon_t \mathbf{z}_t(\epsilon) \rangle \right\} \\
&\geq \mathbb{E}_\epsilon \left\{ -\inf_{f \in \mathcal{F}} \sum_{t=1}^n \langle f, \epsilon_t \mathbf{z}_t(\epsilon) \rangle \right\} \\
&= \mathcal{R}^{seq}(\mathcal{F};\mathbf{z})
\end{aligned}
$$

where the player's loss disappears because the conditional distribution of $\epsilon_t \mathbf{z}_t(\epsilon)$ is zero mean. Since the lower bound holds for an arbitrary tree $\mathbf{z}$, the statement follows. □

15

Examples: Complexity of Linear and Kernel Classes, Neural Networks

In the past five chapters, we have developed powerful theoretical tools, and it is now time to reap the benefits. Before proceeding, let us first recall the high-level picture. In Chapter 7, we had motivated the study of the suprema of several stochastic processes. We argued that a handle on the empirical or Rademacher process gives a finite sample guarantee on the performance of Empirical Risk Minimization, while the martingale and tree processes provide an upper bound on the minimax value of the sequential prediction problem. After this realization, we turned to the study of the suprema of stochastic processes. In Chapter 9 we showed some rather straightforward upper bounds on the suprema over finite index sets. The two chapters following that were devoted to extending the finite-class results to Rademacher processes indexed by infinite classes of functions in binary and in real-valued cases. In the following two chapters, we studied the suprema of the tree process indexed by infinite classes of, respectively, binary or real-valued functions. We defined various notions of complexity of the function class that capture the behavior of the suprema of these stochastic processes.

We now proceed to calculate the upper bounds on the expected suprema of the Rademacher and tree processes. While we do not prove lower bounds, it can be shown that, for the examples presented here, the behavior of the suprema of the Rademacher and tree processes is almost identical. Hence, the guarantees for the statistical learning framework with i.i.d. data are the same as those for regret minimization in the setting of individual sequences.

15.1 Prediction with Linear Classes

In the supervised setting, we have pairs (x, y) of predictor-response variables. Consider the square, absolute and logistic loss functions $\ell(f(x), y)$ paired with linear functions $f(x) = \langle f, x \rangle$:

$$\left(\langle f, x \rangle - y\right)^2, \quad |\langle f, x \rangle - y|, \quad \log\left(1 + \exp\left\{-y\langle f, x \rangle\right\}\right)$$

For $\mathcal{X}$ and $\mathcal{F}$ being unit balls in dual norms, it is natural to assume that $\mathcal{Y} = [-1, 1]$ since that is the range of functions $f(x)$. Then, the square loss is evaluated on the interval $[-2, 2]$, and on this interval it is a 2-Lipschitz function of $f(x)$ for any y. Clearly, the absolute loss is a 1-Lipschitz function, and so is the logistic loss.

For any such L-Lipschitz loss, Proposition 14.9 tells us that the value of the sequential prediction problem

$$\mathcal{V}(\mathcal{F}) \leq 2L\mathcal{R}^{seq}(\mathcal{F}).$$

The same result holds for the setting of statistical learning in view of Lemma 12.9:

$$\mathcal{V}^{iid}(\mathcal{F}) \leq 2L\mathcal{R}^{iid}(\mathcal{F}).$$

We can now use upper bounds on $\mathcal{R}^{seq}(\mathcal{F})$ and $\mathcal{R}^{iid}(\mathcal{F})$ for the examples of linear function classes considered in Chapter 10.

Observe that for square loss, the rate $4/\sqrt{n}$ guaranteed through the argument above seems to mismatch the rate $\mathcal{O}(d/n)$ obtained in Section 4. In fact, each of these bounds works in its own regime: for $d > \sqrt{n}$ the former kicks in, while for small d the $\mathcal{O}(d/n)$ rate might be better. A more careful analysis, which goes beyond the scope of this course, allows one to interpolate between these two rates.

15.2 Kernel Methods

Linear predictors discussed above are attractive from both algorithmic and from the analysis perspectives. At first glance, however, linear predictors might appear to be limited in terms of their representativeness. The kernel methods are aimed to alleviate this very problem. The basic idea is to first map data points into a feature space (usually of higher—or even infinite—dimension) and to use linear predictors in that feature space. To illustrate the point, let us consider the two dimensional example shown in Figure 15.1. The two classes are arranged in circles,

and are not linearly separable. In fact, any linear predictor will misclassify a constant proportion of the data points. However, if we used a circle as the decision boundary, the data would be perfectly separable. Such a decision rule is of the form $f(x) = \text{sign}(x[1]^2 + x[2]^2 - r^2)$, and can be written as a linear form once we define the features appropriately. In particular, consider the feature space given by the map $x \mapsto \phi(x)$ where $\phi(x) = (x[1]^2, x[2]^2, x[1] \cdot x[2], x[1], x[2], 1)$. The classifier based on the circle $x[1]^2 + x[2]^2 - r^2$ can now be written as a half-space defined by the vector $f = (1, 1, 0, 0, 0, -r^2) \in \mathbb{R}^6$. This half-space in the feature space is specified by $\langle f, x \rangle = x[1]^2 + x[2]^2 - r^2$ and it perfectly classifies the data. We see that by employing a mapping to a higher dimensional feature space one can use linear classifiers in the feature space and increase the expressive power of linear predictors.
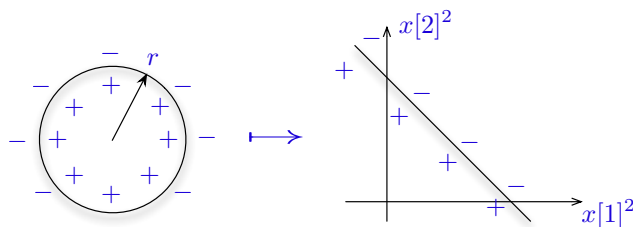
Figure 15.1: The problem becomes linearly separable once an appropriate feature mapping ϕ is chosen.

Specifically for kernel methods we consider some input space $\mathcal{X}$ and an associated feature space mapping ϕ that maps input in $\mathcal{X}$ to a possibly infinite dimensional Hilbert space. We further assume that for every $x \in \mathcal{X}$, $\|\phi(x)\| \leq R$ for some finite R (the norm is the norm on the Hilbert space). The function class $\mathcal{F}$ we consider consists of all the elements of the Hilbert space such that $\|f\| \leq B$, and the predictors we consider are of form $\langle f, \phi(x) \rangle$, as motivated in the above example.

As mentioned earlier, the feature space is often taken to be infinite dimensional. In this case, one may wonder how to even use such linear functions on computers with finite memory. After all, we cannot store or process infinite sequences. Fortunately, there is no need to store the functions! The trick that makes this possible is commonly called *the kernel trick*. In Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, there exists a function called *the kernel function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$* such that $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$. Given such a kernel function, the *Representer*

Theorem acts as a cornerstone for kernel methods. Roughly speaking, the Representer Theorem implies that any algorithm that only operates on the functions f through the inner products can do all the computations with the kernel function. This means the optimization objective for the algorithm can depend on terms like $\langle x, \phi(x_t) \rangle$ or the norm $\langle f, f \rangle$, and the solution $\widehat{\boldsymbol{y}}_t \in \mathcal{H}$ in the RKHS at time t chosen by the learning algorithm can be represented as a linear combination of feature vectors of the data points. That is $\widehat{\boldsymbol{y}}_t$ can be written as $\widehat{\boldsymbol{y}}_t = \sum_{i=1}^{t-1} \alpha_i \phi(x_i)$. Viewed as a linear function in the feature space

$$\widehat{\boldsymbol{y}}_t(x) = \langle \widehat{\boldsymbol{y}}_t, \phi(x) \rangle = \sum_{i=1}^{t-1} \alpha_i \langle \phi(x_i), \phi(x) \rangle = \sum_{i=1}^{t-1} \alpha_i k(x_i, x)$$

Thanks to the Representer Theorem, the main take home message is that one never needs to explicitly deal with feature space but only with kernel functions over data seen so far. As for this chapter, the main thing to note is that one can treat the Rademacher complexity of the kernel class as the Rademacher complexity of the linear class in the feature space with the Hilbert norm of f bounded by B and Hilbert norm of data bounded by R. This is basically the same as the L_2/L_2 example in Chapter 10, and so one gets the bound

$$\mathcal{R}^{iid}(\mathcal{F}) \leq \mathcal{R}^{seq}(\mathcal{F}) \leq \sqrt{\frac{R^2 B^2}{n}} \ .$$

The above upper bound shows that one may go well beyond the linear classes in terms of expressiveness. The kernel methods are arguably the most useful practical methods.

15.3 Neural Networks

Neural networks can be viewed as a class of functions which are built in a hierarchical manner. Let $\mathcal{F} \subset [-1, 1]^{\mathcal{X}}$ be a base function class (that is, a set of possible functions at the input layer of the neural network). In every node of every subsequent layer, we take a linear combination of outputs of the nodes from the previous layer, and this linear combination is subsequently passed through a Lipschitz transformation function $\sigma : \mathbb{R} \mapsto \mathbb{R}$, producing the output of the node. Typically the transformation function is a sigmoid and performs a form of a soft thresholding. The motivation for such a construction is to mimic neural activity. Each neuron

is excited by the input connections from other neurons, and its own output is a (modulated) function of the aggregate excitation.

We now formally describe a multilayer neural network with some arbitrary base class $\mathcal{F}$. Consider any base function class $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$, and assume (in order to simplify the proof) that $0 \in \mathcal{F}$. We now recursively define $\mathcal{F}_i$ as: $\mathcal{F}_1 = \mathcal{F}$ and for each $2 \le i \le k$

$$\mathcal{F}_i = \left\{ x \mapsto \sum_j w_j \sigma\left(f_j(x)\right) \; : \; f_j \in \mathcal{F}_{i-1}, \|w\|_1 \le R \right\}$$

where σ is an L-Lipschitz transformation function such that $\sigma(0) = 0$. The following lemma upper bounds the Rademacher complexity of the neural network class in terms of the Rademacher complexity of the base class.

Lemma 15.1. *For any base function class $\mathcal{F} \subset [-1,1]^{\mathcal{X}}$ that contains the zero function,*

$$\mathcal{R}^{iid}(\mathcal{F}_k) \le \mathcal{R}^{seq}(\mathcal{F}_k) \le (RL)^{k-1} \, \mathcal{O}\left(\log^{\frac{3}{2}(k-1)}(n)\right) \mathcal{R}^{seq}(\mathcal{F})$$

Proof. We shall prove that for any $i \in \{2, \dots, k\}$,

$$\mathcal{R}_n^{seq}(\mathcal{F}_i) \le LR \, \mathcal{O}(\log^{3/2}(n)) \, \mathcal{R}_n^{seq}(\mathcal{F}_{i-1})$$

To see this note that

$$\mathcal{R}_n^{seq}(\mathcal{F}_i) = \sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[\sup_{\substack{w:\|w\|_1 \le R \\ f_j \in \mathcal{F}_{i-1}}} \sum_{t=1}^n \epsilon_t \left(\sum_j w_j \sigma\left(f_j(\mathbf{x}_t(\epsilon))\right) \right) \right]$$

Interchanging the sum over t and the sum over j, and then using Hölder's inequality, we get an upper bound of

$$\sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[\sup_{\substack{w:\|w\|_1 \le R \\ f_j \in \mathcal{F}_{i-1}}} \|w\|_1 \max_j \left| \sum_{t=1}^n \epsilon_t \sigma\left(f_j(\mathbf{x}_t(\epsilon))\right) \right| \right] \le \sup_{\mathbf{x}} \mathbb{E}_{\epsilon}\left[R \sup_{f \in \mathcal{F}_{i-1}} \left| \sum_{t=1}^n \epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right) \right| \right]$$

$$(15.1)$$

We now claim that we can remove the absolute value at the expense of an extra factor of 2. Indeed,

$$\sup_{f \in \mathcal{F}_{i-1}} \left| \sum_{t=1}^n \epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right) \right| \le \max\left\{ \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^n \epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right), \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^n -\epsilon_t \sigma\left(f(\mathbf{x}_t(\epsilon))\right) \right\}$$

Since $\sigma(0) = 0$ and $0 \in \mathcal{F}$, we also conclude that $0 \in \mathcal{F}_i$ for any layer i. We can thus replace the maximum by the sum of the two terms. The supremum over $\mathbf{x}$ in (15.1) can then be split into two equal terms:

$$\sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[R \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^n \epsilon_t \sigma \left(f(\mathbf{x}_t(\epsilon)) \right) \right] + \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[R \sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^n -\epsilon_t \sigma \left(f(\mathbf{x}_t(\epsilon)) \right) \right]$$

thus proving the assertion. In view of Lemma 14.8, we pass to the upper bound of

$$2RL\,\mathcal{O}(\log^{3/2} n) \sup_{\mathbf{x}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^n \epsilon_t f(\mathbf{x}_t(\epsilon)) \right] = 2RL\,\mathcal{O}(\log^{3/2} n)\,\mathcal{R}_n^{seq}(\mathcal{F}_{i-1})$$

The statement of the lemma follows by applying the above argument repeatedly for all layers. □

A key property to note about the above bound is that while the number of layers of the neural network enters the bound, the number of nodes in each layer does not, and so in each layer we could have any number of nodes without compromising the bound.

15.4 Discussion

Both excess risk (in statistical learning) and regret (in sequential prediction) are based on the comparison of learner's performance to the performance within a class $\mathcal{F}$. That is, the goal is to minimize the difference

$$\text{(our loss)} - \text{(loss of the best element in } \mathcal{F})$$

Of course, competing with very simple $\mathcal{F}$ is not interesting, as the performance of the comparator is unlikely to be good. Conversely, being able to have vanishing excess risk or vanishing regret with respect to a *large* class $\mathcal{F}$ implies good performance in a wide variety of situations, as we are more likely to capture the underlying phenomenon that generates the data.

The course so far has focused on understanding a key question: what are the relevant complexities of $\mathcal{F}$ that make it possible to have vanishing excess loss or vanishing regret? The key point of this discussion is that these complexities are not exactly what we might think of when we look at a particular $\mathcal{F}$ and its expressiveness. For instance, take a finite $\mathcal{F}$ and take its convex hull. Lemma 7.14 implies that the Rademacher complexities of the convex hull is equal to that of the

finite collection. However, the convex hull of $\mathcal{F}$ is more expressive. Consider the example of neural networks which correspond to complicated nested compositions and convex combinations. Despite the apparent complexity, as we showed in Lemma 15.1, the inherent complexity is simply that of the base class. In the case of kernel methods, a bound on Rademacher complexity is possible in infinite dimensional spaces, which shows once again that the notion of "inherent complexity" relevant to the problem of excess loss or regret minimization is not exactly what we might think of. In some sense, this is great news, as we can take expressive ("large") classes $\mathcal{F}$ while still guaranteeing that the inherent complexity is manageable.

16

Large Margin Theory for Classification

17

Regression with Square Loss: From Regret to Nonparametric Estimation

Part III

Algorithms

18

Algorithms for Sequential Prediction: Finite Classes

Within the setting of Statistical Learning Theory, the upper bounds on the supremum of the Rademacher process give us a mandate to use the Empirical Risk Minimization algorithm, thanks to the inequality (7.2). But what can be said about the setting of Sequential Prediction? The upper bounds on the minimax value guarantee *existence* of a strategy that will successfully minimize regret, but how do we go about finding this strategy? Recall that to upper bound the minimax value, we jumped over quite a few hurdles: we appealed to the minimax theorem and then performed sequential symmetrization, replacing the mixed strategies of the opponent by a binary tree and coin flips. In this process, we seem to have lost the hope of finding a method for minimizing regret.

Of course, we can say that the algorithm for sequential prediction is obtained by solving the long minimax problem (5.15) at each step. But this is not very satisfying, as we would like something more succinct and useful in practice. George Pólya's famous quote comes to mind: "If you can't solve a problem, then there is an easier problem you can solve: find it."

Let us consider the simplest case we can think of: the supervised setting of binary prediction with a finite benchmark set $\mathcal{F}$. More precisely, let $\mathcal{X}$ be some set, $\mathcal{Y} = \mathcal{D} = \{\pm 1\}$, $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$, and the loss $\ell(y_1, y_2) = \mathbf{I}\{y_1 \neq y_2\}$. Lemma 9.5 (coupled with Lemma 13.1 to erase the indicator loss) yields a bound

$$\mathcal{V}^{seq}(\mathcal{F}) \leq 2\sqrt{\frac{2\log|\mathcal{F}|}{n}} \, , \tag{18.1}$$

and the immediate question is: what is the algorithm that achieves it?

Before answering this question, consider an even simpler problem in order to gain some intuition.

18.1 The Halving Algorithm

Imagine that among the functions in $\mathcal{F}$ there is one that exactly gives the true labels y_t, an assumption we termed *realizable* on page 26. That is, there is $f \in \mathcal{F}$ such that, no matter what x_t is presented by Nature, the label y_t is required to satisfy $y_t = f(x_t)$. If we know there is a perfect predictor in $\mathcal{F}$, we can prune away elements of $\mathcal{F}$ as soon as they disagree with some observed y_t. It is then clear that we have a strategy that makes at most $\text{card}(\mathcal{F}) - 1$ mistakes.

However, we can do better. Suppose that at each round we update the set $\mathcal{F}_t$ of possible perfect predictors at time t given the new information. At the first round, $\mathcal{F}_1 = \mathcal{F}$. Upon observing x_t we predict $\widehat{y}_t$ according to the majority of the functions in $\mathcal{F}_t$. If $\widehat{y}_t = y_t$, we set $\mathcal{F}_{t+1} = \mathcal{F}_t$. Otherwise, we set $\mathcal{F}_{t+1} = \{f \in \mathcal{F}_t : f(x_t) = y_t\}$. Clearly, any time we made a mistake, $\text{card}(\mathcal{F}_{t+1}) \leq \text{card}(\mathcal{F}_t)/2$. Suppose we made m mistakes after n rounds. Then $1 \leq \text{card}(\mathcal{F}_n) \leq \text{card}(\mathcal{F})/2^m$ and thus

$$m \leq \log_2(\text{card}(\mathcal{F})),$$

making regret (to the zero-loss comparator) no larger than $n^{-1}\log_2(\text{card}(\mathcal{F}))$. It is not hard to construct an example with $\log_2(\text{card}(\mathcal{F}))$ as a lower bound on the number of mistakes suffered by the learner.

We mention the Halving Algorithm because it has various incarnations in a range of disciplines. For instance, the idea of choosing the point that brings the most information (in our case, the majority vote) can be seen in Binary Search and the Method of Centers of Gravity in Optimization. If each new point decreases the "size" of the version space by a multiplicative factor, the rate of localizing the solution is exponential.

18.2 The Exponential Weights Algorithm

What if there is no element of $\mathcal{F}$ that predicts the sequence perfectly? We can no longer throw out parts of $\mathcal{F}$. The idea is to keep an eye on all the functions in $\mathcal{F}$, yet decrease our confidence in those that make many mistakes. The algorithm we

now present is a "soft version" of the Halving Algorithm, and it is arguably the most famous method in the framework of sequential prediction. It works not only for binary-valued prediction, but also for the real-valued case and for non-supervised problems.

In later lectures, we will appeal to Lemma 5.1 and argue that our sequential strategy can be deterministic. In the present case of a finite $\mathcal{F}$, however, the assumptions of the lemma are not satisfied, as $\mathcal{F}$ is not a convex set. In fact, a mixed strategy is necessary, and we will call it $q_t \in \Delta(\mathcal{F})$, following Eq. (5.15). Here, $\Delta(\mathcal{F})$ is a set of distributions on $N = \text{card}(\mathcal{F})$ elements. Let us enumerate the functions in $\mathcal{F}$ as $f^1, \dots, f^N$.

Exponential Weights Algorithm (EWA), Supervised Learning

Initialize: $q_1 = (1/N, \dots, 1/N)$, $\eta = \sqrt{\frac{8 \ln N}{n}}$

At time t:

    Observe x_t, sample $i_t \sim q_t$, and predict $\widehat{\boldsymbol{y}}_t = f^{i_t}(x_t)$

    Observe y_t and update

$$q_{t+1}(i) \propto q_t(i) \times \exp\left\{-\eta \ell(f^i(x_t), y_t)\right\} \quad \text{for all} \quad i \in \{1, \dots, N\}$$

Lemma 18.1. *Suppose $\text{card}(\mathcal{F}) = N$. For the supervised learning problem with any loss function ℓ with the range in $[0, 1]$, the Exponential Weights Algorithm guarantees*

$$\mathbb{E}\left\{\frac{1}{n} \sum_{t=1}^{n} \ell(\widehat{\boldsymbol{y}}_t, y_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t)\right\} \leq \sqrt{\frac{\ln N}{2n}} \tag{18.2}$$

no matter how the sequence $(x_1, y_1), \dots, (x_n, y_n)$ is chosen.

Proof of Lemma 18.1 *(see e.g. [16]).* Denote the cumulative loss of expert i by $L_t^i = \sum_{s=1}^{t} \ell(f^i(x_s), y_s)$. Let $W_t = \sum_{i=1}^{N} \exp\left\{-\eta L_t^i\right\}$ with $W_0 = N$. We then have

$$\ln \frac{W_n}{W_0} = \ln \sum_{i=1}^{N} \exp(-\eta L_n^i) - \ln N \geq \ln\left(\max_{i=1,\dots,N} \exp(-\eta L_n^i)\right) - \ln N = -\eta \min_{i=1,\dots,N} L_n^i - \ln N.$$

On the other hand,

$$
\begin{aligned}
\ln \frac{W_t}{W_{t-1}} &= \ln \frac{\sum_{i=1}^{N} \exp\{-\eta L_t^i\}}{\sum_{i=1}^{N} \exp\{-\eta L_{t-1}^i\}} \\
&= \ln \frac{\sum_{i=1}^{N} \exp(-\eta \ell(f_t^i, y_t)) \cdot \exp(-\eta L_{t-1}^i)}{\sum_{i=1}^{N} \exp(-\eta L_{t-1}^i)} \\
&= \ln \mathbb{E}_{i_t \sim q_t} \exp\left(-\eta \ell(f^{i_t}(x_t), y_t)\right)
\end{aligned}
$$

We now employ a useful fact that for a random variable $X \in [a, b]$,

$$
\ln \mathbb{E} e^{sX} \leq s\mathbb{E}X + \frac{s^2(b-a)^2}{8}
$$

for any $s \in \mathbb{R}$ (see [16] for more details). This inequality implies

$$
\ln \mathbb{E}_{i_t} \exp\left(-\eta \ell(f^{i_t}(x_t), y_t)\right) \leq -\eta \mathbb{E}_{i_t} \ell(f^{i_t}(x_t), y_t) + \frac{\eta^2}{8}
$$

for the loss $0 \leq \ell(\cdot, \cdot) \leq 1$. Summing the last inequality over $t = 1, \ldots, n$, and observing that the logarithms telescope,

$$
\ln \frac{W_n}{W_0} \leq -\eta \sum_{t=1}^{n} \mathbb{E}_{i_t} \ell(f^{i_t}(x_t), y_t) + \frac{\eta^2 n}{8}.
$$

Combining the upper and lower bounds for $\ln W_{n+1}/W_1$,

$$
\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{i_t} \ell(f^{i_t}(x_t), y_t) - \min_{i \in \{1, \ldots, N\}} \frac{1}{n} \sum_{t=1}^{n} \ell(f^i(x_t), y_t) \leq \frac{\ln N}{n\eta} + \frac{\eta}{8}.
$$

Balancing the two terms with $\eta = \sqrt{\frac{8 \ln N}{n}}$ and taking the expectation gives the bound. $\qquad \square$

Notice a strange feature of the Exponential Weights Algorithm: q_t is decided upon before seeing x_t. In other words, the structure of $\mathcal{F}$ (agreement and disagreement among the functions) is not exploited. Consequently, the Exponential Weights Algorithm has the same regret guarantee in a more abstract setting of unsupervised learning, where at each time step the learner picks $\hat{y}_t \in \mathcal{F}$, Nature simultaneously chooses $z_t \in \mathcal{Z}$, the learner suffers loss $\ell(\hat{y}_t, z_t)$, and z_t is observed. Suppose the loss ℓ is again taking values in $[0, 1]$. The algorithm then becomes

---

Exponential Weights Algorithm (EWA), Unsupervised Learning

Initialize: $q_1 = (1/N, \dots, 1/N)$, $\eta = \sqrt{\frac{8 \ln N}{n}}$

At time t:

   Sample $i_t \sim q_t$, and predict $\widehat{\boldsymbol{y}}_t = f^{i_t} \in \mathcal{F}$

   Observe z_t and update

$$q_{t+1}(i) \propto q_t(i) \times \exp\left\{-\eta \ell(f^i, z_t)\right\} \quad \text{for all} \quad i \in \{1, \dots, N\}$$

It is easy to see that the same upper bound

$$\mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^n \ell(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^n \ell(f, z_t)\right\} \leq \sqrt{\frac{\ln N}{2n}} \tag{18.3}$$

holds in this case too.

Another variant of the problem bears the name "Prediction with Expert Advice", as we may view the side information x_t as a vector of advice of N *experts*. Then, the choice of elements of $\mathcal{F}$ corresponds to the choice of which expert to follow, and we may take $\mathcal{F} = \{\mathbf{e}_1, \dots, \mathbf{e}_N\}$, the vertices of the probability simplex Δ_N. As $\Delta(\mathcal{F}) = \Delta_N$, the mixed strategy q_t is simply a distribution over the coordinates $\{1, \dots, N\}$. Let us now mention a variant where the expert advice is actually being combined to form a prediction. More precisely, at time t, each expert $i \in \{1, \dots, N\}$ offers a prediction $x_t^i \in [0, 1]$, the learner forms a decision $\widehat{\boldsymbol{y}}_t \in [0, 1]$, observes the outcome $y_t \in [0, 1]$ and incurs loss $\ell(\widehat{\boldsymbol{y}}_t, y_t)$ for some cost function ℓ, *convex in the first argument*. At the end of n days, we would like to have suffered loss not much worse than that of the *best* expert, without knowing who is the best until the very end. In this case, the algorithm is

---

Exponential Weights Algorithm (EWA), Prediction with Expert Advice

Initialize: $q_1 = (1/N, \dots, 1/N)$, $\eta = \sqrt{\frac{8 \ln N}{n}}$

At time t:

   Observe $x_t \in [0, 1]^N$

   Predict $\widehat{\boldsymbol{y}}_t = q_t^\top x_t$

   Observe y_t and update

$$q_{t+1}(i) \propto q_t(i) \times \exp\left\{-\eta \ell(x_t^i, y_t)\right\} \quad \text{for all} \quad i \in \{1, \dots, N\}$$

We then have a regret bound

$$\mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t) - \inf_{i \in \{1,\ldots,N\}}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(x_t^i, y_t)\right\} \leq \sqrt{\frac{\ln N}{2n}} \tag{18.4}$$

The fact that EWA treats the experts as unrelated entities that generate predictions has both positive and negative implications. On the positive side, this fact increases the applicability of the algorithm. On the negative side, the method does not take into account the correlations between experts. If experts give very similar predictions, we would expect smaller regret, yet this fact is not reflected in the upper bound. Exploiting the structure of $\mathcal{F}$ is a very interesting topic, and it will be studied later in the course. We can already guess that covering of $\mathcal{F}$ needs to come into the picture, as it indicates the effective number of distinct functions.

A variant of Exponential Weights is also employed for statistical aggregation of estimators [51, 29, 19], treating them as fixed entities. Since the structure of $\mathcal{F}$ is not exploited by the Exponential Weights method, the resulting estimator is able to mimic the best one no matter how the estimators are originally constructed. More generally, it is instructive to think of the Exponential Weights algorithm as a "union bound" algorithm, a point that will be explained in a few lectures.

✎ **Exercise 18.1** (★★)**.** Prove an analogue of Lemma 18.1 for a countable $\mathcal{F}$. More precisely, show that for any distribution π on $\mathcal{F}$,

$$\mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t)\right\} \leq \inf_{f \in \mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(f(x_t), y_t) + \frac{1 + \ln(1/\pi(f))}{\sqrt{8n}}\right\} \tag{18.5}$$

✎ **Exercise 18.2** (★★★)**.** For the previous exercise, prove an upper bound of the form (18.5) but with $\log(1/\pi(f))$ under the square root.

✎ **Exercise 18.3** (★★)**.** Suppose $\mathcal{F} = \cup_{i \in \mathbb{N}}\mathcal{F}_i$ is a countable union (of potentially uncountable sets), and the sequential prediction problem with each $\mathcal{F}_i$ is Uniformly Sequentially Consistent (recall the definitions in Chapter 6). Prove that the sequential prediction problem with respect to $\mathcal{F}$ is Universally Sequentially Consistent. Hint: Use the previous exercise.

19

Algorithms for Sequential Prediction: Binary Classification with Infinite Classes

The binary prediction problem within the supervised learning setting is as follows: on round t, some side information x_t is revealed, the learner picks $\widehat{\boldsymbol{y}}_t \in \{\pm 1\}$, and the outcome y_t is revealed. Let us consider the case of prediction with hyperplanes

$$\left\{ g(x) = \text{sign}(\langle f, x \rangle) : f \in \mathcal{F} \subset \mathbb{R}^d \right\}$$

in $\mathbb{R}^d$. However, we immediately recognize that this class has infinite Littlestone's dimension, since the example of thresholds on an interval (given in Section 8.4) can be easily embedded into this problem if $d \geq 2$.

Interestingly, the problem of infinite Littlestone's dimension can be circumvented if we assume a *margin condition*. Analogously to the previous lecture, where we assumed that there exists an $f \in \mathcal{F}$ that perfectly classifies the data, we now make an assumption that some $f \in \mathcal{F}$ perfectly classifies data with an extra margin of $\gamma > 0$. Let us state this more formally.

Assumption 19.1. *Given sets $\mathcal{F}$ and $\mathcal{X}$, the γ-margin condition (for some $\gamma > 0$) is given by*

$$\exists f \in \mathcal{F} \;\; s.t. \;\; \forall t \in [n], \;\; y_t \langle f, x_t \rangle > \gamma.$$

19.1 Halving Algorithm with Margin

Let $\mathcal{F} = \mathsf{B}_2^d$. Because of the margin condition, we can discretize the uncountable set $\mathcal{F}$ of all experts (i.e. the sphere) and pick a finite number of them for our prob-

lem. It is easy to show that if we discretize to a fine enough level, there will be an expert that perfectly classifies the data and then we can use the Halving Algorithm over this discretized set. The following lemma does exactly this. Before we begin, consider the discretization of the interval $[-1, 1]$ at scale $\gamma/2d$ given by $B = \{-1, -1 + \gamma/2d, -1 + \gamma/d, \ldots, 1\}$.

Lemma 19.2. *Let $\mathcal{X} \subset [-1, 1]^d$. Under the γ-margin condition, the total number of mistakes committed by the Halving Algorithm with the set of half spaces specified by the discretization $\mathcal{F}_\gamma = B^d$ is at most*

$$d \log\left(\frac{4d}{\gamma} + 1\right)$$

Proof. By the margin assumption there exists an $f^* \in \mathcal{F}$, such that $\forall t \in [n]$, $y_t \langle f^*, x_t \rangle > \gamma$. Since $\mathcal{F}_\gamma$ is a $\gamma/2d$ discretization on each co-ordinate, there exists $f_\gamma^* \in \mathcal{F}_\gamma$ such that

$$\forall i \in [d], \ |f^*[i] - f_\gamma^*[i]| \le \gamma/2d .$$

Hence, for any $t \in [n]$,

$$\left| y_t \langle f^*, x_t \rangle - y_t \langle f_\gamma^*, x_t \rangle \right| = \left| \langle f^* - f_\gamma^*, x_t \rangle \right| \le \sum_{i=1}^{d} |f^*[i] - f_\gamma^*[i]| \le \gamma/2 .$$

Combining with the margin assumption this implies that for all $t \in [n]$, $y_t \langle f_\gamma^*, x_t \rangle > \gamma/2$. In short, the set of half-spaces specified by $\mathcal{F}_\gamma$ has an element f_γ^* that perfectly separates the data. Cardinality of the set is given by $|\mathcal{F}_\gamma| = \left(\frac{4d}{\gamma} + 1\right)^d$. Using the Halving Algorithm, we get that the number of mistakes is bounded by

$$\log_2 |\mathcal{F}_\gamma| = d \log\left(\frac{4d}{\gamma} + 1\right) .$$

$\square$

The nice aspect of this reduction to a finite case is its simplicity. It is transparent how the margin assumption allows us to discretize and only consider a finite number of experts. Another good side is the $\log \gamma^{-1}$ dependence on the margin γ. The bad aspects: the majority algorithm over $O(\gamma^{-d})$ experts is computationally infeasible, and the dependence of the regret bound on d is linear.

19.2   The Perceptron Algorithm

Luckily, the bad computational performance of the discretization algorithm can be avoided. The Perceptron algorithm is a simple and efficient method for the problem of classification with a margin assumption, and it has some remarkable history. Invented by Rosenblatt in 1957, this algorithm can be called the father of neural networks and the starting point of machine learning. In their 1969 book, Minsky and Papert showed that perceptrons are limited in what they can learn: in particular, they cannot learn the XOR function. This realization is attributed to a decline in the field for at least a decade. Meanwhile, in 1964, Aizerman, Braverman and Rozonoer introduced a kernelized version of the perceptron and showed its relationship to stochastic approximation methods that minimize a risk functional (with respect to an unknown distribution). Their results gained popularity only in the 80's.

Perceptron Algorithm

Initialize: $f_1 = 0$.

At each time step $t = 1, \ldots, n$

　　Receive $x_t \in \mathbb{R}^d$

　　Predict $\widehat{y}_t = \text{sign}(\langle f_t, x_t \rangle)$

　　Observe $y_t \in \{-1, 1\}$

　　Update $f_{t+1} = f_t + y_t x_t$ if $\text{sign}(\langle f_t, x_t \rangle) \neq y_t$, and set $f_{t+1} = f_t$ otherwise.

Note that no update is performed when the correct label is predicted.

Lemma 19.3. *Under the γ-margin condition, the number of mistakes committed by the Perceptron Algorithm is at most $\left(\frac{R}{\gamma}\right)^2$ where $R = \max_t \|x_t\|$.*

Proof. Let f^* be the normal to the hyperplane that separates all data by the margin γ. Let us look at how the "correlation" between f^* and f_t is evolving. If a mistake is made on round t,

$$\langle f^*, f_{t+1} \rangle = \langle f^*, f_t + y_t x_t \rangle = \langle f^*, f_t \rangle + y_t \langle f^*, x_t \rangle \geq \langle f^*, f_t \rangle + \gamma.$$

Thus, every time there is a mistake, the inner product of our hypothesis f_t with the unknown f^* increases by γ. If m mistakes have been made over n rounds, we have $\langle f^*, f_{n+1} \rangle \geq \gamma m$ since we started with a zero vector. Now, the main question

166

is whether the increase in the inner product is due to the smaller angle between the vectors (i.e. we are indeed getting close to the unknown f^* in terms of the direction), or is it because of the increase in the length of $\|f_t\|$?

While the "correlation" with f^* increases with every mistake, we can show that the size of the hypothesis $\|f_t\|$ cannot increase too fast. If there is a mistake on round t,

$$\|f_{t+1}\|^2 = \|f_t + y_t x_t\|^2 = \|f_t\|^2 + 2y_t \langle f_t, x_t \rangle + \|x_t\|^2 \leq \|f_t\|^2 + \|x_t\|^2$$

Then after n rounds and m mistakes,

$$\|f_{n+1}\|^2 \leq mR^2.$$

Combining the two arguments,

$$\gamma m \leq \langle f^*, f_{n+1} \rangle \leq \|f^*\| \cdot \|f_{n+1}\| \leq \sqrt{mR^2}$$

assuming f^* is a unit vector, and so

$$m \leq \left(\frac{R}{\gamma} \right)^2.$$

$\square$

19.3   The Winnow Algorithm

20

Algorithms for Online Convex Optimization

20.1 Online Linear Optimization

Suppose $\mathcal{F} = \mathcal{X} = B_2 \subset \mathbb{R}^d$, $\ell(f, x) = \langle f, x \rangle$, and the decision set $\mathcal{D} = \mathcal{F}$. At every t, the learner predicts $\widehat{\boldsymbol{y}}_t \in \mathcal{D}$, observes x_t, and suffers a cost $\langle \widehat{\boldsymbol{y}}_t, x_t \rangle$. The goal is to minimize regret defined as

$$\frac{1}{n} \sum_{t=1}^{n} \langle \widehat{\boldsymbol{y}}_t, x_t \rangle - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \langle f, x_t \rangle . \tag{20.1}$$

Eq. (10.2) tells us that the value of the game is upper bounded by $1/\sqrt{n}$ (the constant 2 is not necessary). So, what is the strategy that gives this regret guarantee?

The first idea is to reduce the problem to the case of a finite number of experts. What allows us to do so is the structure of $\mathcal{F}$: two nearby vectors $f, g \in \mathcal{F}$ have similar loss. Let us sketch a possible reduction. We can find a cover of the unit ball $\mathcal{F}$ with centers in $\mathcal{F}' \subset \mathcal{F}$, at the granularity $1/\sqrt{n}$ in the ℓ_2 distance. It is easy to check that regret against $\mathcal{F}$ is within $1/\sqrt{n}$ from regret with respect to the comparator class $\mathcal{F}$. The regret bound for the Exponential Weights algorithm on the finite set $\mathcal{F}'$ is $O\left(\sqrt{\ln |\mathcal{F}'| / n}\right)$, as given in the last section. But the size of the cover $\mathcal{F}'$ is at least $\Omega(n^{d/2})$, and so the resulting regret bound is $O\left(\sqrt{d \ln(n) / n}\right)$ at best. Unfortunately, this is off our target $1/\sqrt{n}$ in Eq. (10.2). Even more importantly, the Exponential Weights algorithm on $\mathcal{F}'$ is not an efficient algorithm. Evidently, this is because Exponential Weights treats each expert as a separate entity, despite the fact that the nearby elements of $\mathcal{F}'$ incur similar loss. While the discretization

168

takes into account the linear structure of the problem, the Exponential Weights algorithm does not. The proposed reduction is not satisfying, but the idea is worth mentioning. For many problems in sequential prediction, an Exponential Weights algorithm over the discretization often yields a near-optimal—yet (possibly) an inefficient—method.

How can we take algorithmic advantage of the linear structure of the problem? First thing to note is that the loss function is convex (linear) in $\hat{\boldsymbol{y}}_t$, and the set $\mathcal{F}$ is convex. Hence, by Lemma 5.1, there is a deterministic method for choosing $\hat{\boldsymbol{y}}_t$'s. (In contrast to the Exponential Weights algorithm which works in the space of distributions, the present problem should be solvable with a deterministic method.) The regret objective in (20.1) has a flavor of an optimization problem, so it is reasonable to check if the simple deterministic method

$$\hat{\boldsymbol{y}}_{t+1} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{s=1}^{t} \langle f, x_s \rangle$$

works. This method, known as *Follow the Leader* simply chooses the best decision given the observed data. Follow the Leader is basically the Empirical Risk Minimization method for each observed history. Unlike the i.i.d. case of statistical learning where ERM is a sound procedure, the sequential prediction problem with linear functions is not solved by this simple method, as the next (classical) example shows.

Example 17. Suppose $\mathcal{F} = \mathcal{X} = [-1, 1]$ and $\boldsymbol{\ell}(f, x) = f \cdot x$. Suppose Nature presents $x_1 = 0.5$ and $x_{2k} = -1$, $x_{2k+1} = 1$ for $k = 1, \ldots, \lfloor n/2 \rfloor$. The choice f_1 makes little difference, so suppose $f_1 = 0$. Clearly, the FTL algorithm generates a sequence $f_{2k} = -1$ and $f_{2k+1} = 1$, inadvertently matching the x_t sequence, and thus the average loss of FTL is 1. On the other hand, taking $f = 0$ in the comparator term makes it 0, and thus regret is at least a constant.

20.2   Gradient Descent

Let us now check whether a simple gradient update works. To this end, define

$$\hat{\boldsymbol{y}}_{t+1} = \Pi_{\mathcal{F}}(\hat{\boldsymbol{y}}_t - \eta x_t)$$

where $\Pi_{\mathcal{F}}(g) = \inf_{f \in \mathcal{F}} \|f - g\|$ the Euclidean projection onto the set $\mathcal{F}$. Fix any $f^* \in \mathcal{F}$ and write

$$\|\widehat{\boldsymbol{y}}_{t+1} - f^*\|^2 = \|\Pi_{\mathcal{F}}(\widehat{\boldsymbol{y}}_t - \eta x_t) - f^*\|^2 \leq \|\widehat{\boldsymbol{y}}_t - \eta x_t - f^*\|^2 = \|\widehat{\boldsymbol{y}}_t - f^*\|^2 + \eta^2 \|x_t\|^2 - 2\eta \langle \widehat{\boldsymbol{y}}_t - f^*, x_t \rangle$$

The inequality in the above chain follows from the simple fact that an element outside the set cannot be closer than its projection to an element in the set. Rearranging,

$$2\eta \langle \widehat{\boldsymbol{y}}_t - f^*, x_t \rangle \leq \|\widehat{\boldsymbol{y}}_t - f^*\|^2 - \|\widehat{\boldsymbol{y}}_{t+1} - f^*\|^2 + \eta^2 \|x_t\|^2$$

Summing over $t = 1, \ldots, n$ and dividing by 2η,

$$\sum_{t=1}^{n} \langle \widehat{\boldsymbol{y}}_t - f^*, x_t \rangle \leq (2\eta)^{-1} \left(\|\widehat{\boldsymbol{y}}_1 - f^*\|^2 - \|\widehat{\boldsymbol{y}}_{n+1} - f^*\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{n} \|x_t\|^2 \leq \frac{1}{2\eta} + \frac{n\eta}{2} = \sqrt{n},$$

for $\eta = 1/\sqrt{n}$. We used $\widehat{\boldsymbol{y}}_1 = 0$ and the fact that $\|x_t\| \leq 1$ for any t. Since this holds for any $f^* \in \mathcal{F}$, we have proved the following:

Lemma 20.1. *Let $\mathcal{F} = \mathcal{X} = B_2^d$, and $\ell(f, x) = \langle f, x \rangle$. Then the deterministic strategy of gradient descent*

$$\widehat{\boldsymbol{y}}_{t+1} = \Pi_{\mathcal{F}}(\widehat{\boldsymbol{y}}_t - \eta x_t)$$

with projection yields the regret bound of $1/\sqrt{n}$, where $\widehat{\boldsymbol{y}}_1 = 0$ and $\eta = 1/\sqrt{n}$.

Notably, the gradient descent method is efficient, and its regret guarantee matches the non-constructive upper bound on the supremum of the Rademacher process, given in Eq. (10.2). It should also be noted that d does not appear in the upper bound, and thus the same guarantee holds for a unit ball B_2 in an infinite dimensional Hilbert space.

20.3   Follow the Regularized Leader and Mirror Descent

Follow the Regularized Leader (FTRL)
Input: Regularization function $\mathcal{R}$, learning rate $\eta > 0$.
For $t \geq 0$,
$$\widehat{\boldsymbol{y}}_{t+1} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{s=1}^{t} \langle f, x_s \rangle + \eta^{-1} \mathcal{R}(f)$$

Recall from (10.6) that a function $\mathcal{R}$ is σ-strongly convex over $\mathcal{F}$ with respect to $\|\cdot\|$ if

$$\mathcal{R}(a) \geq \mathcal{R}(b) + \langle \nabla \mathcal{R}(b), a - b \rangle + \frac{\sigma}{2} \|a - b\|^2$$

for all $a, b \in \mathcal{F}$. If $\mathcal{R}$ is non-differentiable, $\nabla \mathcal{R}(b)$ can be replaced with any subgradient $\nabla \in \partial \mathcal{R}(b)$.

Definition 20.2. The Bregman divergence with respect to a convex function $\mathcal{R}$ is defined as

$$\mathcal{D}_{\mathcal{R}}(a, b) = \mathcal{R}(a) - \mathcal{R}(b) - \langle \nabla \mathcal{R}(b), a - b \rangle$$

Clearly, if $\mathcal{R}$ is σ-strongly convex with respect to $\|\cdot\|$, then

$$\mathcal{D}_{\mathcal{R}}(a, b) \geq \frac{\sigma}{2} \|a - b\|^2 \tag{20.2}$$

The definition of Bregman divergence immediately leads to the following useful equality:

$$\langle \nabla \mathcal{R}(a) - \nabla \mathcal{R}(b), c - b \rangle = \mathcal{D}_{\mathcal{R}}(c, b) + \mathcal{D}_{\mathcal{R}}(b, a) - \mathcal{D}_{\mathcal{R}}(c, a) \tag{20.3}$$

The *convex conjugate* of the function $\mathcal{R}$ is defined as

$$\mathcal{R}^*(u) = \sup_a \langle u, a \rangle - \mathcal{R}(a), \tag{20.4}$$

and this transformation is known as the Legendre-Fenchel transformation.

Recall the definition of the dual norm in (10.5).

Mirror Descent

Input: $\mathcal{R}$ σ-strongly convex w.r.t. $\|\cdot\|$, learning rate $\eta > 0$

$$\widehat{\boldsymbol{y}}_{t+1} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \langle f, x_t \rangle + \eta^{-1} \mathcal{D}_{\mathcal{R}}(f, \widehat{\boldsymbol{y}}_t) \tag{20.5}$$

or, equivalently,

$$\tilde{\boldsymbol{y}}_{t+1} = \nabla \mathcal{R}^*(\nabla \mathcal{R}(\widehat{\boldsymbol{y}}_t) - \eta x_t) \quad \text{and} \quad \widehat{\boldsymbol{y}}_{t+1} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \mathcal{D}_{\mathcal{R}}(f, \tilde{\boldsymbol{y}}_{t+1}) \tag{20.6}$$

20.3 Follow the Regularized Leader and Mirror Descent

Lemma 20.3. *Let $\mathcal{F}$ be a convex set in a separable Banach space $\mathcal{B}$ and $\mathcal{X}$ be a convex set in the dual space $\mathcal{B}^*$. Let $\mathcal{R} : \mathcal{B} \mapsto \mathbb{R}$ be a σ-strongly convex function on $\mathcal{F}$ with respect to some norm $\|\cdot\|$. For any strategy of Nature,*

$$\frac{1}{n}\sum_{t=1}^{n}\langle \hat{\boldsymbol{y}}_t, x_t\rangle - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\langle f, x_t\rangle \le R_{\max}X_{\max}\sqrt{\frac{2}{\sigma n}} \tag{20.7}$$

where $R_{\max}^2 = \sup_{f,g\in\mathcal{F}}\mathcal{R}(f) - \mathcal{R}(g)$, and $X_{\max} = \sup_{x\in\mathcal{X}}\|x\|_*$ for the dual norm $\|\cdot\|_*$.

Proof. Fix any $f^* \in \mathcal{F}$ and $\eta > 0$ to be chosen later. Then

$$\eta\sum_{t=1}^{n}\langle \hat{\boldsymbol{y}}_t - f^*, x_t\rangle = \eta\sum_{t=1}^{n}\langle \hat{\boldsymbol{y}}_t - \tilde{\boldsymbol{y}}_{t+1}, x_t\rangle + \eta\sum_{t=1}^{n}\langle \tilde{\boldsymbol{y}}_{t+1} - f^*, x_t\rangle \tag{20.8}$$

Using the inequality $\langle f, x\rangle \le \|f\| \cdot \|x\|_*$, which follows directly from the definition of the dual norm,

$$\eta\langle \hat{\boldsymbol{y}}_t - \tilde{\boldsymbol{y}}_{t+1}, x_t\rangle \le \eta\|\hat{\boldsymbol{y}}_t - \tilde{\boldsymbol{y}}_{t+1}\|\,\|x_t\|_* = \left(\sqrt{\sigma}\,\|\hat{\boldsymbol{y}}_t - \tilde{\boldsymbol{y}}_{t+1}\|\right)\left(\frac{\eta}{\sqrt{\sigma}}\,\|x_t\|_*\right)$$

The inequality $ab \le a^2/2 + b^2/2$ now yields an upper bound

$$\frac{\sigma}{2}\left\|\hat{\boldsymbol{y}}_t - \tilde{\boldsymbol{y}}_{t+1}\right\|^2 + \frac{\eta^2}{2\sigma}\|x_t\|_*^2 \le \mathcal{D}_{\mathcal{R}}\left(\tilde{\boldsymbol{y}}_{t+1}, \hat{\boldsymbol{y}}_t\right) + \frac{\eta^2}{2\sigma}X_{\max}^2$$

The definition of the Mirror Descent update and the equality (20.3) imply

$$\begin{aligned}
\eta\langle \tilde{\boldsymbol{y}}_{t+1} - f^*, x_t\rangle &= \langle \tilde{\boldsymbol{y}}_{t+1} - f^*, \nabla\mathcal{R}(\hat{\boldsymbol{y}}_t) - \nabla\mathcal{R}(\tilde{\boldsymbol{y}}_{t+1})\rangle \\
&= \mathcal{D}_{\mathcal{R}}\left(f^*, \hat{\boldsymbol{y}}_t\right) - \mathcal{D}_{\mathcal{R}}\left(f^*, \tilde{\boldsymbol{y}}_{t+1}\right) - \mathcal{D}_{\mathcal{R}}\left(\tilde{\boldsymbol{y}}_{t+1}, \hat{\boldsymbol{y}}_t\right) \\
&\le \mathcal{D}_{\mathcal{R}}\left(f^*, \hat{\boldsymbol{y}}_t\right) - \mathcal{D}_{\mathcal{R}}\left(f^*, \hat{\boldsymbol{y}}_{t+1}\right) - \mathcal{D}_{\mathcal{R}}\left(\tilde{\boldsymbol{y}}_{t+1}, \hat{\boldsymbol{y}}_t\right)
\end{aligned}$$

where the last inequality is due to the fact that for the projection $\hat{\boldsymbol{y}}_{t+1}$ is closer to any point f^* in the set in terms of the Bregman divergence than the unprojected point $\tilde{\boldsymbol{y}}_{t+1}$. Adding the last inequality for $t = 1,\dots,n$, we obtain

$$\eta\sum_{t=1}^{n}\langle \tilde{\boldsymbol{y}}_{t+1} - f^*, x_t\rangle \le \mathcal{D}_{\mathcal{R}}\left(f^*, \hat{\boldsymbol{y}}_1\right) - \mathcal{D}_{\mathcal{R}}\left(f^*, \hat{\boldsymbol{y}}_{n+1}\right) - \sum_{t=1}^{n}\mathcal{D}_{\mathcal{R}}\left(\tilde{\boldsymbol{y}}_{t+1}, \hat{\boldsymbol{y}}_t\right) \tag{20.9}$$

Combining all the terms,

$$\eta\sum_{t=1}^{n}\langle \hat{\boldsymbol{y}}_t - f^*, x_t\rangle \le \mathcal{D}_{\mathcal{R}}\left(f^*, \hat{\boldsymbol{y}}_1\right) - \mathcal{D}_{\mathcal{R}}\left(f^*, \hat{\boldsymbol{y}}_{n+1}\right) + \frac{n\eta^2}{2\sigma}X_{\max}^2$$

$$\le R_{\max}^2 + \frac{n\eta^2}{2\sigma}X_{\max}^2$$

$\square$

172

Example 18. Consider the regularization function

$$\mathcal{R}(f) = \frac{1}{2}\|f\|^2$$

In this case, the mirror space coincides with the original space, as the gradient mapping $\nabla\mathcal{R}(f) = f$, and the Mirror Descent algorithm becomes simply the Gradient Descent method.

Example 19. If we take

$$\mathcal{R}(f) = \sum_{i=1}^{N}(f_i \log f_i - f_i)$$

defined over $\mathbb{R}_+^N$, the mirror space is defined by the gradient mapping $\nabla\mathcal{R}(f) = \log f$ where the logarithm is taken coordinate-wise. The inverse mapping is $\nabla\mathcal{R}^* = (\nabla\mathcal{R})^{-1}$ is precisely the mapping $a \mapsto \exp(a)$ for $a \in \mathbb{R}^N$. The unprojected point $\tilde{\boldsymbol{y}}_{t+1}$ is then defined by

$$\tilde{\boldsymbol{y}}_{t+1}(i) = \exp\left\{\log(\hat{\boldsymbol{y}}_t(i)) - \eta x_t(i)\right\} = \hat{\boldsymbol{y}}_t(i)\exp\left\{-\eta x_t(i)\right\}$$

and the projection onto the simplex with respect to the KL divergence is equivalent to normalization. Hence, we recover the Exponential Weights algorithm.

20.4   From Linear to Convex Functions

21

Example: Binary Sequence Prediction and the Mind Reading Machine

We now continue the example discussed in Chapter 2. This chapter is based on the paper of Blackwell [9], as well as the detailed development of Lerche and Sarkar [36].

Recall the basic problem of predicting a $\{0,1\}$ sequence $z_1, z_2, \ldots$. At each stage t, the learner chooses $\widehat{y}_t \in \mathcal{D} = \{0,1\}$ and Nature reveals $z_t \in \mathcal{Z} = \{0,1\}$. Suppose the cost $\ell(\widehat{y}_t, z_t)$ is the indicator of a mistake $\mathbf{I}\{\widehat{y}_t \neq z_t\}$. For any t, let the average number of *correct predictions* after t rounds be denoted by

$$\bar{c}_t = \frac{1}{t} \sum_{s=1}^{t} \mathbf{I}\{\widehat{y}_s = z_s\} \ .$$

Further, let $\bar{z}_t = \frac{1}{t} \sum_{s=1}^{t} z_s$ be the frequency of 1's in the sequence so far.

As discussed in Chapter 2, if the sequence $z_1, z_2, \ldots$ chosen by Nature is actually generated i.i.d. from some Bernoulli distribution with bias $p \in [0,1]$, the simple *Follow the Leader* strategy (see Section 20.3)

$$\widehat{y}_{t+1} = \operatorname*{argmin}_{f \in \{0,1\}} \sum_{s=1}^{t} \mathbf{I}\{f \neq z_s\}$$

works well. The i.i.d. result, however, is quite limited. For instance, for the alternating sequence $0, 1, 0, 1, \ldots$ the FTL method yields close to zero proportion of correct predictions, while there is a clear pattern that can be learned.

The failure of FTL on the alternating sequence $0, 1, 0, 1, \ldots$ is akin to Example 17 which shows that FTL does not work for linear optimization. Of course, we should

174

not even be looking at deterministic strategies for bit prediction, as the set $\mathcal{F}$ is not convex. A randomized prediction strategy is necessary.

Luckily, we already have all the tools we need to prove vanishing regret for individual sequences. While we only prove the *in-expectation* convergence to simplify the presentation, the almost-sure analogues are not much more difficult.

21.1 Prediction with Expert Advice

Fix two (somewhat simple-minded) experts: one always advises 1 and the other 0. Let q_t be the mixed strategy over $\mathcal{F}$ which we may equivalently represent as the probability $q_t \in [0,1]$ of predicting 1 (or, choosing the first expert). For the Exponential Weights Algorithm with $N = 2$ experts, Lemma 18.1 then guarantees

$$\mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\{\widehat{y}_t \neq z_t\}\right\} - \inf_{f \in \mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\{f \neq z_t\} \leq \sqrt{\frac{\ln 2}{2n}} \tag{21.1}$$

or, equivalently,

$$\max\{\bar{z}_t, 1 - \bar{z}_t\} - \mathbb{E}\{\bar{c}_n\} = \sup_{f \in \mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\mathbf{I}\{f = z_t\} - \mathbb{E}\{\bar{c}_n\} \leq \sqrt{\frac{\ln 2}{2n}} \tag{21.2}$$

which is the desired result in expectation. For the almost-sure result, we may use the high-probability bound of Proposition 22.1 and the so-called doubling trick [16]. And that's all!

21.2 Blackwell's method

We now present a different approach to achieving no regret for individual sequences of bits. This method is somewhat longer to develop than the Exponential Weights hammer that we used, yet it provides some very interesting insights. Moreover, the proof will be a stepping stone to Chapter 27, the Blackwell's Approachability Theorem, a powerful generalization of von Neumann's minimax theorem.

The goal (2.1) of the learner can be rephrased very elegantly in the language of geometry. Define $L_t = (\bar{z}_t, \bar{c}_t) \in [0,1]^2$ and

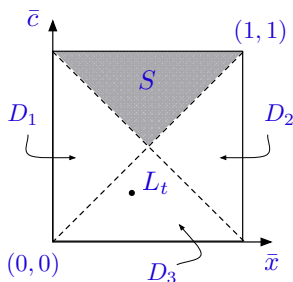$$S = \left\{(z,c) \in [0,1]^2 : u > \max\{c, 1-c\}\right\},$$

Figure 21.1: Partition of the unit square into four regions. The desired set is S.

as depicted in Figure 21.1. As we make predictions $\widehat{\boldsymbol{y}}_t$ and Nature reveals the bits z_t, the point L_t moves within the unit square. We would like to have a strategy that will steer the point towards S no matter what the sequence is. The goal (2.1) can be written in terms of the distance to the set S:

$$\lim_{t \to \infty} d(L_t, S) = 0 \quad \text{almost surely} .$$

The opponent producing the bit sequence, on the other hand, is trying to make sure L_t stays away from S.

Our "steering" strategy can be broken down to three cases according to the position of L_t within the square. For each of the cases, we show that the distance to the set decreases. First, it is easy to check that $\bar{z}_{t+1} - \bar{z}_t = \frac{1}{t+1}(z_{t+1} - \bar{z}_t)$ and hence we may write

$$L_{t+1} = L_t + \frac{1}{t+1}(z_{t+1} - \bar{z}_t, c_{t+1} - \bar{c}_t) .$$

Let $\delta_{t+1} \triangleq \frac{1}{t+1}(z_{t+1} - \bar{z}_t, c_{t+1} - \bar{c}_t)$. The goal is now to show that, in expectation (or almost surely) the update δ_{t+1} moves the point closer to S.

Case 1: If $L_t \in S$, there is nothing to do. We may predict $\widehat{\boldsymbol{y}}_{t+1} = 0$ if $\bar{z}_t < 1/2$, and predict $\widehat{\boldsymbol{y}}_{t+1} = 1$ otherwise.

Case 2: If $L_t \in D_1$, we predict $\widehat{\boldsymbol{y}}_{t+1} = 0$ deterministically. In this case, if $z_{t+1} = 0$ then $\delta_{t+1} = \frac{1}{t+1}(-\bar{z}_t, 1 - \bar{c}_t)$. If, on the other hand, $z_{t+1} = 1$, we have $\delta_{t+1} = \frac{1}{t+1}(1 - \bar{z}_t, -\bar{c}_t)$. These two vectors δ_{t+1} are shown in Figure 21.2. It is easy to check that the distance to the set drops by a multiplicative factor of $1/(t+1)$ irrespective of z_{t+1}. To see this, drop a perpendicular from L_t and from L_{t+1} onto the boundary of S and use properties of similar triangles.

176
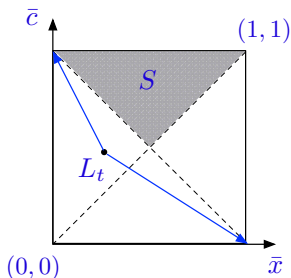
Figure 21.2: When L_t is in the region D_1, the prediction is $\widehat{\boldsymbol{y}}_{t+1} = 0$.

The case $L_t \in D_2$ is identical to this case, except here we deterministically predict $\widehat{\boldsymbol{y}}_{t+1} = 1$.

Case 3: The most interesting situation is $L_t \in D_3$. Observe that our strategy so far has been deterministic, so it better be the case that we use a randomized strategy in D_3. This is indeed true, and the mixed strategy q_{t+1} is defined in a very peculiar fashion. Draw a line from L_t to the vertex $v := (1/2, 1/2)$ of S, and let q_{t+1} be its intersection with the horizontal axis, as shown in Figure 21.3. In the next
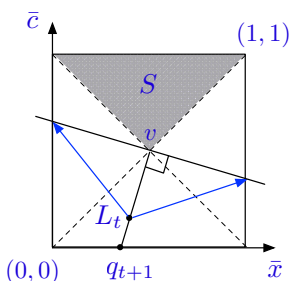
Figure 21.3: Construction of the mixed strategy q_t when $L_t \in D_3$.

lecture we will see why this construction is natural. At this point, let us see why it brings us closer to the set. Clearly, the distance from L_t to the set S is the same as the distance to $(1/2, 1/2)$, while the distance from L_{t+1} to the set is at least the distance to this vertex. Hence, for $\boldsymbol{\ell}_{t+1} = (z_{t+1}, c_{t+1})$, we can write

$$(t+1)^2 d(L_{t+1}, S)^2 \le (t+1)^2 \|L_{t+1} - v\|^2 = \|t(L_t - v) + (\boldsymbol{\ell}_{t+1} - v)\|^2 \qquad (21.3)$$

$$= t^2 d(L_t, S)^2 + \|\boldsymbol{\ell}_{t+1} - v\|^2 + 2t \langle L_t - v, \boldsymbol{\ell}_{t+1} - v \rangle \qquad (21.4)$$

And now, for the magic. We claim that $\langle L_t - v, \ell_{t+1} - v \rangle = 0$ in expectation (conditioned on the value z_{t+1} and the past). Indeed,

$$\mathbb{E}_{\widehat{y}_{t+1} \sim q_{t+1}} \ell_{t+1} = \begin{cases} (0, 1 - q_{t+1}) & \text{if } z_{t+1} = 0 \\ (1, q_{t+1}) & \text{if } z_{t+1} = 1 \end{cases} \tag{21.5}$$

which correspond to the two intersections of the line A (see Figure 21.3) with the sides of the unit square. In both cases, $\mathbb{E}_{\widehat{y}_{t+1}} \langle L_t - v, \ell_{t+1} - v \rangle = 0$ since the line A is perpendicular to $L_t - v$.

Finally, $\|\ell_{t+1} - v\|^2 \leq 1/2$, and summing Eq. (21.3) for $t = 1, \ldots, n-1$ and canceling the terms we get

$$n^2 d(L_n, S)^2 \leq d(L_1, S)^2 + n/2 + \sum_{t=1}^{n} t \cdot M_t \tag{21.6}$$

with M_t a bounded martingale difference sequence. An application of Hoeffding-Azuma gives a high-probability bound and an application of Borel-Cantelli yields the almost sure statement.

21.3   Follow the Regularized Leader

Since the Exponential Weights Algorithm will never produce deterministic strategies, it is clear that the Blackwell's method is genuinely different. The reason for non-deterministic strategies of the Exponential Weights method is simple: the update rule never pushes the mixed strategy to the corner of the probability simplex, always keeping a nonzero weight even on hopelessly bad experts.

What if instead of the Follow the Regularized Leader with entropic regularization, which yields the Exponential Weights Algorithm as shown in Example 19, we use Euclidean regularization, which yields a gradient-descent type update as shown in Example 18?

To phrase the Follow the Regularized Leader (FTRL) method, we need to define the loss as a linear or convex function. Of course, the indicator loss $\mathbf{I}\{\widehat{y}_t \neq z_t\}$ is neither of these, but the trick is to consider the linearized problem where the choice of the learner is $q_t \in [0, 1]$, interpreted as the probability of predicting $\widehat{y}_t = 1$. Since $\mathbf{I}\{\widehat{y}_t \neq z_t\} = \widehat{y}_t + z_t - 2\widehat{y}_t z_t$ for $\widehat{y}_t, z_t \in \{0, 1\}$, the expected loss can be written as

$$\mathbb{E}_{\widehat{y}_t \sim q_t} \mathbf{I}\{\widehat{y}_t \neq z_t\} = q_t \cdot (1 - 2z_t) + z_t =: \ell(q_t, z_t)$$

for $q_t \in [0,1]$. We can now define the FTRL method

$$q_{t+1} = \operatorname*{argmin}_{q \in [0,1]} \left\{ \sum_{s=1}^{t} q \cdot (1 - 2z_s) + \eta^{-1} \frac{1}{2} \|q - 1/2\|^2 \right\} \tag{21.7}$$

with the Euclidean regularizer centered at $q = 1/2$. The unconstrained problem over the real line has solution at $\tilde{q}_{t+1} = \frac{1}{2} - \eta \sum_{s=1}^{t} (1 - 2z_s)$ which is subsequently clipped (or projected to) the interval $[0,1]$. If this clipping happens, the strategy becomes deterministic, either $q_t = 1$ or $q_t = 0$. Clipping is needed precisely when $|\sum_{s=1}^{t} (1 - 2z_s)| < \frac{1}{2\eta}$, or, equivalently,

$$|\bar{z}_t - 1/2| \geq \frac{1}{4t\eta} \ .$$

As shown in Figure 21.4, the FTRL strategy gives deterministic prediction when the empirical frequency $\bar{z}_t$ is outside the band of width $\frac{1}{2t\eta}$, centered at $1/2$. The typical guarantee of $O(1/\sqrt{n})$ for the regret of FTRL sets $\eta = c/\sqrt{n}$ for some constant c (or, in a time-changing manner, $\eta_t = c/\sqrt{t}$) and thus the width of the region where the prediction is randomized is of the order $1/\sqrt{n}$.
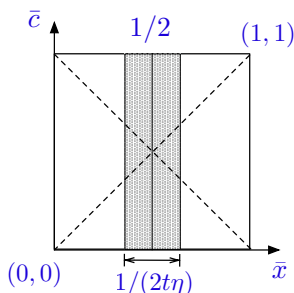
Figure 21.4: FTRL with a fixed learning rate η gives a mixed strategy only in the small band around $1/2$.

Since FTRL enjoys small regret, we have derived yet another strategy that attains the goal in (2.1). Observe, however, that the FTRL strategy with the Euclidean regularizer is different from both the Exponential Weights Algorithm, and from Blackwell's method. If we are to replicate the behavior of the latter, the information about the frequency of our correct predictions $\bar{c}_t$ needs to be taken into account. This information (or, *statistic* about the past) cannot be deduced from $\bar{z}_t$ alone, and so FTRL or Mirror Descent methods seem to be genuinely distinct from Blackwell's algorithm based on $(\bar{z}_t, \bar{c}_t)$.

179

It turns out that the extra information about the frequency $\bar{c}_t$ of correct predictions can be used to set the learning rate η in Follow the Regularized Leader. With the time-changing learning rate

$$\eta_t = \frac{1}{4t|\bar{c}_t - 1/2|}$$

the FTRL method

$$q_{t+1} = \underset{q \in [0,1]}{\operatorname{argmin}} \left\{ \sum_{s=1}^{t} q \cdot (1 - 2z_s) + \eta_t^{-1} \frac{1}{2} \|q - 1/2\|^2 \right\} \tag{21.8}$$

becomes exactly the Blackwell's algorithm. Let's see why this is so. First, let us check the case when the optimum at (21.8) is achieved at the pure strategy $q_t = 1$ or $q_t = 0$. As argued earlier, this occurs when

$$|\bar{z}_t - 1/2| \geq \frac{1}{4t\eta_t} = |\bar{c}_t - 1/2| \ .$$

This exactly corresponds to the regions D_1 and D_2 in Figure 21.1, thus matching the behavior of Blackwell's method. It remains to check the behavior in D_3. Setting the derivative in (21.8) to zero,

$$q_t = \frac{1}{2} + (2t\eta_t)(\bar{z}_t - 1/2) = \frac{1}{2} + \frac{\bar{z}_t - 1/2}{2|\bar{c}_t - 1/2|} \ .$$

We need to check that this is the same value as that obtained geometrically in Figure 21.3. It is indeed the case, as can be seen from similar triangles: the ratio of $q_t - 1/2$ to $1/2$ is equal to the ratio of $\bar{z}_t - 1/2$ to $|\bar{c}_t - 1/2|$.

21.4 Discussion

We have described three different methods for the problem of $\{0,1\}$-sequence prediction. The Exponential Weights Algorithm puts exponentially more weight on the bit that occurs more often, but never lets go of the other bit, and the strategy is always randomized. The Follow the Regularized Leader method with a Euclidean regularizer produces a randomized strategy only in the narrow band around $1/2$. A variant of FTRL which adapts the learning rate with respect to the proportion of correct predictions yields a randomized strategy in a triangular region D_3 (Figure 21.1), and deterministic prediction otherwise. Further, this behavior matches the Blackwell's method, based on a geometric construction.

In the worst case, the performance (or, regret) of the three methods described above is the same, up to a multiplicative constant. However, the methods that use the extra information about the proportion $\bar{c}_t$ of correct predictions may have better convergence properties for "benign" sequences. Such adaptive procedures have been analyzed in the literature.

The beauty of Blackwell's proof is its geometric simplicity and the rather surprising construction of the mixed strategy q_t. As we will see in the next lecture, the construction appears out of a generalization of the minimax theorem.

Of course, our application of the Exponential Weights Algorithm or Follow the Regularized Leader is a shorter proof, but remember that we spent some time building these hammers. Is there a similar hammer based on the ideas of approaching a desired set? The answer is yes, and it is called the Blackwell's Approachability Theorem. Using this theorem, one can actually prove a wide range of results in repeated games that go beyond the "regret" formulation. What is interesting, Blackwell's Approachability itself can be proved using online convex optimization algorithms whose performance is defined in terms of regret. This result points to an equivalence of these big "hammers". Finally, the sequential symmetrization tools we had developed earlier in the course can be used to prove Blackwell approachability in even more generality, without exhibiting an algorithm.

21.5 Can we *derive* an algorithm for bit prediction?

Let us consider the minimax regret formulation:

$$
\mathcal{V} = \min_{q_1 \in [0,1]} \max_{z_1} \mathbb{E}_{\widehat{y}_1 \sim q_1} \dots \min_{q_n \in [0,1]} \max_{z_n} \mathbb{E}_{\widehat{y}_n \sim q_n} \left\{ \frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{\widehat{y}_t \neq z_t\} - \min_{f \in \{0,1\}} \frac{1}{n} \sum_{t=1}^{n} \mathbf{I}\{f \neq z_t\} \right\}
$$

Observe that $n\mathcal{V}$ can be written recursively as

$$
\mathcal{V}(z_1, \dots, z_{t-1}) = \min_{q_t \in [0,1]} \max_{z_t} \mathbb{E}_{\widehat{y}_t \sim q_t} \left\{ \mathbf{I}\{\widehat{y}_t \neq z_t\} + \mathcal{V}(z_1, \dots, z_t) \right\}
$$

with

$$
\mathcal{V}(z_1, \dots, z_n) = - \min_{f \in \{0,1\}} \sum_{t=1}^{n} \mathbf{I}\{f \neq z_t\}.
$$

21.5 Can we *derive* an algorithm for bit prediction?

We leave it as an exercise to check that

$$n\mathcal{V} = \mathcal{V}(\emptyset)$$

with these definitions. The recursion can in fact be solved exactly:

$$
\begin{aligned}
\mathcal{V}(z_1,\ldots,z_{t-1}) &= \min_{q_t\in[0,1]} \max_{z_t\in\{0,1\}} \mathop{\mathbb{E}}_{\widehat{\mathbf{y}}_t\sim q_t} \left\{ \mathbf{I}\{\widehat{\mathbf{y}}_t \neq z_t\} + \mathcal{V}(z_1,\ldots,z_t) \right\} \\
&= \min_{q_t\in[0,1]} \max_{z_t\in\{0,1\}} \left\{ q_t\cdot(1-2z_t) + z_t + \mathcal{V}(z_1,\ldots,z_t) \right\} \\
&= \min_{q_t\in[0,1]} \max \left\{ -q_t + 1 + \mathcal{V}(z_1,\ldots,z_{t-1},1)\,,\, q_t + 0 + \mathcal{V}(z_1,\ldots,z_{t-1},0) \right\}
\end{aligned}
$$

The solution is obtained by equating the two terms

$$q_t^* = \frac{1}{2} + \frac{\mathcal{V}(z_1,\ldots,z_{t-1},1) - \mathcal{V}(z_1,\ldots,z_{t-1},0)}{2} \tag{21.9}$$

which gives

$$\mathcal{V}(z_1,\ldots,z_{t-1}) = \frac{1}{2} + \mathbb{E}\mathcal{V}(z_1,\ldots,z_{t-1},b_t)\,, \tag{21.10}$$

b_t is fair coin. Then, working backwards, the minimax value of the problem is

$$
\begin{aligned}
\mathcal{V} &= \frac{n}{2} + \mathbb{E}\mathcal{V}(b_1,\ldots,b_n) \\
&= \frac{n}{2} + \mathbb{E}\left\{ -\min_{f\in\{0,1\}} \sum_{t=1}^{n} \mathbf{I}\{f \neq b_t\} \right\} \\
&= \mathbb{E}\max_{f\in\{0,1\}} \left\{ \sum_{t=1}^{n} \mathbb{E}\mathbf{I}\{f \neq b\} - \mathbf{I}\{f \neq b_t\} \right\} \\
&= \mathbb{E}\max_{f\in\{0,1\}} \left\{ \sum_{t=1}^{n} \epsilon_t f \right\}
\end{aligned}
$$

where ϵ_t are ± 1 Rademacher random variables. From Hoeffding's inequality and a union bound,

$$\mathcal{V} \leq \sqrt{\frac{\ln 2}{n}},$$

which is the bound we obtained with exponential weights algorithm.

21.5 Can we *derive* an algorithm for bit prediction?

We now turn to the question of obtaining computationally feasible algorithms. A calculation similar to the above gives

$$\mathcal{V}(z_1,\ldots,z_t) = \mathbb{E} \max_{f\in\{0,1\}} \left\{ \sum_{s=t+1}^{n} \epsilon_s f - \sum_{s=1}^{t} \mathbf{I}\{f \neq z_s\} \right\}$$

that can be plugged into the algorithm

$$q_t^* = \operatorname*{argmin}_{q_t\in[0,1]} \max_{z_t\in\{0,1\}} \mathbb{E}_{\widehat{\mathbf{y}}_t\sim q_t} \left\{ \mathbf{I}\{\widehat{\mathbf{y}}_t \neq z_t\} + \mathcal{V}(z_1,\ldots,z_t) \right\}$$

However, the expression for $\mathcal{V}(z_1,\ldots,z_t)$ is still computationally expensive, as it involves averaging over random signs. The key idea is to introduce relaxations

$$\mathcal{V}(z_1,\ldots,z_t) \leq \mathbf{Rel}_n(z_1,\ldots,z_t)$$

Not any upper bound will work, but one can see that it is enough to ensure *admissibility*:

$$\mathbf{Rel}_n(z_1,\ldots,z_{t-1}) \geq \min_{q_t\in[0,1]} \max_{z_t} \mathbb{E}_{\widehat{\mathbf{y}}_t\sim q_t} \left\{ \mathbf{I}\{\widehat{\mathbf{y}}_t \neq z_t\} + \mathbf{Rel}_n(z_1,\ldots,z_t) \right\} \qquad (*)$$

and

$$\mathbf{Rel}_n(z_1,\ldots,z_n) \geq - \min_{f\in\{0,1\}} \sum_{t=1}^{n} \mathbf{I}\{f \neq z_t\}$$

Any algorithm that solves $(*)$ guarantees performance bound of $\mathbf{Rel}_n(\emptyset)$. The main question now is: How do we come up with relaxations that lead to computationally feasible algorithms?

Note that for maximum, a tight upper bound is achieved by the so-called softmax. For any $\eta > 0$,

$$\begin{aligned}
\mathcal{V}(z_1,\ldots,z_t) &= \mathbb{E} \max_{f\in\{0,1\}} \left\{ \sum_{s=t+1}^{n} \epsilon_s f - \sum_{s=1}^{t} \mathbf{I}\{f \neq z_s\} \right\} \\
&\leq \frac{1}{\eta} \ln \mathbb{E} \sum_{f\in\{0,1\}} \exp\left\{ \eta \sum_{s=t+1}^{n} \epsilon_s f - \eta \sum_{s=1}^{t} \mathbf{I}\{f \neq z_s\} \right\} \qquad (21.11)
\end{aligned}$$

Let us now fix $f \in \{0,1\}$ and write

$$\mathbb{E} \exp\left\{ \eta \sum_{s=t+1}^{n} \epsilon_s f - \eta \sum_{s=1}^{t} \mathbf{I}\{f \neq z_s\} \right\} = \exp\left\{ -\eta \sum_{s=1}^{t} \mathbf{I}\{f \neq z_s\} \right\} \times \mathbb{E} \exp\left\{ \eta \sum_{s=t+1}^{n} \epsilon_s f \right\}$$

$$(21.12)$$

By Lemma [9.2](#),

$$\mathbb{E}\exp\left\{\eta\sum_{s=t+1}^{n}\epsilon_s f\right\} \le \frac{\eta^2(n-t)}{2}$$

(the same analysis also holds for $f \in \{-1,1\}$ and so the constant can be improved). Plugging into (21.11),

$$\mathcal{V}(z_1,\ldots,z_t) \le \inf_{\eta>0}\left\{\frac{1}{\eta}\ln\sum_{f\in\{0,1\}}\exp\left\{-\eta\sum_{s=1}^{t}\mathbf{I}\{f\ne z_s\}\right\} + \frac{\eta}{2}(n-t)\right\} \qquad (21.13)$$

✎ **Exercise 21.1** ($\star$)**.** Prove that the right-hand side of (21.13) is an admissible relaxation and it leads to a parameter-free version of the Exponential Weights algorithm.

21.6   The Mind Reading Machine

The fact that the entropy of the source is related being able to predict the sequences has long been recognized. It is, therefore, no surprise that Shannon was interested in the problem of prediction of binary sequences. Quoting from (Feder, Merhav, and Gutman, 2005)[1]:

> In the early 50's, at Bell Laboratories, David Hagelbarger built a simple "mind reading" machine, whose purpose was to play the "penny matching" game. In this game, a player chooses head or tail, while a "mind reading" machine tries to predict and match his choice. Surprisingly, as Robert Lucky tells in his book "Silicon Dreams", Hagelbarger's simple, 8-state machine, was able to match the "pennies" of its human opponent 5,218 times over the course of 9,795 plays. Random guessing would lead to such a high success rate with a probability less than one out of 10 billion! Shannon, who was interested in prediction, information, and thinking machines, closely followed Hagelbarger's machine, and eventually built his own stripped-down version of the machine, having the same states, but one that used a simpler strategy at each state. As the legend goes, in a duel between the two machines, Shannon's machine won by a slight margin! No one knows if this was due to a superior algorithm or just a chance happening associated

---

[1] http://backup.itsoc.org/review/meir/node1.html

with the specific sequence at that game. In any event, the success of both these machines against "untrained" human opponents was explained by the fact that the human opponents cannot draw completely random bits. Certainly, as Shannon himself noted, his machine was beatable by the best possible player by a ratio of three-to-one. This raises the question, can one design a better "mind reading" machine? What is the best one can do in predicting the next outcome of the opponent? How is it related to the "randomness" of the opponent sequence?

The techniques we developed in this course can be directly applied for building such a machine. Of course, it needs to take advantage of the fact that the sequence input by an (untrained) person is likely to be non-random. From our earlier proof, we already see that we should be able to predict the sequence better than chance if there are more 0'1 or 1's.

22

Algorithmic Framework for Sequential Prediction

So far in the course, we have seen very general non-constructive upper bounds on the value of the prediction problem, as well as specialized algorithms that seem to magically match some of these bounds in terms of their regret. The algorithms have been developed over the course of several decades, yet for each new setting the researcher is tasked with a rather difficult problem of coming up with a new procedure almost from scratch. The online convex optimization framework successfully guides in the development of algorithms for many convex problems, but what is really behind all the methods? It turns out that, based on the non-constructive upper bounds, we can develop an algorithmic framework that captures a very wide range of known methods. The framework also gives a general prescription for the development of new algorithms. We refer to [43] for more details.

We first study the general prediction problem with an abstract set $\mathcal{D}$ of learner's moves and $\mathcal{Z}$ being the set of moves of Nature. We will devote a later section to the more specific problem of supervised learning.

Recall that the online protocol dictates that on every round $t = 1, \ldots, n$ the learner and Nature simultaneously choose $\widehat{\boldsymbol{y}}_t \in \mathcal{D}$, $z_t \in \mathcal{Z}$, and observe each other's actions. Unless specified otherwise, we will assume $\mathcal{D} = \mathcal{F}$. For the sake of brevity, let us consider the unnormalized minimax value of the prediction problem:

$$\mathcal{V}_n(\mathcal{F}) = \inf_{q_1 \in \Delta(\mathcal{D})} \sup_{z_1 \in \mathcal{Z}} \mathbb{E}_{\widehat{\boldsymbol{y}}_1 \sim q_1} \ldots \inf_{q_n \in \Delta(\mathcal{D})} \sup_{z_n \in \mathcal{Z}} \mathbb{E}_{\widehat{\boldsymbol{y}}_n \sim q_n} \left[\sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \boldsymbol{\ell}(f, z_t) \right]$$

(22.1)

The minimax formulation immediately gives rise to the optimal algorithm that solves the minimax expression at every round t. That is, after witnessing $z_1, \ldots, z_{t-1}$, the algorithm returns

$$\underset{q \in \Delta(\mathcal{D})}{\operatorname{argmin}} \left\{ \sup_{z_t} \underset{\hat{\mathbf{y}}_t \sim q}{\mathbb{E}} \ldots \inf_{q_n} \sup_{z_n} \underset{\hat{\mathbf{y}}_n}{\mathbb{E}} \left[\sum_{i=t}^{n} \boldsymbol{\ell}(\hat{\mathbf{y}}_i, z_i) - \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} \boldsymbol{\ell}(f, z_i) \right] \right\} \tag{22.2}$$

$$= \underset{q \in \Delta(\mathcal{D})}{\operatorname{argmin}} \left\{ \sup_{z_t} \underset{\hat{\mathbf{y}}_t \sim q}{\mathbb{E}} \left[\boldsymbol{\ell}(\hat{\mathbf{y}}_t, z_t) + \inf_{q_{t+1}} \sup_{z_{t+1}} \underset{\hat{\mathbf{y}}_{t+1}}{\mathbb{E}} \ldots \inf_{q_n} \sup_{z_n} \underset{\hat{\mathbf{y}}_n}{\mathbb{E}} \left[\sum_{i=t+1}^{n} \boldsymbol{\ell}(\hat{\mathbf{y}}_i, z_i) - \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} \boldsymbol{\ell}(f, z_i) \right] \right] \right\}$$

Henceforth, if the quantification in inf and sup is omitted, it will be understood that z_t, $\hat{\mathbf{y}}_t$, p_t, q_t range over $\mathcal{Z}$, $\mathcal{D}$, $\Delta(\mathcal{Z})$, $\Delta(\mathcal{D})$, respectively. Moreover, $\mathbb{E}_{z_t}$ is with respect to p_t while $\mathbb{E}_{\hat{\mathbf{y}}_t}$ is with respect to q_t. The first sum in (22.2) starts at $i = t$ since the partial loss $\sum_{i=1}^{t-1} \boldsymbol{\ell}(\hat{\mathbf{y}}_i, z_i)$ has been fixed. We now notice a recursive form for defining the value of the game. For any $t \in [n-1]$ and any given prefix $z_1, \ldots, z_t \in \mathcal{Z}$ define the *conditional value*

$$\mathcal{V}_n(z_1, \ldots, z_t) \triangleq \inf_{q \in \Delta(\mathcal{D})} \sup_{z \in \mathcal{Z}} \left\{ \underset{\hat{\mathbf{y}} \sim q}{\mathbb{E}} \left[\boldsymbol{\ell}(\hat{\mathbf{y}}, z) \right] + \mathcal{V}_n(z_1, \ldots, z_t, z) \right\}$$

where

$$\mathcal{V}_n(z_1, \ldots, z_n) \triangleq -\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \boldsymbol{\ell}(f, z_t) \quad \text{and} \quad \mathcal{V}_n(\mathcal{F}) = \mathcal{V}_n(\emptyset).$$

The *minimax optimal* algorithm specifying the mixed strategy of the player can be written succinctly

$$q_t = \underset{q \in \Delta(\mathcal{D})}{\operatorname{argmin}} \sup_{z \in \mathcal{Z}} \left\{ \mathbb{E}_{\hat{\mathbf{y}} \sim q}[\boldsymbol{\ell}(\hat{\mathbf{y}}, z)] + \mathcal{V}_n(z_1, \ldots, z_{t-1}, z) \right\}. \tag{22.3}$$

This dynamic programming formulation has appeared in the literature, but now we have tools to study the conditional value of the game. We will show that various upper bounds on $\mathcal{V}_n(z_1, \ldots, z_{t-1}, z)$ yield an array of algorithms, some with better computational properties than others. In this way, the non-constructive approach we developed earlier in the course to upper bound the value of the game directly translates into algorithms.

The minimax algorithm in (22.3) can be interpreted as choosing the best decision that takes into account the present loss and the worst-case future. We then realize that the conditional value of the game serves as a "regularizer", and thus well-known online learning algorithms such as Exponential Weights, Mirror Descent and Follow-the-Regularized-Leader arise as relaxations rather than a "method that just works".

22.1   Relaxations

A *relaxation* $\mathbf{Rel}_n\,()$ is a sequence of functions $\mathbf{Rel}_n\,(z_1,\dots,z_t)$ for each $t \in [n]$. A relaxation will be called *admissible* if for any $z_1,\dots,z_n \in \mathcal{Z}$,

$$\mathbf{Rel}_n\,(z_1,\dots,z_{t-1}) \geq \inf_{q_t \in \Delta(\mathcal{D})} \sup_{z_t \in \mathcal{Z}} \left\{ \underset{\widehat{\boldsymbol{y}}_t \sim q_t}{\mathbb{E}} \left[\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) \right] + \mathbf{Rel}_n\,(z_1,\dots,z_{t-1},z_t) \right\} \quad (22.4)$$

for all $t \in [n-1]$, and

$$\mathbf{Rel}_n\,(z_1,\dots,z_n) \geq - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \boldsymbol{\ell}(f,z_t).$$

A simple inductive argument shows that $\mathcal{V}_n\,(z_1,\dots,z_t) \leq \mathbf{Rel}_n\,(z_1,\dots,z_t)$ for any t and any $z_1,\dots,z_n$. Thus, $\mathcal{V}_n$ is the smallest admissible relaxation. This can indeed be another definition for the value of the game.

A strategy q that minimizes the expression in (22.4) defines a minimax optimal algorithm for the relaxation $\mathbf{Rel}$. However, minimization need not be exact: any q that satisfies the admissibility condition (22.4) is a valid method, and we will say that such an algorithm is *admissible with respect to the relaxation* $\mathbf{Rel}$.

Meta-Algorithm

Parameters: Admissible relaxation $\mathbf{Rel}$

At each $t = 1$ to n, compute

$$q_t = \operatorname*{argmin}_{q \in \Delta(\mathcal{D})} \sup_{z_t \in \mathcal{Z}} \left\{ \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q}[\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t)] + \mathbf{Rel}_n\,(z_1,\dots,z_{t-1},z_t) \right\} \quad (22.5)$$

and play $\widehat{\boldsymbol{y}}_t \sim q_t$. Receive z_t from Nature.

Let $\mathbf{Reg}_n$ stand for the unnormalized regret

$$\mathbf{Reg}_n \triangleq \sum_{t=1}^{n} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \boldsymbol{\ell}(f, z_t)$$

and let $\mathbf{Rel}_n\,(\mathcal{F}) \triangleq \mathbf{Rel}_n\,(\varnothing)$

Proposition 22.1. *Let* $\mathbf{Rel}\,()$ *be an admissible relaxation. For any admissible algorithm with respect to* $\mathbf{Rel}\,()$*, including the* Meta-Algorithm*, irrespective of the strategy of the adversary,*

$$\sum_{t=1}^{n} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \boldsymbol{\ell}(f, z_t) \leq \mathbf{Rel}_n\,(\mathcal{F}) \;, \quad (22.6)$$

and therefore,

$$\mathbb{E}[\boldsymbol{Reg}_n] \leq \mathbf{Rel}_n(\mathcal{F}) \ .$$

We also have that

$$\mathcal{V}_n(\mathcal{F}) \leq \mathbf{Rel}_n(\mathcal{F}) \ .$$

If $a \leq \ell(f, z) \leq b$ for all $f \in \mathcal{F}, z \in \mathcal{Z}$, the Hoeffding-Azuma inequality yields, with probability at least $1 - \delta$,

$$\boldsymbol{Reg}_n \leq \mathbf{Rel}_n(\mathcal{F}) + (b - a)\sqrt{n/2 \cdot \log(2/\delta)} \ .$$

Further, if for all $t \in [n]$, the admissible strategies q_t are deterministic,

$$\boldsymbol{Reg}_n \leq \mathbf{Rel}_n(\mathcal{F}) \ .$$

Proof of Proposition 22.1. By definition,

$$\sum_{t=1}^{n} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \ell(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, z_t) \leq \sum_{t=1}^{n} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \ell(\widehat{\boldsymbol{y}}_t, z_t) + \mathbf{Rel}_n(z_1, \ldots, z_n) \ .$$

Peeling off the n-th expected loss, we have

$$\sum_{t=1}^{n} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \ell(\widehat{\boldsymbol{y}}_t, z_t) + \mathbf{Rel}_n(z_1, \ldots, z_n) \leq \sum_{t=1}^{n-1} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \ell(\widehat{\boldsymbol{y}}_t, z_t) + \left\{ \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \ell(\widehat{\boldsymbol{y}}_t, z_t) + \mathbf{Rel}_n(z_1, \ldots, z_n) \right\}$$

$$\leq \sum_{t=1}^{n-1} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \ell(\widehat{\boldsymbol{y}}_t, z_t) + \mathbf{Rel}_n(z_1, \ldots, z_{n-1})$$

where we used the fact that q_n is an admissible algorithm for this relaxation, and thus the last inequality holds for any choice z_n of the opponent. Repeating the process, we obtain

$$\sum_{t=1}^{n} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \ell(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f, z_t) \leq \mathbf{Rel}_n(\mathcal{F}) \ .$$

We remark that the left-hand side of this inequality is random, while the right-hand side is not. Since the inequality holds for any realization of the process, it also holds in expectation. The inequality

$$\mathcal{V}_n(\mathcal{F}) \leq \mathbf{Rel}_n(\mathcal{F})$$

holds by unwinding the value recursively and using admissibility of the relaxation. The high-probability bound is an immediate consequences of (22.6) and the Hoeffding-Azuma inequality for bounded martingales. The last statement is immediate. $\square$

22.1 Relaxations

For many problems a tight relaxation (sometimes within a factor of 2) is achieved through symmetrization. Define the *conditional Sequential Rademacher complexity*

$$\mathcal{R}_n(z_1,\ldots,z_t) = \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f\in\mathcal{F}} \left[2\sum_{s=t+1}^{n} \epsilon_s \boldsymbol{\ell}(f,\mathbf{z}_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \boldsymbol{\ell}(f,z_s) \right]. \quad (22.7)$$

Here the supremum is over all $\mathcal{Z}$-valued binary trees of depth $n-t$. One may view this complexity as a partially symmetrized version of the sequential Rademacher complexity $\mathcal{R}^{seq}(\mathcal{F})$. We shall refer to the term involving the tree $\mathbf{z}$ as the "future" and the term being subtracted off – as the "past". This indeed corresponds to the fact that the quantity is conditioned on the already observed $z_1,\ldots,z_t$, while for the future we have the worst possible binary tree.[1]

Proposition 22.2. *Conditional Sequential Rademacher complexity is an admissible relaxation.*

Proof. Denote $L_t(f) = \sum_{s=1}^{t} \boldsymbol{\ell}(f,z_s)$. The first step of the proof is an application of the minimax theorem (we assume the necessary conditions hold):

$$\inf_{q_t\in\Delta(\mathcal{D})} \sup_{z_t\in\mathcal{Z}} \left\{ \mathbb{E}_{\widehat{\boldsymbol{y}}_t\sim q_t} \left[\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t,z_t)\right] + \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f\in\mathcal{F}} \left[2\sum_{s=t+1}^{n} \epsilon_s\boldsymbol{\ell}(f,\mathbf{z}_{s-t}(\epsilon_{t+1:s-1})) - L_t(f)\right] \right\}$$

$$= \sup_{p_t\in\Delta(\mathcal{Z})} \inf_{\widehat{\boldsymbol{y}}_t\in\mathcal{F}} \left\{ \mathbb{E}_{z_t\sim p_t} \left[\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t,z_t)\right] + \mathbb{E}_{z_t\sim p_t} \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f\in\mathcal{F}} \left[2\sum_{s=t+1}^{n} \epsilon_s\boldsymbol{\ell}(f,\mathbf{z}_{s-t}(\epsilon_{t+1:s-1})) - L_t(f)\right] \right\}$$

For the sake of brevity, let us use the notation $A(f) = 2\sum_{s=t+1}^{n} \epsilon_s\boldsymbol{\ell}(f,\mathbf{z}_{s-t}(\epsilon_{t+1:s-1}))$. Then, for any $p_t \in \Delta(\mathcal{Z})$, the infimum over $\widehat{\boldsymbol{y}}_t$ of the above expression is equal to

$$\mathbb{E}_{z_t\sim p_t} \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f\in\mathcal{F}} \left[A(f) - L_{t-1}(f) + \inf_{\widehat{\boldsymbol{y}}_t\in\mathcal{F}} \mathbb{E}_{z_t\sim p_t} \left[\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t,z_t)\right] - \boldsymbol{\ell}(f,z_t) \right]$$

$$\leq \mathbb{E}_{z_t\sim p_t} \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f\in\mathcal{F}} \left[A(f) - L_{t-1}(f) + \mathbb{E}_{z_t\sim p_t} \left[\boldsymbol{\ell}(f,z_t)\right] - \boldsymbol{\ell}(f,z_t) \right]$$

$$\leq \mathbb{E}_{z_t,z_t'\sim p_t} \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f\in\mathcal{F}} \left[A(f) - L_{t-1}(f) + \boldsymbol{\ell}(f,z_t') - \boldsymbol{\ell}(f,z_t) \right]$$

---

[1] It is somewhat cumbersome to write out the indices on $\mathbf{z}_{s-t}(\epsilon_{t+1:s-1})$ in (22.7), so we will instead use $\mathbf{z}_s(\epsilon)$ for $s = 1,\ldots,n-t$, whenever this does not cause confusion.

We now argue that the independent z_t and z_t' have the same distribution p_t, and thus we can introduce a random sign ϵ_t. The above expression then equals to
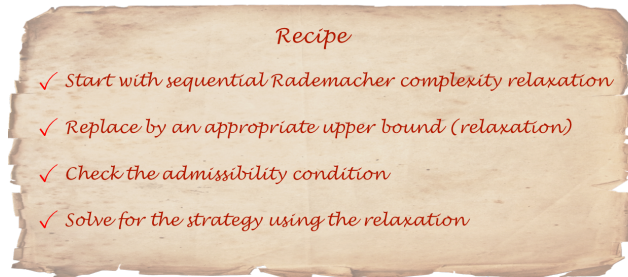
$$\mathop{\mathbb{E}}_{z_t, z_t' \sim p_t} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left[A(f) - L_{t-1}(f) + \epsilon_t (\boldsymbol{\ell}(f, z_t') - \boldsymbol{\ell}(f, z_t)) \right]$$

$$\leq \sup_{z_t, z_t' \in \mathcal{Z}} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left[A(f) - L_{t-1}(f) + \epsilon_t (\boldsymbol{\ell}(f, z_t') - \boldsymbol{\ell}(f, z_t)) \right]$$

where we upper bounded the expectation by the supremum. Splitting the resulting expression into two parts, we arrive at the upper bound of

$$2 \sup_{z_t \in \mathcal{Z}} \mathbb{E}_{\epsilon_t} \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left[\frac{1}{2}(A(f) - L_{t-1}(f)) + \epsilon_t \boldsymbol{\ell}(f, z_t) \right] = \mathcal{R}_n(z_1, \ldots, z_{t-1}) .$$

The last equality is easy to verify, as we are effectively adding a root z_t to the two subtrees, for $\epsilon_t = +1$ and $\epsilon_t = -1$, respectively.

$$\square$$

The conditional sequential Rademacher complexity can be thought of as a tight relaxation and a good starting point for developing a computationally-attractive method. The recipe is then as follows:

Recipe

✓ *Start with sequential Rademacher complexity relaxation*

✓ *Replace by an appropriate upper bound (relaxation)*

✓ *Check the admissibility condition*

✓ *Solve for the strategy using the relaxation*

We now show that several well-known methods arise by following the recipe. Next few chapters are devoted to delineating certain techniques for following the recipe and to deriving new prediction methods.

22.1.1 Follow the Regularized Leader / Dual Averaging

In the setting of online linear optimization, the loss is $\boldsymbol{\ell}(f, z) = \langle f, z \rangle$. Let $\mathcal{F} = \mathcal{Z} = B_2^d$ a unit ball in $\mathbb{R}^d$. The conditional sequential Rademacher complexity can be

written as

$$\mathcal{R}_n(z_1,\ldots,z_t) = \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left\langle f, \ 2 \sum_{s=t+1}^{n} \epsilon_s \mathbf{z}_{s-t}(\epsilon_{t+1:s-1}) - \sum_{s=1}^{t} z_s \right\rangle \tag{22.8}$$

which can be written simply as

$$\sup_{\mathbf{z}} \mathbb{E} \left\| 2 \sum_{s=t+1}^{n} \epsilon_s \mathbf{z}_{s-t}(\epsilon_{t+1:s-1}) - \sum_{s=1}^{t} z_s \right\|$$

(see Eq. (10.1)), where the norm is Euclidean. Of course, this relaxation is not computationally attractive because of the supremum over $\mathbf{z}$, so we need to pass to an upper bound that has nicer properties. We will remove the dependence on $\mathbf{z}$ via probabilistic inequalities.

Observe that for any $\mathbf{z}$, the expected norm in the above expression is upper bounded via Jensen's inequality by

$$\left(\mathbb{E} \left\| 2 \sum_{s=t+1}^{n} \epsilon_s \mathbf{z}_{s-t}(\epsilon_{t+1:s-1}) - \sum_{s=1}^{t} z_s \right\|^2 \right)^{1/2} = \left(\left\| \sum_{s=1}^{t} z_s \right\|^2 + 4 \sum_{s=t+1}^{n} \mathbb{E} \|\mathbf{z}_{s-t}(\epsilon_{t+1:s-1})\|^2 \right)^{1/2}$$

$$\leq \left(\left\| \sum_{s=1}^{t} z_s \right\|^2 + 4(n-t) \right)^{1/2} = \mathbf{Rel}_n(z_1,\ldots,z_t)$$

where the first equality follows by expanding the square and letting the expectation act on the cross-terms, and the inequality — because the tree $\mathbf{z}$ is B_2^d-valued. We now take the last expression as the relaxation. To show admissibility, we need to prove that

$$\inf_{\widehat{\mathbf{y}}_t \in \mathcal{F}} \sup_{z_t \in \mathcal{Z}} \left\{ \langle \widehat{\mathbf{y}}_t, z_t \rangle + \left(\left\| \sum_{s=1}^{t} z_s \right\|^2 + 4(n-t) \right)^{1/2} \right\} \leq \left(\left\| \sum_{s=1}^{t-1} z_s \right\|^2 + 4(n-t+1) \right)^{1/2} \tag{22.9}$$

Let $v = \sum_{s=1}^{t-1} z_s$. For the choice

$$\widehat{\mathbf{y}}_t = -\frac{v}{2\left(\|v\|^2 + 4(n-t+1) \right)^{1/2}}, \tag{22.10}$$

the left-hand side of (22.9) is

$$\sup_{z_t \in \mathcal{Z}} \left\{ -\frac{\langle v, z_t \rangle}{2\left(\|v\|^2 + 4(n-t+1) \right)^{1/2}} + \left(\|v + z_t\|^2 + 4(n-t) \right)^{1/2} \right\} \tag{22.11}$$

$$\leq \sup_{z_t \in \mathcal{Z}} \left\{ -\frac{\langle v, z_t \rangle}{2\left(\|v\|^2 + 4(n-t+1) \right)^{1/2}} + \left(\|v\|^2 + \langle v, z_t \rangle + 4(n-t+1) \right)^{1/2} \right\} \tag{22.12}$$

Observe that z_t only enters through the inner product with v. We can therefore write $z_t = r\frac{v}{\|v\|}$ and the optimization problem as

$$\sup_{r\in[-1,1]}\left\{-\frac{r\|v\|}{2\left(\|v\|^2+4(n-t+1)\right)^{1/2}}+\left(\|v\|^2+r\|v\|+4(n-t+1)\right)^{1/2}\right\}$$

Setting the derivative to zero, we conclude that the supremum is attained at $r=0$. With this value, the required inequality (22.9) is proved.

The algorithm in (22.10) is, in fact, an optimal method for the above relaxation. Rather than being pulled out of a hat, it can be derived as an optimal solution with a few more lines of algebra. Observe that the algorithm is closely related to Follow the Regularized Leader and Dual Averaging, modulo the normalization step that always keeps the solution in side the set. The technique readily extends beyond unit Euclidean balls, in which case the absence of a projection is a big computational advantage over the usual first-order methods.

22.1.2   Exponential Weights

As another example of a relaxation that is easily derived as an upper bound on the sequential Rademacher complexity, consider the case when $\mathcal{F}$ is a finite set and $|\ell(f,z)|\le 1$. This example generalizes the case of bit prediction in Section 21.5, and the reader is referred to that section for the proof with a not-as-heavy notation.

Let $L_t(f)=\sum_{s=1}^{t}\ell(f,z_t)$ and $A(f)=2\sum_{i=1}^{n-t}\epsilon_i\ell(f,\mathbf{z}_i(\epsilon))$. For any $\lambda>0$ and any tree $\mathbf{z}$, the sequential Rademacher complexity

$$
\begin{aligned}
\mathbb{E}_\epsilon\max_{f\in\mathcal{F}}\left\{2\sum_{i=1}^{n-t}\epsilon_i\ell(f,\mathbf{z}_i(\epsilon))-\sum_{s=1}^{t}\ell(f,z_t)\right\} &\le \frac{1}{\lambda}\log\left(\mathbb{E}_\epsilon\max_{f\in\mathcal{F}}\exp\left(A(f)-\lambda L_t(f)\right)\right)\\
&\le\frac{1}{\lambda}\log\left(\mathbb{E}_\epsilon\sum_{f\in\mathcal{F}}\exp\left(A(f)-\lambda L_t(f)\right)\right)\\
&=\frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\lambda L_t(f)\right)\mathbb{E}_\epsilon\left\{\prod_{i=1}^{n-t}\exp\left(2\lambda\epsilon_i\ell(f,\mathbf{z}_i(\epsilon))\right)\right\}\right)
\end{aligned}
$$

We now upper bound the expectation over the "future" tree by the worst-case path, resulting in the upper bound

$$\frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\lambda L_t(f)\right)\times\exp\left(2\lambda^2\max_{\epsilon_1,\dots\epsilon_{n-t}\in\{\pm 1\}}\sum_{i=1}^{n-t}\ell(f,\mathbf{z}_i(\epsilon))^2\right)\right)\le\frac{1}{\lambda}\log\left(\sum_{f\in\mathcal{F}}\exp\left(-\lambda L_t(f)\right)\right)+2\lambda(n-t)$$

193

by Lemma 9.2. Notice that the last step removes the **z** tree. Since the above calculation holds for any $\lambda > 0$, we define the relaxation

$$\mathbf{Rel}_n(z_1, \ldots, z_t) = \inf_{\lambda > 0} \left\{ \frac{1}{\lambda} \log \left(\sum_{f \in \mathcal{F}} \exp \left(-\lambda \sum_{i=1}^{t} \ell(f, z_i) \right) \right) + 2\lambda(n - t) \right\} \qquad (22.13)$$

that gives to a computationally tractable algorithm. Let us first prove that the relaxation is admissible with the Exponential Weights algorithm as an admissible algorithm. Let λ^* be the optimal value in the definition of $\mathbf{Rel}_n(z_1, \ldots, z_{t-1})$. Then, by suboptimality of λ^* for $\mathbf{Rel}_n(z_1, \ldots, z_t)$

$$\inf_{q_t \in \Delta(\mathcal{D})} \sup_{z_t \in \mathcal{Z}} \left\{ \mathbb{E}_{f \sim q_t} \left[\ell(f, z_t) \right] + \mathbf{Rel}_n(z_1, \ldots, z_t) \right\}$$

$$\leq \inf_{q_t \in \Delta(\mathcal{D})} \sup_{z_t \in \mathcal{Z}} \left\{ \mathbb{E}_{f \sim q_t} \left[\ell(f, z_t) \right] + \frac{1}{\lambda^*} \log \left(\sum_{f \in \mathcal{F}} \exp \left(-\lambda^* L_t(f) \right) \right) + 2\lambda^*(n - t) \right\}$$

Let us upper bound the infimum by a particular choice of q which is the exponential weights distribution

$$q_t(f) = \exp(-\lambda^* L_{t-1}(f)) / Z_{t-1}$$

where $Z_{t-1} = \sum_{f \in \mathcal{F}} \exp \left(-\lambda^* L_{t-1}(f) \right)$. By [16, Lemma A.1],

$$\frac{1}{\lambda^*} \log \left(\sum_{f \in \mathcal{F}} \exp \left(-\lambda^* L_t(f) \right) \right) = \frac{1}{\lambda^*} \log \left(\mathbb{E}_{f \sim q_t} \exp \left(-\lambda^* \ell(f, z_t) \right) \right) + \frac{1}{\lambda^*} \log Z_{t-1}$$

$$\leq -\mathbb{E}_{f \sim q_t} \ell(f, z_t) + \frac{\lambda^*}{2} + \frac{1}{\lambda^*} \log Z_{t-1}$$

Hence,

$$\inf_{q_t \in \Delta(\mathcal{D})} \sup_{z_t \in \mathcal{Z}} \left\{ \mathbb{E}_{f \sim q_t} \left[\ell(f, z_t) \right] + \mathbf{Rel}_n(z_1, \ldots, z_t) \right\} \leq \frac{1}{\lambda^*} \log \left(\sum_{f \in \mathcal{F}} \exp \left(-\lambda^* L_{t-1}(f) \right) \right) + 2\lambda^*(n - t + 1)$$

$$= \mathbf{Rel}_n(z_1, \ldots, z_{t-1})$$

by the optimality of λ^*. The bound can be improved by a factor of 2 for some loss functions, since it will disappear from the definition of sequential Rademacher complexity.

The Chernoff-Cramèr inequality tells us that (22.13) is the tightest possible relaxation. The proof reveals that the only inequality is the softmax which is also

present in the proof of the maximal inequality for a finite collection of random variables. In this way, exponential weights is an algorithmic realization of a maximal inequality for a finite collection of random variables. The connection between probabilistic (or concentration) inequalities and algorithms runs much deeper.

We point out that the exponential-weights algorithm arising from the relaxation (22.13) is a *parameter-free* algorithm. The learning rate λ^* can be optimized (via one-dimensional line search) at each iteration with almost no cost. This can lead to improved performance as compared to the classical methods that set a particular schedule for the learning rate.

Our aim at this point was to show that the associated relaxations arise naturally (typically with a few steps of algebra) from the sequential Rademacher complexity. It should now be clear that upper bounds, such as the Dudley Entropy integral, can be turned into a relaxation, provided that admissibility is proved. Our ideas have semblance of those in Statistics, where an information-theoretic complexity can be used for defining penalization methods.

22.2   Supervised Learning

In the supervised setting of sequential prediction, the learner observes $x_t \in \mathcal{X}$, makes a prediction $\widehat{y}_t \in \mathcal{D}$ (or $q_t \in \Delta(\mathcal{D})$) and observes $y_t \in \mathcal{Y}$. We write the minimax value as

$$\left\langle\!\!\left\langle \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta(\mathcal{D})} \sup_{y_t \in \mathcal{Y}} \mathbb{E}_{\widehat{y}_t \sim q_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \left[\sum_{t=1}^{n} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right] \tag{22.14}$$

Let us briefly detail the relaxation framework for this supervised case. A relaxation will be called *admissible* if for any $(x_1, y_1), \ldots, (x_n, y_n) \in (\mathcal{X} \times \mathcal{Y})$,

$$\mathbf{Rel}_n\big((x_1, y_1), \ldots, (x_{t-1}, y_{t-1})\big) \geq \sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta(\mathcal{D})} \sup_{y_t \in \mathcal{Y}} \left\{ \mathbb{E}_{\widehat{y}_t \sim q_t} \left[\ell(\widehat{y}_t, y_t) \right] + \mathbf{Rel}_n\big((x_1, y_1), \ldots, (x_t, y_t)\big) \right\}$$

$$\tag{22.15}$$

for all $t \in [n]$, and

$$\mathbf{Rel}_n\big((x_1, y_1), \ldots, (x_n, y_n)\big) \geq - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t).$$

Since x_t is revealed at the beginning of round t, our strategy can be computed based on x_t:

$$q_t = \underset{q \in \Delta(\mathcal{D})}{\operatorname{argmin}} \sup_{y_t} \left\{ \underset{\widehat{\boldsymbol{y}}_t \sim q}{\mathbb{E}} \left[\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t) \right] + \mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_t, y_t) \big) \right\} \tag{22.16}$$

Whenever $\mathcal{Y}$ is bounded and both $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t)$ and $\mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_t, y_t) \big)$ are convex in y_t, the supremum in Eq. (22.16) becomes a maximum between the two extreme values for y_t, significantly simplifying both calculation of the strategy q_t and verifying admissibility. Another example where this simplification happens is in classification (binary or multi-class).

Binary Classification In the case of binary-valued labels $\mathcal{D} = \mathcal{Y} = \{0, 1\}$, the objective takes on a simpler form. Suppose that the loss is $\boldsymbol{\ell}(\widehat{\boldsymbol{y}}, y) = \mathbf{I}\{\widehat{\boldsymbol{y}} \neq y\}$. Then $\mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \mathbf{I}\{\widehat{\boldsymbol{y}}_t, y_t\} = |y_t - q_t|$ and the optimization problem becomes

$$q_t = \underset{q \in \Delta(\mathcal{D})}{\operatorname{argmin}} \max \big\{ 1 - q + \mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_t, 1) \big), q + \mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_t, 0) \big) \big\}$$

$$\tag{22.17}$$

This development is very similar to (21.9). The minimum is

$$q_t = \frac{1}{2} + \frac{\mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_t, 1) \big) - \mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_t, 0) \big)}{2} \tag{22.18}$$

if $q_t \in [0, 1]$, and otherwise needs to be clipped to this interval. For simplicity, assume no clipping is required. Plugging in the value of q_t into the right-hand side of (22.15), we obtain

$$\sup_{x_t \in \mathcal{X}} \inf_{q_t \in \Delta(\mathcal{D})} \sup_{y_t \in \mathcal{Y}} \left\{ \underset{\widehat{\boldsymbol{y}}_t \sim q_t}{\mathbb{E}} \left[\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t) \right] + \mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_t, y_t) \big) \right\}$$

$$\leq \sup_{x_t \in \mathcal{X}} \left\{ \frac{1}{2} + \frac{1}{2} \big(\mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_t, 1) \big) + \mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_t, 0) \big) \big) \right\}$$

$$= \sup_{x_t \in \mathcal{X}} \left\{ \frac{1}{2} + \mathbb{E}_{b_t} \mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_t, b_t) \big) \right\} \tag{22.19}$$

for a fair coin $b_t \in \{0, 1\}$. And so checking admissibility of the relaxation and strategy (22.16) reduces to verifying that

$$\frac{1}{2} + \sup_{x_t \in \mathcal{X}} \mathbb{E}_{b_t} \mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_t, b_t) \big) \leq \mathbf{Rel}_n \big((x_1, y_1), \ldots, (x_{t-1}, y_{t-1}) \big) \tag{22.20}$$

For the case of $\mathcal{Y} = \{-1, +1\}$, the Bernoulli variable b_t is replaced by the Rademacher random variable ϵ_t throughout.

23

Algorithms Based on Random Playout, and Follow the Perturbed Leader

23.1 The Magic of Randomization

Recall the generic form (22.5) of a Meta Algorithm for a given relaxation. Suppose the relaxation is of the form $\textbf{Rel}_n\,(z_1,\ldots,z_t) = \mathbb{E}\Phi(\textsc{w}, z_{1:t})$, for some function Φ and the expectation is over some random variable $\textsc{w}$. The optimal randomized strategy for this relaxation is

$$q^* = \operatorname*{argmin}_{q}\,\sup_{z_t}\left\{\mathbb{E}_{\hat{\boldsymbol{y}}\sim q}\boldsymbol{\ell}(\hat{\boldsymbol{y}}, z_t) + \mathbb{E}_{\textsc{w}\sim p}\Phi(\textsc{w}, z_{1:t})\right\}. \tag{23.1}$$

However, computing the expectation might be computationally costly. With this in mind, consider a deceptively simple-minded strategy $\tilde{q}$: first draw $\textsc{w} \sim p$ and then compute

$$q(\textsc{w}) \triangleq \operatorname*{argmin}_{q}\,\sup_{z_t}\left\{\mathbb{E}_{\hat{\boldsymbol{y}}\sim q}\boldsymbol{\ell}(\hat{\boldsymbol{y}}, z_t) + \Phi(\textsc{w}, z_{1:t})\right\}. \tag{23.2}$$

or its clipped version if $q(\textsc{w})$ falls outside $[0, 1]$.

We then verify that the value of the objective in (23.1) evaluated at $\tilde{q}$ is

$$\begin{aligned}
\sup_{z_t}\left\{\mathbb{E}_{\hat{\boldsymbol{y}}\sim\tilde{q}}\boldsymbol{\ell}(\hat{\boldsymbol{y}}, z_t) + \mathbb{E}_{\textsc{w}\sim p}\Phi(\textsc{w}, z_{1:t})\right\} &= \sup_{z_t}\left\{\mathbb{E}_{\textsc{w}\sim p}\mathbb{E}_{\hat{\boldsymbol{y}}\sim q(\textsc{w})}\boldsymbol{\ell}(\hat{\boldsymbol{y}}, z_t) + \mathbb{E}_{\textsc{w}\sim p}\Phi(\textsc{w}, z_{1:t})\right\} \\
&\leq \mathbb{E}_{\textsc{w}\sim p}\sup_{z_t}\left\{\mathbb{E}_{\hat{\boldsymbol{y}}\sim q(\textsc{w})}\boldsymbol{\ell}(\hat{\boldsymbol{y}}, z_t) + \Phi(\textsc{w}, z_{1:t})\right\} \\
&= \mathbb{E}_{\textsc{w}\sim p}\inf_{q}\sup_{z_t}\left\{\mathbb{E}_{\hat{\boldsymbol{y}}\sim q}\boldsymbol{\ell}(\hat{\boldsymbol{y}}, z_t) + \Phi(\textsc{w}, z_{1:t})\right\}\,.
\end{aligned}$$

$$\tag{23.3}$$

Recall that to show admissibility of the relaxation and the randomized strategy $\tilde{q}$, we need to ensure that the above expression is upper bounded by the relaxation with one less outcome. Thankfully, in view of the last expression, we may do so *conditionally* on w, as if there were no random variables in the definition of the relaxation at all! This general approach of obtaining randomized methods for relaxations defined via an expectation appears to be quite powerful. The randomized strategy can be seen as "mimicking" the randomization in the relaxation. Many of the methods developed later in the course are based on this simple technique, which we will term "random playout". In game theory, random playout is a strategy employed for estimating the value of a board position. Luckily, the strategy has a solid basis in regret minimization. In particular, we will show that a well-known Follow the Perturbed Leader algorithm is an example of such a randomized strategy.

Our starting point is, once again, the conditional sequential Rademacher complexity

$$\mathcal{R}_n(z_1,\ldots,z_t) = \sup_{\mathbf{z}} \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left[2 \sum_{s=t+1}^{n} \epsilon_s \boldsymbol{\ell}(f, \mathbf{z}_{s-t}(\epsilon_{t+1:s-1})) - \sum_{s=1}^{t} \boldsymbol{\ell}(f, z_s) \right] \quad (23.4)$$

However, we cannot employ the above arguments right away, as the relaxation involves a supremum over a tree. In what follows, we provide several interesting situations where the supremum can be replaced by an expectation.

23.2 Linear Loss

We now show a somewhat surprising result: it is often the case in finite-dimensional problems with linear loss that the supremum over the trees can be replaced by an *i.i.d. distribution* that is "almost as bad". We first phrase this fact as an assumption, and then verify it for several cases.

Following the abstraction of Section 20.3, suppose $\mathcal{Z}$ is a unit ball in a separable Banach space $(\mathcal{B}, \|\cdot\|)$ and $\mathcal{F}$ is a unit ball in the dual norm $\|\cdot\|_*$.

Assumption 23.1. *There exists a distribution $\mathscr{D} \in \Delta(\mathcal{Z})$ and constant $C \geq 2$ such that for any $w \in \mathcal{B}$*

$$\sup_{z \in \mathcal{Z}} \mathbb{E} \|w + 2\epsilon_1 z\| \leq \mathbb{E}_{z \sim \mathscr{D}} \mathbb{E} \|w + C\epsilon_1 z\| \quad (23.5)$$

where ϵ_1 is a Rademacher random variable.

The assumption says that there exists a universal (that is, independent of w) distribution $\mathcal{D}$ for the problem, such that the expected length of a one-step random walk from any point w under this distribution is almost as large as the expected length that would be achieved with the worst two-point symmetric distribution that can depend on w.

✎ **Exercise 23.1** (⋆)**.** Prove that under Assumption 23.1, the sequential Rademacher complexity of $\mathcal{F}$ is within a constant factor from the classical i.i.d. Rademacher complexity with respect to $\mathcal{D}^n$.

The assumption holds in finite dimensional spaces. Not surprisingly, the algorithm based on this assumption has a regret guarantee upper bounded by the i.i.d. Rademacher complexity with respect to $\mathcal{D}^n$. Let us now state a general lemma; its proof is deferred to Section 23.2.3.

Lemma 23.2. *Under the Assumption 23.1, the relaxation*

$$\mathbf{Rel}_n\left(z_1,\ldots,z_t\right) = \underset{z_{t+1},\ldots,z_n \sim \mathcal{D}}{\mathbb{E}}\,\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left[C \sum_{i=t+1}^{n} \epsilon_i \left\langle f, z_i \right\rangle - \sum_{i=1}^{t} \left\langle f, z_i \right\rangle \right] \qquad (23.6)$$

is admissible and a randomized strategy that ensures admissibility is given by: at time t, draw $z_{t+1},\ldots,z_n \sim \mathcal{D}$ and Rademacher random variables $\epsilon = (\epsilon_{t+1},\ldots,\epsilon_n)$, and then define

$$\widehat{\boldsymbol{y}}_t = \underset{\widehat{\boldsymbol{y}} \in \mathcal{D}}{\operatorname{argmin}} \sup_{z \in \mathcal{Z}} \left\{ \left\langle \widehat{\boldsymbol{y}}, z \right\rangle + \left\| C \sum_{i=t+1}^{n} \epsilon_i z_i - \sum_{i=1}^{t-1} z_i - z \right\| \right\} \qquad (23.7)$$

The expected regret for the method is bounded by the classical Rademacher complexity:

$$\mathbb{E}\boldsymbol{Reg}_n \leq C\,\mathbb{E}_{z_{1:n} \sim \mathcal{D}}\,\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \left\langle f, z_t \right\rangle \right],$$

It will be shown in the examples below that the update (23.7) is of the nicer form

$$\widehat{\boldsymbol{y}}_t = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\langle f, \sum_{i=1}^{t-1} z_i - C \sum_{i=t+1}^{n} \epsilon_i z_i \right\rangle \qquad (23.8)$$

This is the famous Follow the Perturbed Leader algorithm [30, 16]:

> **Follow the Perturbed Leader**
>
> $$\widehat{\boldsymbol{y}}_t = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{t-1} \langle f, z_i \rangle + \eta^{-1} \langle f, r \rangle$$

where r is a random vector and η is some learning rate parameter. It is now clear that the random vector r arises as a sum of random draws from the distribution that is "almost as bad" as the worst-case tree in the conditional sequential Rademacher complexity.

We now look at specific examples of ℓ_2/ℓ_2 and ℓ_1/ℓ_∞ cases and provide closed form solution of the randomized algorithms.

23.2.1 Example: Follow the Perturbed Leader on the Simplex

Here, we consider the setting similar to that in [30]. Let $\mathcal{F} = \mathsf{B}_1^N$ and $\mathcal{Z} = \mathsf{B}_\infty^N$. In [30], the set $\mathcal{F}$ is the probability simplex and $\mathcal{Z} = [0,1]^N$ but these are subsumed by the ℓ_1/ℓ_∞ case. We claim that:

Lemma 23.3. *Assumption 23.1 is satisfied with a distribution $\mathcal{D}$ that is uniform on the vertices of the cube $\{\pm 1\}^N$ and $C = 6$.*

Proof of Lemma 23.3. Let $w \in \mathbb{R}^N$ be arbitrary. Throughout this proof, let $\epsilon \in \{\pm 1\}$ be a single Rademacher random variable. The norm in (23.5) is the ℓ_∞ norm, and so we need to show

$$\max_{z \in \{\pm 1\}^N} \mathbb{E}_\epsilon \max_{i \in [N]} |w_i + 2\epsilon z_i| \le \underset{z \sim \mathcal{D}}{\mathbb{E}} \, \mathbb{E}_\epsilon \max_{i \in [N]} |w_i + 6\epsilon z_i| \tag{23.9}$$

Let $i^* = \underset{i}{\operatorname{argmax}} \, |w_i|$ and $j^* = \underset{i \neq i^*}{\operatorname{argmax}} \, |w_i|$ be the coordinates with largest and second-largest magnitude. If $|w_{i^*}| - |w_{j^*}| \ge 4$, the statement follows since, for any $z \in \{\pm 1\}^N$ and $\epsilon \in \{\pm 1\}$,

$$\max_{i \neq i^*} |w_i + 2\epsilon z_i| \le \max_{i \neq i^*} |w_i| + 2 \le |w_{i^*}| - 2 \le |w_{i^*} + 2\epsilon z_{i^*}|,$$

and thus

$$\max_{z \in \{\pm 1\}^N} \mathbb{E}_\epsilon \max_{i \in [N]} |w_i + 2\epsilon z_i| = \max_{z \in \{\pm 1\}^N} \mathbb{E}_\epsilon |w_{i^*} + 2\epsilon z_{i^*}| = |w_{i^*}| \le \mathbb{E}_{z,\epsilon} |w_{i^*} + 6\epsilon z_{i^*}| \le \mathbb{E}_{z,\epsilon} \max_i |w_i + 6\epsilon z_i|.$$

It remains to consider the case when $|w_{i^*}| - |w_{j^*}| < 4$. We have that

$$\mathbb{E}_{z,\epsilon} \max_{i \in [N]} |w_i + 6\epsilon z_i| \ge \mathbb{E}_{z,\epsilon} \max_{i \in \{i^*, j^*\}} |w_i + 6\epsilon z_i| \ge \frac{1}{2}(|w_{i^*}| + 6) + \frac{1}{4}(|w_{i^*}| - 6) + \frac{1}{4}(|w_{j^*}| + 6) \ge |w_{i^*}| + 2$$

$$\ge \max_{z \in \{\pm 1\}^N} \mathbb{E}_{\epsilon} \max_{i \in [N]} |w_i + 2\epsilon z_i|,$$

where $1/2$ is the probability that $\epsilon z_{i^*} = \operatorname{sign}(w_{i^*})$, the second event of probability $1/4$ is the event that $\epsilon z_{i^*} \ne \operatorname{sign}(w_{i^*})$ and $\epsilon z_{j^*} \ne \operatorname{sign}(w_{j^*})$, while the third event of probability $1/4$ is that $\epsilon z_{i^*} \ne \operatorname{sign}(w_{i^*})$ and $\epsilon z_{j^*} = \operatorname{sign}(w_{j^*})$. $\qquad \square$

In fact, one can pick *any symmetric distribution $\mathscr{D}$ on the real line and use $\mathscr{D}^N$* for the perturbation. Assumption 23.1 is then satisfied, as we show in the following lemma.

Lemma 23.4. *If $\tilde{\mathscr{D}}$ is any symmetric distribution over the real line, then Assumption 23.1 is satisfied by using the product distribution $\mathscr{D} = \tilde{\mathscr{D}}^N$. The constant C required is any $C \ge 6/\mathbb{E}_{z \sim \tilde{\mathscr{D}}} |z|$.*

The above lemma is especially attractive when used with standard normal distribution because in that case as sum of normal random variables is again normal. Hence, instead of drawing $z_{t+1}, \ldots, z_n \sim N(0, 1)$ on round t, one can simply draw just one vector $z_t \sim N(0, \sqrt{n-t})$ and use it for perturbation. In this case constant C is bounded by 8.

While we have provided simple distributions to use for perturbation, the form of update in Equation (23.7) is not in a convenient form. The following lemma shows a simple Follow the Perturbed Leader type algorithm with the associated regret bound.

Lemma 23.5. *Suppose $\mathcal{D} = \mathcal{F} = B_1^N$, $\mathcal{Z} = B_\infty^N$, and let $\tilde{\mathscr{D}}$ be any symmetric distribution. Consider the randomized algorithm that at each round t freshly draws Rademacher random variables $\epsilon_{t+1}, \ldots, \epsilon_n$ and freshly draws $z_{t+1}, \ldots, z_n \sim \tilde{\mathscr{D}}^N$ (each co-ordinate drawn independently from $\tilde{\mathscr{D}}$) and picks*

$$\widehat{y}_t = \operatorname*{argmin}_{\widehat{y} \in \mathcal{D}} \left\langle \widehat{y}, \sum_{i=1}^{t-1} z_i - C \sum_{i=t+1}^{n} \epsilon_i z_i \right\rangle \tag{23.10}$$

where $C = 6/\mathbb{E}_{z \sim \tilde{\mathscr{D}}} [|z|]$. The randomized algorithm enjoys a bound on the expected regret given by

$$\mathbb{E}\boldsymbol{Reg}_n \le C \mathop{\mathbb{E}}_{z_{1:n} \sim \tilde{\mathscr{D}}^N} \mathbb{E}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t z_t \right\|_\infty + 4 \sum_{t=1}^{n} \mathbf{P}_{y_{t+1:n} \sim \tilde{\mathscr{D}}} \left(C \left| \sum_{i=t+1}^{n} y_i \right| \le 4 \right)$$

Notice that for $\tilde{\mathscr{D}}$ being the $\{\pm 1\}$ coin flips or standard normal distribution, the probability

$$\mathbf{P}_{y_{t+1},\dots,y_n \sim \tilde{\mathscr{D}}}\left(C \left| \sum_{i=t+1}^{n} y_i \right| \leq 4 \right)$$

is exponentially small in $n-t$ and so $\sum_{t=1}^{n} \mathbf{P}_{y_{t+1},\dots,y_n \sim \tilde{\mathscr{D}}}\left(C \left| \sum_{i=t+1}^{n} y_i \right| \leq 4 \right)$ is bounded by a constant. For these cases, we have

$$\mathbb{E}\mathbf{Reg}_n \leq O\left(\mathbb{E}_{z_{1:n} \sim \tilde{\mathscr{D}}^N} \mathbb{E}_\epsilon \left\| \sum_{t=1}^{n} \epsilon_t z_t \right\|_\infty \right) = O\left(\sqrt{n \log N} \right)$$

This yields the logarithmic dependence on the dimension, matching that of the Exponential Weights algorithm.

23.2.2   Example: Follow the Perturbed Leader on Euclidean Balls

We now consider the case when $\mathcal{F}$ and $\mathcal{Z}$ are both the unit ℓ_2 ball. We can use as perturbation the uniform distribution on the surface of unit sphere, as the following lemma shows.

Lemma 23.6. *Let $\mathcal{F} = \mathcal{Z} = \mathsf{B}_2^N$. Then Assumption 23.1 is satisfied with a uniform distribution $\mathscr{D}$ on the surface of the unit sphere with constant $C = 4\sqrt{2}$.*

Again as in the previous example the form of update in Equation (23.7) is not in a convenient form and this is addressed in the following lemma.

Lemma 23.7. *Let $\mathcal{D} = \mathcal{F} = \mathcal{Z} = \mathsf{B}_2^N$ and $\mathscr{D}$ be the uniform distribution on the surface of the unit sphere. Consider the randomized algorithm that at each round (say round t) freshly draws $z_{t+1}, \dots, z_n \sim \mathscr{D}$ and picks*

$$\widehat{y}_t = \frac{-\sum_{i=1}^{t-1} z_i + C\sum_{i=t+1}^{n} z_i}{\sqrt{\left\| -\sum_{i=1}^{t-1} z_i + C\sum_{i=t+1}^{n} \epsilon_i z_i \right\|_2^2 + 1}}$$

where $C = 4\sqrt{2}$. The randomized algorithm enjoys a bound on the expected regret given by

$$\mathbb{E}\mathbf{Reg}_n \leq C\,\mathbb{E}_{z_1,\dots,z_n \sim \mathscr{D}} \left\| \sum_{t=1}^{n} z_t \right\|_2 \leq 4\sqrt{2n}$$

Importantly, the bound does not depend on the dimensionality of the space.

23.2.3 Proof of Lemma 23.2

To show admissibility using the particular randomized strategy q_t described in Lemma 23.5, we need to show that

$$\sup_{z_t} \left\{ \mathbb{E}_{\widehat{\boldsymbol{y}} \sim q_t} \left[\langle \widehat{\boldsymbol{y}}, z_t \rangle \right] + \mathbf{Rel}_n (z_1, \ldots, z_t) \right\} \leq \mathbf{Rel}_n (z_1, \ldots, z_{t-1})$$

The strategy q_t proposed by the lemma is such that we first draw $z_{t+1}, \ldots, z_n \sim \mathscr{D}$ and $\epsilon_{t+1}, \ldots \epsilon_n$ Rademacher random variables, and then calculate $\widehat{\boldsymbol{y}}_t = \widehat{\boldsymbol{y}}_t(z_{t+1:n}, \epsilon_{t+1:n})$ as in (23.7). In view of (23.3),

$$\sup_{z_t} \left\{ \mathbb{E}_{\widehat{\boldsymbol{y}} \sim q_t} \left[\langle \widehat{\boldsymbol{y}}, z_t \rangle \right] + \mathbf{Rel}_n (z_1, \ldots, z_t) \right\} \leq \underset{\substack{\epsilon_{t+1:n} \\ z_{t+1:n}}}{\mathbb{E}} \inf_{g \in \mathcal{F}} \sup_{z_t} \left\{ \langle g, z_t \rangle + \left\| C \sum_{i=t+1}^{n} \epsilon_i z_i - \sum_{i=1}^{t} z_i \right\| \right\}$$

$$(23.11)$$

Let $w = C \sum_{i=t+1}^{n} \epsilon_i z_i - \sum_{i=1}^{t-1} z_i$. We now appeal to the minimax theorem:

$$\inf_{g \in \mathcal{F}} \sup_{z_t} \left\{ \langle g, z_t \rangle + \| w - z_t \| \right\} = \inf_{g \in \mathcal{F}} \sup_{p_t \in \Delta(\mathbb{Z})} \mathbb{E}_{z_t \sim p_t} \left\{ \langle g, z_t \rangle + \| w - z_t \| \right\}$$

$$= \sup_{p \in \Delta(\mathbb{Z})} \inf_{g \in \mathcal{F}} \left\{ \mathbb{E}_{z_t \sim p} \langle g, z_t \rangle + \mathbb{E}_{z_t \sim p} \| w - z_t \| \right\}$$

The minimax theorem holds because loss is convex in g and $\mathcal{F}$ is a compact convex set and the term in the expectation is linear in p_t, as it is an expectation. The last expression is equal to

$$\sup_{p \in \Delta(\mathbb{Z})} \mathbb{E}_{z_t \sim p} \sup_{f \in \mathcal{F}} \left[\langle f, w \rangle + \inf_{g \in \mathcal{F}} \mathbb{E}_{z_t \sim p} \left[\langle g, z_t \rangle \right] - \langle f, z_t \rangle \right] \leq \sup_{p \in \Delta(\mathbb{Z})} \mathbb{E}_{z_t \sim p} \sup_{f \in \mathcal{F}} \left[\langle f, w \rangle + \mathbb{E}_{z_t \sim p} \left[\langle f, z_t \rangle \right] - \langle f, z_t \rangle \right]$$

$$\leq \sup_{z_t \in \mathbb{Z}} \mathbb{E}_{\epsilon_t} \sup_{f \in \mathcal{F}} \left[\langle f, w \rangle + 2\epsilon \langle f, z_t \rangle \right]$$

$$\leq \sup_{z_t \in \mathbb{Z}} \mathbb{E}_{\epsilon_t} \| w + 2\epsilon z_t \|$$

using the familiar symmetrization technique. In view of Assumption 23.1,

$$\sup_{z_t \in \mathbb{Z}} \mathbb{E}_{\epsilon_t} \| w + 2\epsilon z_t \| \leq \underset{z_t \sim \mathscr{D}}{\mathbb{E}} \underset{\epsilon_t}{\mathbb{E}} \| w + C \epsilon_t z_t \| = \underset{z_t \sim \mathscr{D}}{\mathbb{E}} \underset{\epsilon_t}{\mathbb{E}} \left\| C \sum_{i=t}^{n} \epsilon_i z_i - \sum_{i=1}^{t-1} z_i \right\|$$

Plugging this into Eq. (23.11) completes the proof of admissibility.

23.3   Supervised Learning

Recall the relaxation framework in Section 22.2. Suppose that the relaxation in Eq. (22.16) is of the form $\mathbb{E}\Phi(w, x_{1:t}, y_{1:t})$ for some random variable w with distribution p. For simplicity, consider the classification scenario with $\mathcal{Y} = \{0, 1\}$. Consider the following randomized strategy $\tilde{q}_t$: draw $w \sim p$, and set

$$q(w) = \frac{1}{2} + \frac{\Phi(w, (x_1, y_1), \ldots, (x_t, 1)) - \Phi(w, (x_1, y_1), \ldots, (x_t, 0))}{2}, \tag{23.12}$$

or its clipped version if $q(w)$ falls outside $[0, 1]$. The randomized strategy $\tilde{q}_t$ combines the idea of randomization described in the beginning of this chapter together with the closed-form solution of (22.16).

Admissibility of the above strategy (under the appropriate conditions on Φ) will be shown along the following general lines. An adaptation of the argument leading to (23.3) for supervised learning gives that for any x_t, the value of the objective for the randomized strategy $\tilde{q}$ is

$$\sup_{y_t} \left\{ \mathbb{E}_{\widehat{y} \sim \tilde{q}} \ell(\widehat{y}, y_t) + \mathbb{E}_{w \sim p} \Phi(w, x_{1:t}, y_{1:t}) \right\} \leq \mathbb{E}_{w \sim p} \inf_q \sup_{y_t} \left\{ \mathbb{E}_{\widehat{y} \sim q} \ell(\widehat{y}, y_t) + \Phi(w, x_{1:t}, y_{1:t}) \right\} .$$

$$\tag{23.13}$$

In the case of binary classification with indicator loss, we turn to the analysis in Eq. (22.19). We then see that the value of the objective for $\tilde{q}$ is at most

$$\sup_{x_t} \mathbb{E}_{w \sim p} \inf_q \sup_{y_t} \left\{ \mathbb{E}_{\widehat{y} \sim q} \mathbf{I}\left\{ \widehat{y} \neq y_t \right\} + \Phi(w, x_{1:t}, y_{1:t}) \right\} \leq \sup_{x_t} \mathbb{E}_{w \sim p} \left\{ \frac{1}{2} + \mathbb{E}_{b_t} \Phi(w, x_{1:t}, y_{1:t-1}, b_t) \right\}$$

$$\tag{23.14}$$

and thus verifying admissibility reduces to checking

$$\frac{1}{2} + \sup_{x_t} \mathbb{E}_{b_t} \mathbb{E}_{w \sim p} \Phi(w, x_{1:t}, y_{1:t-1}, b_t) \leq \mathbb{E}\Phi(w', x_{1:t-1}, y_{1:t-1}) \tag{23.15}$$

for some w'. This step can be verified under a condition akin Assumption 23.1. Alternatively, this inequality holds in the so-called transductive (or, fixed design) setting. This is the subject of the next chapter.

24

Algorithms for Fixed Design

24.1 ... And the Tree Disappears

In this chapter, we will study the supervised learning problem under the assumption that $\mathcal{X}$ is a finite set and the side information x_t at time t is chosen from this set without replacement. Once this side information is revealed, the supervised protocol is as before: the learner makes a prediction $\hat{\boldsymbol{y}} \in \mathcal{D} \subset \mathbb{R}$ (or chooses a randomized strategy q_t) and the outcome $y_t \in \mathbb{R}$ is subsequently revealed. We will refer to this setting as "fixed design".

As we will see shortly, in the protocol described above, the tree $\mathbf{x}$ —an obstacle in obtaining computationally efficient algorithms—disappears.

In the beginning of the course, we wrote down minimax regret for arbitrary sequences $(x_1, y_1), \ldots, (x_n, y_n)$. How do we ensure that x's are not repeated in this sequence? For this, we need to introduce a notion of *restrictions* on sequences.

Abstractly, at each time step t, we may define a set of allowed choices for x_t by $C_t \subseteq \mathcal{X}$. This restriction can be viewed as a set-valued function $(x_1, \ldots, x_{t-1}) \mapsto C_t(x_1, \ldots, x_{t-1})$. The notation makes explicit the fact that the set may depend on the history (and in fact can also depend on the y sequence, yet the analysis becomes more involved [46]). The case of observing non-repeating x's corresponds to choosing

$$C_t(x_1, \ldots, x_{t-1}) = \mathcal{X} \setminus \{x_1, \ldots, x_{t-1}\}. \tag{24.1}$$

205

We may now write down the minimax value as

$$\left\langle\!\!\left\langle \sup_{x_t \in C(x_{1:t-1})} \inf_{q_t} \sup_{y_t} \mathbb{E}_{\widehat{\boldsymbol{y}}_t} \right\rangle\!\!\right\rangle_{t=1}^n \left\{ \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, y_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(f(x_t), y_t) \right\} \qquad (24.2)$$

Then, a generalization of Proposition 14.9 holds:

Proposition 24.1. *For a supervised problem with constraints and with L-Lipschitz and convex loss $\boldsymbol{\ell}(\cdot, y)$ for any $y \in \mathcal{Y}$, the minimax regret in (24.2) is upper bounded by*

$$2L \sup_{\mathbf{x} \in \mathcal{C}} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \epsilon_t f(\mathbf{x}_t(\epsilon)) \qquad (24.3)$$

where $\mathcal{C}$ is a set of $\mathcal{X}$-valued trees $\mathbf{x}$ that respect the constraints on every path:

$$\forall 1 \le t \le n, \quad \mathbf{x}_t(\epsilon) \in C_t(\mathbf{x}_1, \mathbf{x}_2(\epsilon), \dots, \mathbf{x}_{t-1}(\epsilon))$$

for any $\epsilon \in \{\pm 1\}^n$.

We now show that in the case of constraints (24.1), the supremum can be equivalently taken over constant-level trees:

Lemma 24.2. *For $n = |\mathcal{X}|$ and constraints (24.1), the supremum over $\mathbf{x} \in \mathcal{C}$ in Proposition 24.1 is achieved at a constant-level tree. Hence, the upper bound of Proposition (24.1) is equal to*

$$2L \widehat{\mathcal{R}}^{iid}(\mathcal{F}; x_1, \dots, x_n).$$

In view of Lemma 24.2, the regret behavior in fixed design setting of online supervised learning is governed by i.i.d., rather than sequential, complexities. This important observation translates into efficient algorithms, as the tree can be removed from conditional sequential Rademacher relaxation:

Lemma 24.3. *Consider fixed design setting of online supervised learning with convex L-Lipschitz loss $\boldsymbol{\ell}(\cdot, y)$ for all $y \in \mathcal{Y} \subset \mathbb{R}$. Then*

$$\mathcal{R}_n((x_1, y_1), \dots, (x_t, y_t)) = \mathbb{E}_{\epsilon_{t+1:n}} \sup_{f \in \mathcal{F}} \left[2L \sum_{x_s \in \mathcal{X} \setminus \{x_1, \dots, x_t\}} \epsilon_s f(x_s) - \sum_{s=1}^t \boldsymbol{\ell}(f(x_s), y_s) \right] \qquad (24.4)$$

is an admissible relaxation.

In addition to the Lipschitz-loss setting, we can consider the indicator loss $\ell(\widehat{y}, y) = \mathbf{I}\{\widehat{y} \neq y\}$. If $\mathcal{D} = \mathcal{Y} = \{\pm 1\}$, we may substitute $\mathbf{I}\{\widehat{y} \neq y\} = \frac{1}{2}(1 - \widehat{y}y)$ which is 1/2-Lipschitz. The relaxation becomes

$$\mathcal{R}_n((x_1, y_1), \ldots, (x_t, y_t)) = \mathbb{E} \sup_{f \in \mathcal{F}} \left[\sum_{x_s \in \mathcal{X} \setminus \{x_1, \ldots, x_t\}} \epsilon_s f(x_s) - \sum_{s=1}^{t} \frac{1}{2}(1 - y_s f(x_s)) \right] \quad (24.5)$$

It is easy to verify that this relaxation is indeed admissible. In view of Eq.(23.15) adapted to the case of $\mathcal{Y} = \{+1, -1\}$, it is enough to check that

$$\frac{1}{2} + \sup_{x_t \in C(x_{1:t-1})} \mathbb{E}_{\epsilon_t} \mathbb{E} \sup_{f \in \mathcal{F}} \left[\sum_{x_s \in \mathcal{X} \setminus \{x_1, \ldots, x_t\}} \epsilon_s f(x_s) - \sum_{s=1}^{t-1} \frac{1}{2}(1 - y_s f(x_s)) - \frac{1}{2}(1 - \epsilon_t f(x_t)) \right]$$

$$\leq \mathbb{E} \sup_{f \in \mathcal{F}} \left[\sum_{x_s \in \mathcal{X} \setminus \{x_1, \ldots, x_{t-1}\}} \epsilon_s f(x_s) - \sum_{s=1}^{t-1} \frac{1}{2}(1 - y_s f(x_s)) \right] \quad (24.6)$$

✎ **Exercise 24.1** (⋆)**.** Prove inequality (24.6).

Several remarks are in order. First, observe that the relaxation is of the form studied in chapter 23.1, and thus randomized methods are directly applicable. Second, if $\mathcal{F}$ is a class of linear functions on $\mathcal{X}$, the mixed strategy for predicting $\widehat{y}_t$ can be obtained by standard linear optimization methods.

24.2 Static Experts

Cesa-Bianchi and Lugosi [15] observed that minimax regret for the problem of sequential prediction with Static Experts is given by the classical Rademacher averages of the class of experts. This, in turn, motivated much of the work presented in this course.

Let $\mathcal{F} \subset [0,1]^n$. Each $f \in \mathcal{F}$ is viewed as an expert that conveys a prediction $f_t \in [0,1]$. This prediction is, therefore, a function of time only. Taking $\mathcal{Y} = \{0,1\}$, we may view each prediction f_t as a mixed strategy for predicting the next outcome. The expected indicator loss $\mathbb{E}_{y \sim f_t} \mathbf{I}\{y \neq y_t\}$ can be written simply as $\ell(f_t, y_t) = |f_t - y_t|$. Taking the decision set $\mathcal{D} = [0,1]$, we may write regret as

$$\frac{1}{n} \sum_{t=1}^{n} |\widehat{y}_t - y_t| - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} |f_t - y_t|.$$

The easiest way to map the problem into our general framework is to view the time index $x_t = t$ as the side information from $\mathcal{X} = \{1, \ldots, n\}$, and $f(x_t) = f_t$. Now, the constraints on the presentation of x_t are given trivially by

$$C_t(x_1, \ldots, x_{t-1}) = \{t\}$$

which is even more restrictive than (24.1). By Lemma 24.2, minimax regret is upper bounded by classical Rademacher averages of $\mathcal{F}$. Furthermore, one may obtain efficient algorithms using the relaxation in Lemma 24.3.

24.3 Social Learning / Network Prediction

Consider a weighted graph $G = (V, E, W)$, where V is the set of vertices, E the set of edges, and $W : E \to [-1, 1]$ the weights on the edges. We may think of the graph as a social network with positive and negative relationships between agents. We are interested in the problem of online node classification on this graph.
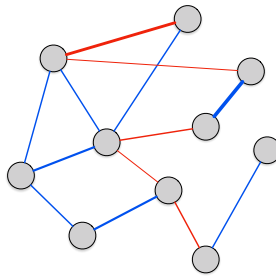

Figure 24.1: Weighted graph $G = (V, E, W)$. Edge color and thickness corresponds to the sign and magnitude of the weight.

More precisely, at each time step, the identity of the vertex $v_t \in V$ is revealed. We view this identity as the side information provided to the learner. Following the assumption of the previous section, the nodes are not repeated: once prediction has been made and a label has been observed, we do not come back to the same node.

24.4 Matrix Completion / Netflix Problem

Adaptive Algorithms

25.1   Adaptive Relaxations

So far in the course, the upper bounds on regret have been uniform for all sequences. There is a sense, however, that certain sequences can be "easier" and some – "harder". In this sense, we might try to prove upper bounds of the

$$\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(f, z_t) \le \psi(z_1,\ldots,z_n) \tag{25.1}$$

for some function ψ that might not be constant. Within various contexts, such upper bounds have appeared in the literature for ψ being some notion of a "variance" of the sequence, a "length" of the sequence, and so on.

The reader should correctly observe the semblance of (25.1) to the notion studied in the very first lecture (see Eq. (2.2)). We will come back to this example in just a bit, but let us first state the straightforward extension of the relaxation framework to the regret against a benchmark $\phi(z_1,\ldots,z_n)$:

$$\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(\widehat{\boldsymbol{y}}_t, z_t) \le \phi(z_1,\ldots,z_n) \tag{25.2}$$

where (25.1) is a particular choice of the benchmark $\phi(z_1,\ldots,z_n) = \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(f, z_t) + \psi(z_1,\ldots,z_n)$.

We say $\mathbf{Rel}_n\,()$ is an admissible relaxation if for any $z_1,\ldots,z_n$, the initial condition

$$\mathbf{Rel}_n\,(z_1,\ldots,z_n) \ge -n\phi(z_1,\ldots,z_n) \tag{25.3}$$

holds, along with the recursive condition

$$\mathbf{Rel}_n(z_1,\dots,z_{t-1}) \geq \inf_{q_t} \sup_{z_t} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \left\{ \ell(\widehat{\boldsymbol{y}}_t, z_t) + \mathbf{Rel}_n(z_1,\dots,z_t) \right\} \tag{25.4}$$

(The case of supervised learning is treated as in (22.15).) It is easy to see that a strategy guaranteeing (25.4) ensures

$$\forall z_1,\dots,z_n \qquad \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \ell(\widehat{\boldsymbol{y}}_t, z_t) - \phi(z_1,\dots,z_n) \leq \mathbf{Rel}_n(\varnothing) \tag{25.5}$$

25.2   Example: Bit Prediction from Lecture 1

We are now ready to give a simple proof of Proposition 2.1 for the case of bit prediction without side information. The claim is that for a stable ϕ (in the sense of Eq. 2.3), there exists a prediction algorithm guaranteeing

$$\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \ell(\widehat{\boldsymbol{y}}_t, y_t) \leq \phi(y_1,\dots,y_n) \tag{25.6}$$

for any sequence of $y_1,\dots,y_n \in \{0,1\}$ if and only if $\mathbb{E}\phi \geq \frac{1}{2}$ under the uniform distribution. The necessity of the latter condition was already shown in Chapter 2. To see the sufficiency, define a relaxation

$$\mathbf{Rel}_n(y_1,\dots,y_t) = -n\mathbb{E}\left[\phi(y_1,\dots,y_n)|y_1,\dots,y_t\right] + n\mathbb{E}\phi - \frac{t}{2} \tag{25.7}$$

We then check that the initial condition

$$\mathbf{Rel}_n(y_1,\dots,y_n) = -n\phi(y_1,\dots,y_n) + n\mathbb{E}\phi - \frac{n}{2} \geq -n\phi(y_1,\dots,y_n) \tag{25.8}$$

is satisfied, and the final regret bound

$$\mathbf{Rel}_n(\varnothing) = -n\mathbb{E}\phi + n\mathbb{E}\phi - \frac{0}{2} = 0 \tag{25.9}$$

as desired. It remains to prove admissibility. To this end, observe that the minimum of

$$\sup_{y_t} \mathbb{E}_{\widehat{\boldsymbol{y}}_t \sim q_t} \left\{ \ell(\widehat{\boldsymbol{y}}_t, y_t) + \mathbf{Rel}_n(y_1,\dots,y_t) \right\}$$

has the solution given by (22.18) (and no clipping is necessary because of the stability assumption on ϕ). Therefore, it is enough to check (22.20), which, for the case of no side information, becomes

$$\frac{1}{2} + \mathbb{E}_{y_t} \mathbf{Rel}_n\left((x_1,y_1),\dots,(x_t,y_t)\right) \leq \mathbf{Rel}_n\left((x_1,y_1),\dots,(x_{t-1},y_{t-1})\right). \tag{25.10}$$

This condition is true by the definition (25.7), concluding the proof.

25.3   Adaptive Gradient Descent

Part IV

Extensions

26 The Minimax Theorem

At the beginning of the course, we phrased a variety of learning problems through the language of minimaxity. Such formulation turned out to be especially fruitful for the problem of Sequential Prediction, where we were able to upper bound the minimax value by a supremum of a certain stochastic process. The very first step of the proof involved exchanging the min and the max. We never fully justified this step, and simply assumed that the necessary conditions are satisfied for us to proceed.

In the simplest (bilinear) form, the minimax theorem is due to von Neumann:

Theorem 26.1. *Let $M \in \mathbb{R}^{p \times m}$ for some $p, m \geq 1$, and let Δ_p and Δ_m denote the set of distributions over the rows and columns, respectively. Then*

$$\min_{a \in \Delta_p} \max_{b \in \Delta_m} a^\top M b = \max_{b \in \Delta_m} \min_{a \in \Delta_p} a^\top M b \tag{26.1}$$

Since von Neumann, there have been numerous extensions of this theorem, generalizing the statement to uncountable sets of rows/columns, and relaxing the assumptions on the payoff function (which is bilinear in the above statement). To make our discussion more precise, consider the following definition:

Definition 26.2. Consider sets $\mathcal{A}$ and $\mathcal{B}$ and a function $\ell : \mathcal{A} \times \mathcal{B} \mapsto \mathbb{R}$. We say that *the minimax theorem holds* for the triple $(\mathcal{A}, \mathcal{B}, \ell)$ if

$$\inf_{a \in \mathcal{A}} \sup_{b \in \mathcal{B}} \ell(a, b) = \sup_{b \in \mathcal{B}} \inf_{a \in \mathcal{A}} \ell(a, b) \tag{26.2}$$

With this definition, von Neumann's result says that the minimax theorem holds for the triple $(\Delta_p, \Delta_m, \ell)$ with $\ell(a, b) = a^\top M b$ for $a \in \Delta_p, b \in \Delta_m$.

In this lecture, we prove a rather general version of the minimax theorem using the Exponential Weights algorithms (the proof is borrowed from [16]), and discuss the important relationship between the minimax theorem and regret minimization. As we show, this relationship is quite subtle.

26.1   When the Minimax Theorem Does Not Hold

We now give two examples for which the minimax theorem does not hold. The first is the famous "Pick the Bigger Integer" game, attributed to Wald [56].

Example 20. The two players in the zero sum game pick positive integers, and the one with the larger number wins. It is not hard to see that the player making the first move has a disadvantage. As soon as her mixed strategy is revealed, the opponent can simply choose a number in the 99th percentile of this distribution in order to win most of the time. The same holds when the order is switched, now to the advantage of the other player, and so the minimax theorem does not hold. More precisely, let $\mathcal{A} = \mathcal{B} = \Delta(\mathbb{N})$ be the two sets, and $\ell(a,b) = \mathbb{E}_{w \sim a, v \sim b} \mathbf{I}\{w \leq v\}$. The minimax theorem does not hold for the triple $(\Delta(\mathbb{N}), \Delta(\mathbb{N}), \ell)$, as the gap between $\inf_a \sup_b \ell(a,b)$ and $\sup_b \inf_a \ell(a,b)$ can be made as close to 1 as we want.

The above example might lead us to believe that the reason the minimax theorem does not hold is because we consider distributions over an unbounded set $\mathbb{N}$. However, the above game can be embedded into the $[0,1]$ interval as follows:

Example 21. The two players in the zero sum game pick real numbers in the interval $[0,1]$. Any player that picks 1 loses (it is a tie if both pick 1); otherwise, the person with the largest real number wins. The setup forces the players into the same situation as in the "Pick the Bigger Integer" example, as the larger number wins while the limit point 1 should be avoided. Let $\mathcal{A} = \mathcal{B} = \Delta([0,1])$ and $\ell(a,b) = \mathbb{E}_{w \sim a, v \sim b} \tilde{\ell}(w,v)$ with

$$\tilde{\ell}(w,v) = \begin{cases} 1 & \text{if } w = 1, v \neq 1, \\ 0 & \text{if } w \neq 1, v = 1, \\ 0.5 & \text{if } w = v = 1 \\ \mathbf{I}\{w \leq v\} & \text{otherwise} \end{cases}$$

We can use the same reasoning as in Example 20 to show that the minimax theorem does not hold for the triple $(\Delta([0,1]), \Delta([0,1]), \ell)$.

26.2   The Minimax Theorem and Regret Minimization

From the previous section, we see that there exist some rather simple cases when the minimax theorem fails to hold. What can we say about regret minimization in these situations? Of course, if the minimax theorem does not hold, our proof technique of upper bounding the value by exchanging the infima and suprema fails. However, could it be that a regret minimizing strategy exists, but we must develop it algorithmically without even considering the value of the prediction problem? Surprisingly, this is not the case. In other words, any time the minimax theorem fails, we have no hope of minimizing regret.

Proposition 26.3. *Let $\mathcal{F}$ and $\mathcal{Z}$ be the sets of moves of the learner and Nature, respectively. Let $\ell : \mathcal{F} \times \mathcal{Z} \mapsto \mathbb{R}$ be a bounded loss function. If the minimax theorem does not hold for the triple $\big(\Delta(\mathcal{F}), \Delta(\mathcal{Z}), \mathbb{E}_{f,z}\ell(f,z)\big)$, the regret is lower bounded by a constant for any n, and thus the problem is not learnable. Equivalently, if there is a regret-minimizing strategy, i.e. the value of the game satisfies*

$$\lim_{n\to\infty} \mathcal{V}^{seq}(\mathcal{F}, n) = 0$$

then it must be that the minimax theorem holds for the triple $\big(\Delta(\mathcal{F}), \Delta(\mathcal{Z}), \mathbb{E}_{f,z}\ell(f,z)\big)$.

Proof of Proposition 26.3. If the one-round game does not have a value, there exists some constant $c > 0$ such that

$$\inf_{q\in\Delta(\mathcal{F})} \sup_{z\in\mathcal{Z}} \mathbb{E}_{f\sim q}\ell(f,z) \geq \sup_{p\in\Delta(\mathcal{Z})} \inf_{f\in\mathcal{F}} \mathbb{E}_{z\sim p}\ell(f,z) + c \tag{26.3}$$

Let $\pi = \{\pi_t\}$ denote a randomized strategy of the learner with $\pi_t : \mathcal{Z}^{t-1} \mapsto \Delta(\mathcal{F})$. Let $\mathbb{E}_\pi$ denote the randomization under the strategy π, and let $\tau = \{\tau_t\}$ be the deterministic strategy of Nature. By definition, the value is

$$\mathcal{V}^{seq}(\mathcal{F}, n) = \inf_{\pi}\sup_{\tau} \mathbb{E}_\pi \left\{ \frac{1}{n}\sum_{t=1}^{n} \ell(f_t, z_t) - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(f, z_t) \right\}$$

$$= \inf_{\pi}\sup_{\tau} \mathbb{E}_\pi \left\{ \frac{1}{n}\sum_{t=1}^{n} \ell(f_t, z_t) - \inf_{f\in\mathcal{F}} \mathbb{E}_{z\sim\hat{p}}\ell(f,z) \right\}$$

$$\geq \inf_{\pi}\sup_{\tau} \mathbb{E}_\pi \left\{ \frac{1}{n}\sum_{t=1}^{n} \ell(f_t, z_t) - \sup_{p}\inf_{f\in\mathcal{F}} \mathbb{E}_{z\sim p}\ell(f,z) \right\}$$

$$= \left\{ \inf_{\pi}\sup_{\tau} \mathbb{E}_\pi \frac{1}{n}\sum_{t=1}^{n} \ell(f_t, z_t) \right\} - \left\{ \sup_{p}\inf_{f\in\mathcal{F}} \mathbb{E}_{z\sim p}\ell(f,z) \right\}$$

Now,

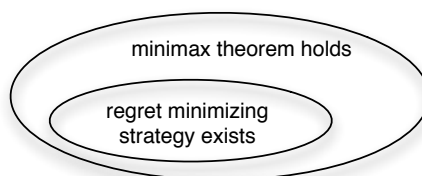$$\inf_{\pi}\sup_{\tau}\mathbb{E}_{\pi}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(f_t,z_t) = \frac{1}{n}\left\langle\!\!\left\langle \inf_{q_t\in\Delta(\mathcal{F})}\sup_{z_t\in\mathcal{Z}}\mathbb{E}_{f_t\sim q_t}\right\rangle\!\!\right\rangle_{t=1}^{n}\left\{\sum_{t=1}^{n}\boldsymbol{\ell}(f_t,z_t)\right\}$$

$$\geq \sup_{p\in\Delta(\mathcal{Z})}\inf_{f\in\mathcal{F}}\mathbb{E}_{z\sim p}\boldsymbol{\ell}(f,z)+c\,,$$

and the statement follows. The converse statement is immediate. □

Proposition 26.3 effectively says that the set of problems that are learnable in the sequential prediction framework is a subset of problems for which the one-shot minimax theorem holds:

nhe inclusion suggests the following intriguing "amplification" strategy for proving regret bounds: first, find a weak learning method with a possibly suboptimal bound that nevertheless ensures vanishing regret; this implies the minimax theorem, which in turn implies that the machinery developed so far in the course can be used to find optimal or near-optimal algorithms (more on this in the next few lectures!)

Of course, Proposition 26.3 only tells us about the "bilinear" form of the minimax theorem, since it only considers mixed strategies over $\mathcal{F}$ and $\mathcal{Z}$. The proof of the Proposition, however, can be extended as follows:

Proposition 26.4. *Let $\mathcal{A}$ and $\mathcal{B}$ be two sets, and $\mathcal{B}$ is convex. Let $\boldsymbol{\ell}:\mathcal{A}\times\mathcal{B}\mapsto\mathbb{R}$ be a bounded function such that $\boldsymbol{\ell}(a,\cdot)$ is concave for any $a\in\mathcal{A}$. If the minimax theorem does not hold for the triple $(\mathcal{A},\mathcal{B},\boldsymbol{\ell})$, then the regret*

$$\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(a_t,b_t)-\inf_{a\in\mathcal{A}}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(a,b_t) \tag{26.4}$$

is lower bounded by a constant for any n, for any regret minimization strategy that deterministically chooses a_t's. Equivalently, if there is a regret-minimizing strategy that deterministically chooses a_t's, then it must be that the minimax theorem holds for the triple $(\mathcal{A},\mathcal{B},\boldsymbol{\ell})$.

✎ **Exercise 26.1** (⋆)**.** Prove Proposition 26.4.

Proposition 26.3 is a consequence of Proposition 26.4, obtained by taking $\mathcal{A} = \Delta(\mathcal{F})$, $\mathcal{B} = \Delta(\mathcal{Z})$, and the loss function as the expectation in the two mixed strategies. In an obvious manner, we are treating the deterministic choice of mixed strategies as equivalent to randomized choices over $\mathcal{F}$ and $\mathcal{Z}$.

Online convex optimization is a class of problems where a deterministic strategy can minimize regret over $\mathcal{A} = \mathcal{F}$ because the loss function $\boldsymbol{\ell}(f, z)$ is convex in $f \in \mathcal{F}$. What is interesting, a regret minimization strategy for online convex optimization can be used to prove more general versions of the minimax theorem than the bilinear form. We can do so by engineering an auxiliary regret minimization game which, with some extra work, implies the minimax theorem. This is shown in the next section.

26.3 Proof of a Minimax Theorem Using Exponential Weights

The following theorem and its proof are taken almost verbatim rom [16].

Theorem 26.5 (Theorem 7.1 in [16])**.** *Suppose $\boldsymbol{\ell} : \mathcal{A} \times \mathcal{B} \mapsto \mathbb{R}$ is a bounded function, where $\mathcal{A}$ and $\mathcal{B}$ are convex and $\mathcal{A}$ is compact. Suppose that $\boldsymbol{\ell}(\cdot, b)$ is convex and continuous for each fixed $b \in \mathcal{B}$ and $\boldsymbol{\ell}(a, \cdot)$ is concave for each $a \in \mathcal{A}$. Then the minimax theorem holds for $(\mathcal{A}, \mathcal{B}, \boldsymbol{\ell})$.*

Proof. The direction

$$\inf_{a \in \mathcal{A}} \sup_{b \in \mathcal{B}} \boldsymbol{\ell}(a, b) \geq \sup_{b \in \mathcal{B}} \inf_{a \in \mathcal{A}} \boldsymbol{\ell}(a, b) \tag{26.5}$$

is immediate, and it holds without any assumptions on the function and its domain.

Without loss of generality, let $\boldsymbol{\ell}(a, b) \in [0, 1]$ for all $a \in \mathcal{A}, b \in \mathcal{B}$. Fix an $\epsilon > 0$ and let $a^{(1)}, \ldots, a^{(N)}$ be centers of an ϵ-cover of $\mathcal{A}$, which is finite by the assumption on compactness of $\mathcal{A}$. We now treat the elements a^i as experts and update an exponential weights distribution according to the following construction. The sequences $a_1, \ldots, a_n$ and $b_0, \ldots, b_n$ are defined recursively, with b_0 chosen arbitrarily. Let

$$a_t = \frac{\sum_{i=1}^{N} a^{(i)} \exp\left\{-\eta \sum_{s=1}^{t-1} \boldsymbol{\ell}(a^{(i)}, b_s)\right\}}{\sum_{i=1}^{N} \exp\left\{-\eta \sum_{s=1}^{t-1} \boldsymbol{\ell}(a^{(i)}, b_s)\right\}}$$

with $\eta = \sqrt{\frac{8\ln N}{n}}$ and b_t chosen so that $\ell(a_t, b_t) \geq \sup_{b \in \mathcal{B}} \ell(a_t, b) - 1/n$. We now appeal to the version of Exponential Weights for Prediction with Expert Advice, and its regret bound given in (18.4):

$$\frac{1}{n}\sum_{t=1}^{n}\ell(a_t, b_t) \leq \min_{i \in \{1,\dots,N\}}\frac{1}{n}\sum_{t=1}^{n}\ell(a^{(i)}, b_t) + \sqrt{\frac{\ln N}{2n}} \tag{26.6}$$

Then,

$$\inf_{a \in \mathcal{A}}\sup_{b \in \mathcal{B}}\ell(a, b) \leq \sup_{b \in \mathcal{B}}\ell\left(\frac{1}{n}\sum_{t=1}^{n}a_t, b\right) \leq \sup_{b \in \mathcal{B}}\frac{1}{n}\sum_{t=1}^{n}\ell(a_t, b) \leq \frac{1}{n}\sum_{t=1}^{n}\sup_{b \in \mathcal{B}}\ell(a_t, b) \tag{26.7}$$

by convexity of ℓ in the first argument. Thanks to the definition of b_t and the regret bound,

$$\frac{1}{n}\sum_{t=1}^{n}\sup_{b \in \mathcal{B}}\ell(a_t, b) \leq \frac{1}{n}\sum_{t=1}^{n}\ell(a_t, b_t) + \frac{1}{n} \leq \min_{i \in \{1,\dots,N\}}\frac{1}{n}\sum_{t=1}^{n}\ell(a^{(i)}, b_t) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n} \tag{26.8}$$

Concavity of ℓ in the second argument implies

$$\min_{i \in \{1,\dots,N\}}\frac{1}{n}\sum_{t=1}^{n}\ell(a^{(i)}, b_t) \leq \min_{i \in \{1,\dots,N\}}\ell\left(a^{(i)}, \frac{1}{n}\sum_{t=1}^{n}b_t\right) \leq \sup_{b \in \mathcal{B}}\min_{i \in \{1,\dots,N\}}\ell\left(a^{(i)}, b\right) \tag{26.9}$$

We conclude that

$$\inf_{a \in \mathcal{A}}\sup_{b \in \mathcal{B}}\ell(a, b) \leq \sup_{y \in \mathcal{Y}}\min_{i \in \{1,\dots,N\}}\ell(a^{(i)}, b) + \sqrt{\frac{\ln N}{2n}} + \frac{1}{n},$$

which implies

$$\inf_{a \in \mathcal{A}}\sup_{b \in \mathcal{B}}\ell(a, b) \leq \sup_{y \in \mathcal{Y}}\min_{i \in \{1,\dots,N\}}\ell(a^{(i)}, b)$$

if we let n go to infinity. It remains to let $\epsilon \to 0$ and use continuity of ℓ to conclude

$$\inf_{a \in \mathcal{A}}\sup_{b \in \mathcal{B}}\ell(a, b) \leq \sup_{b \in \mathcal{B}}\inf_{a \in \mathcal{A}}\ell(a, b) . \tag{26.10}$$

$\square$

✎ **Exercise 26.2** (★★)**.** Suppose $\mathcal{A}$ and $\mathcal{B}$ are unit balls in an infinite dimensional Hilbert space (hence, $\mathcal{A}$ and $\mathcal{B}$ are not compact). Let $\ell : \mathcal{A} \times \mathcal{B} \mapsto \mathbb{R}$ be such that $\ell(\cdot, b)$ is convex and 1-Lipschitz (in the Hilbert space norm) for each fixed $b \in \mathcal{B}$, and $\ell(a, \cdot)$ is concave for each $a \in \mathcal{A}$. Prove that the minimax theorem holds for $(\mathcal{A}, \mathcal{B}, \ell)$.

✎ **Exercise 26.3** (⋆⋆)**.** As in the previous exercise, suppose $\mathcal{A}$ and $\mathcal{B}$ are unit balls in an infinite dimensional Hilbert space. Let $\boldsymbol{\ell} : \mathcal{A} \times \mathcal{B} \mapsto \mathbb{R}$ be such that $\boldsymbol{\ell}(\cdot, b)$ is convex and $|\boldsymbol{\ell}(\cdot, b)| \le 1$ for each fixed $b \in \mathcal{B}$, and $\boldsymbol{\ell}(a, \cdot)$ is concave for each $a \in \mathcal{A}$. Prove that the minimax theorem holds for $(\mathcal{A}, \mathcal{B}, \boldsymbol{\ell})$.

26.4 More Examples

Example 22. We now present a problem for which the minimax theorem holds, yet the associated regret minimization problem is impossible. This supports the strict subset in Figure 26.2.

Consider the example of sequential prediction with thresholds on the unit interval. It was shown in Theorem 8.2 that the problem is not learnable. The question now is whether the minimax theorem holds. more precisely, we let $\mathcal{F} = \mathcal{X} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$, and

$$\boldsymbol{\ell}(f, (x, y)) = \mathbf{I}\big\{\mathbf{I}\big\{x \le f\big\} \ne y\big\} = \mathbf{I}\big\{(x \le f \wedge y = 0) \vee (x > f \wedge y = 1)\big\} \,.$$

Here, we are using f to denote the threshold location instead of θ. Consider the sets of mixed strategies $\Delta(\mathcal{F})$ and $\Delta(\mathcal{X} \times \mathcal{Y})$. Let us inspect

$$\inf_{q \in \Delta(\mathcal{F})} \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathbb{E}_{f \sim q} \boldsymbol{\ell}(f, (x, y))$$

Consider a q that puts half of the mass at $f = 1$ and half at $f = 0$. For any (x, y), the expected loss is $1/2$. To prove the minimax theorem, it remains to show that

$$\sup_{p \in \Delta(\mathcal{X} \times \mathcal{Y})} \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y) \sim p} \boldsymbol{\ell}(f, (x, y)) \ge \frac{1}{2}$$

This can be achieved, for instance, by taking p_x as $\delta_{1/2}$ and $p_{y|x}$ as a fair coin flip.

In fact, using the sufficient conditions in terms of quasi-convexity, presented later in this section, it is possible to directly conclude that the minimax theorem holds.

We conclude that the problem of prediction with thresholds is not learnable, yet the minimax theorem holds.

Example 23. Now consider the example given in Section 13.1:

$$\mathcal{F} = \{f_a : \ a \in [0, 1], \ f_a(x) = 0 \ \forall x \ne a, \ f_a(a) = 1\} \tag{26.11}$$

over the domain $\mathcal{X} = [0,1]$. Let $\mathcal{Y} = \{0,1\}$ and define the loss function $\boldsymbol{\ell}$ as the indicator of a mistake. It was shown that the problem is learnable, and (as Example 16 shows) the Littlestone's dimension is 1. Given Proposition 26.3, it must be the case that the minimax theorem holds.

✎ **Exercise 26.4** (⋆)**.** Give a direct proof that the minimax theorem holds for the triple $(\Delta([0,1]), \Delta([0,1] \times \{0,1\}), \boldsymbol{\ell})$.

Discussion: The two examples presented above together with Example 21 presented earlier paint a fairly interesting picture. In Example 21, the minimax theorem does not hold, and it is not possible to minimize regret. In Example 22, the minimax theorem holds, yet the prediction problem is impossible. Finally, in Example 23, both the minimax theorem holds, and regret can be (trivially) minimized. What is rather curious is that the sets of moves and the loss functions look quite similar across these examples, as a mixture of some point discontinuity and a threshold. While finiteness of the Littlestone's dimension characterizes learnability within the supervised learning class of problems for which the minimax theorem holds, it is not clear under which conditions the minimax theorem itself holds true. The next section presents some sufficient conditions.

26.5   Sufficient Conditions for Weak Compactness

Suppose that $\mathcal{F}$ is a subset of a complete separable metric space and $\mathcal{B}_{\mathcal{F}}$ is the σ-field of Borel subsets of $\mathcal{F}$. Let $\Delta(\mathcal{F})$ denote the set of all probability measures on $\mathcal{F}$. Similarly, let $\Delta(\mathcal{X})$ be the set of all probability measures on $\mathcal{X}$. Under consideration is the question of conditions on $\mathcal{F}$ and $\mathcal{X}$ that guarantee that the minimax theorem holds for the triplet $(\Delta(\mathcal{F}), \Delta(\mathcal{X}), \mathbb{E}_{f,x}\boldsymbol{\ell})$ for a bounded measurable function $\boldsymbol{\ell} : \mathcal{F} \times \mathcal{X} \mapsto \mathbb{R}$. In addressing this question, we appeal to the following result (in a slightly modified form from the original):

Theorem 26.6 (Theorem 3 in [49])**.** *Let A be a nonempty convex weakly compact subset of a Banach space E. Let B be a nonempty convex subset of the dual space, E', of E. Then*

$$\inf_{a \in A} \sup_{b \in B} \langle a, b \rangle = \sup_{b \in B} \inf_{a \in A} \langle a, b \rangle$$

Under a condition we outline below, we can use the above theorem for our purposes. To this end, write

$$\langle g_p, q \rangle = \mathbb{E}_{f \sim q, x \sim p} \ell(f, x) \quad \text{and} \quad G = \{g_p : p \in \Delta(\mathcal{X})\}$$

The desired minimax identity can now be written as

$$\inf_{q \in \Delta(\mathcal{F})} \sup_{g \in G} \langle g, q \rangle = \sup_{g \in G} \inf_{q \in \Delta(\mathcal{F})} \langle g, q \rangle.$$

We can view g_p as a linear functional on $\Delta(\mathcal{F})$, the set of all Borel probability measures on $\mathcal{F}$.

Condition 26.7 (Continuity)**.** *Assume that $\Delta(\mathcal{F})$ is a subset of a Banach space E (that is, all the linear functionals defined by $q \in \Delta(\mathcal{F})$ are* continuous*).*

Under the above assumption, it can be verified that the dual E' contains the set of all bounded Borel measurable functions on $\mathcal{F}$. In this case, $G \subseteq E'$. To appeal to the above theorem it is therefore enough to delineate conditions under which $\Delta(\mathcal{F})$ is weakly compact (clearly, $\Delta(\mathcal{F})$ is convex).

Condition 26.8 (Weak Compactness)**.** $\Delta(\mathcal{F})$ *is weakly compact.*

By a fundamental result of Prohorov, weak compactness of the set of probability measures on a complete separable metric space is equivalent to uniform tightness (see e.g. [11, Theorem 8.6.2.], [54]). Depending on the application, various assumptions on $\mathcal{F}$ can be imposed to guarantee uniform tightness of $\Delta(\mathcal{F})$. First, if $\mathcal{F}$ itself is compact, then $\Delta(\mathcal{F})$ is tight, and hence (under the continuity condition) the minimax theorem holds. In infinite dimensions, however, compactness of $\mathcal{F}$ is a very restrictive property. Thankfully, tightness can be established under more general assumptions on $\mathcal{F}$, as we show next.

By Example 8.6.5 (ii) in [11], a family $\Delta(\mathcal{F})$ of Borel probability measures on a separable *reflexive* Banach space E is uniformly tight (under the weak topology) precisely when there exists a function $V : E \mapsto [0, \infty)$ continuous in the norm topology such that

$$\lim_{\|f\| \to \infty} V(f) = \infty \quad \text{and} \quad \sup_{q \in \Delta(\mathcal{F})} \mathbb{E}_{f \sim q} V(f) < \infty.$$

For instance, if $\mathcal{F}$ is a subset of a ball in E, it is enough to take $V(f) = \|f\|$. We conclude that the minimax theorem holds whenever $\mathcal{F}$ is a subset of a ball in a separable reflexive Banach space, and the continuity condition holds.

27

Two Proofs of Blackwell's Approachability Theorem

The purpose of this lecture is two-fold. The first goal is to introduce a generalization of the minimax theorem to vector-valued payoffs, due to David Blackwell [8]. The theorem has been an invaluable tool for the analysis of repeated games. Its proof is simple and geometric, while the result is quite non-trivial. The second goal of the lecture is to show that the symmetrization ideas we employed for analyzing minimax regret can be extended to notions beyond regret. We exhibit another (and very different) proof, which is non-constructive but also more general than the constructive approach.

Let $\ell : \mathcal{F} \times \mathcal{X} \mapsto \mathbb{R}$ be the loss (payoff) function, which need not be convex. For the purposes of this lecture let us introduce the notation

$$\ell(q, p) \triangleq \mathbb{E}_{f \sim q, x \sim p} \ell(f, x)$$

for $q \in \Delta(\mathcal{F})$ and $p \in \Delta(\mathcal{X})$. Since $\ell(q, p)$ is a bi-linear function in the mixed strategies, the minimax theorem

$$\inf_{q \in \Delta(\mathcal{F})} \sup_{p \in \Delta(\mathcal{X})} \ell(q, p) = \sup_{p \in \Delta(\mathcal{X})} \inf_{q \in \Delta(\mathcal{F})} \ell(q, p) \tag{27.1}$$

holds under quite general assumptions on the (not necessarily finite) $\mathcal{F}$ and $\mathcal{X}$, as we have seen in Chapter 26. We assume that the necessary conditions for the minimax theorem hold.

Suppose we take the point of view of the row player (that is, we are choosing from $\mathcal{F}$). Then the right-hand-side of Eq. (27.1) can be interpreted as "what is the smallest loss we can incur if we know what p the column player chooses". Such

a response can be a pure action. For the left-hand-side, however, we are required to furnish an "oblivious" randomized strategy that will "work" no matter what the opponent chooses. Let the *value* of the expression in (27.1) be denoted by V. The minimax statement, from our point of view, can be written as

$$\exists q \in \Delta(\mathcal{F}) \text{ s.t. } \forall p \in \Delta(\mathcal{X}) \quad \ell(q, p) \leq V \iff \forall p \in \Delta(\mathcal{X}) \exists q \in \Delta(\mathcal{F}) \text{ s.t. } \ell(q, p) \leq V \tag{27.2}$$

It is remarkable that being able to respond to the opponent is equivalent to putting down a randomized strategy ahead of time and not worrying about opponent's choice.

27.1 Blackwell's vector-valued generalization and the original proof

It is often the case that our utility cannot be measured by a single number, as our goals are incomparable. We would like to make decisions that gives us desired payoffs along multiple dimensions. David Blackwell came up with a wonderful generalization of the real-valued minimax theorem.

Consider a vector-valued payoff $\ell : \mathcal{F} \times \mathcal{X} \mapsto \mathcal{B}$ where $\mathcal{B}$ is some Banach space (in the original Blackwell's formulation and in his original proof, $\mathcal{B} = \mathbb{R}^d$; his results have been subsequently extended to Hilbert spaces). It is natural to define the goal as a set $S \subset \mathcal{B}$ of payoffs with which we would be content. Von Neumann's Minimax Theorem can then be seen as a statement about a set $S = (-\infty, c]$: we can ensure that our loss is in S for any $c \leq V$ but not for any $c > V$.

Now that the generalization to vector-valued payoffs is established, is it still true that an equivalent of (27.2) holds? That is

$$\exists q \in \Delta(\mathcal{F}) \text{ s.t. } \forall p \in \Delta(\mathcal{X}) \quad \ell(q, p) \in S \overset{?}{\iff} \forall p \in \Delta(\mathcal{X}) \exists q \in \Delta(\mathcal{F}) \text{ s.t. } \ell(q, p) \in S$$

This turns out to be not necessarily the case even for convex sets S. A simple example (borrowed from [1]) is a game where $\mathcal{F} = \mathcal{X} = \{0, 1\}$, the payoff $\ell(f, x) = (f, x)$ and $S = \{(z, z) : z \in [0, 1]\}$. Clearly, for any $p \in \Delta(\mathcal{X})$, we can choose $q = p$ so that $\ell(q, p) \in S$; however, no silver-bullet strategy $q \in \Delta(\mathcal{F})$ exists that will work for all p.

223

27.1 Blackwell's vector-valued generalization and the original proof

We see that being able to respond to the mixed strategy of the opponent no longer guarantees that we can put down an oblivious mixed strategy that will work no matter what she does. In this respect, the one-dimensional von Neumann's minimax theorem is a very special case.

It is, however, the case that being able to respond is equivalent to being able to put down an oblivious strategy for any half-space $H = \{u : \langle u, a \rangle \le c\}$ containing S:

Theorem 27.1 (Blackwell, 1956 [8])**.**

$$\forall p \ \exists q \ s.t. \ \ell(q, p) \in S \iff \forall H \supset S, \ \exists q \ s.t. \ \forall p \ \ell(q, p) \in H$$

Proof. Suppose there exists a p such that $\ell(q, p) \notin S$ for all q. Then there exists a hyperplane separating S from the set $\{\ell(q, p) : q \in \Delta(\mathcal{F})\}$. Clearly, for this set hyperplane there exists no q ensuring that $\ell(q, p) \in H$ for all p.

The other direction is proved by taking any half-space $H = \{u : \langle u, a \rangle \le c\}$ and defining a scalar auxiliary game $\ell'(q, p) := \langle a, \ell(q, p) \rangle$. Being able to respond to any p with a q that satisfies $\ell(q, p) \in S$ means that we are able to respond in the scalar game such that the loss is at most c. But for scalar games, von Neumann's minimax theorem says that being able to respond is equivalent to existence of an oblivious strategy. This concludes the proof. □

At this point, the utility of the above theorem is not clear. The multi-dimensional nature of the payoff unfortunately means that there is no silver-bullet mixed strategy that will get us what we want. The key insight of Blackwell is that we can get what we want in the long run by having an *adaptive* strategy which tunes to the behavior of the opponent. That is, we can *learn* opponent's behavior and play in a way that will make the average payoff approach the set S.

Definition 27.2. A set S is approachable by the row player if the row player has a strategy such that no matter how the column player plays,

$$\lim_{n \to \infty} d\left(\frac{1}{n} \sum_{t=1}^{n} \ell(f_t, x_t), S\right) = 0 \quad \text{a.s.}$$

In the above definitions, n is the number of stages in the repeated game, f_t is a random draw from our strategy q_t at time t, and x_t is the opponent's move; d is a distance (e.g. the norm in the Banach space).

Approachability is trivial for the scalar game: we can approach the set $S = (-\infty, c]$ if and only if $c \le V$. The same is true for half-space approachability: a

half-space $H = \{u : \langle u, a \rangle \le c\}$ is approachable if and only if there exists q such that for all p, $\langle a, \ell(q, p) \rangle \le c$.

It is remarkable that Blackwell completely characterized situations in which S is approachable. In fact, the situation is quite simple: a set is approachable if and only if we can respond to the mixed strategy of the opponent. That is if and only if $\forall p \; \exists q$, s.t. $\ell(q, p) \in S$. By the previous Theorem, it is also equivalent to approachability of every half-space containing S:

Theorem 27.3 (Blackwell, 1956 [8]). *A closed convex set S is approachable iff every half-space containing S is approachable.*

Proof. Let $L_t = \frac{1}{t} \sum_{s=1}^{t} \ell(f_s, x_s)$. We would like to make this average approach S. The idea is that we can make progress towards the set by attempting to approach a hyperplane perpendicular to the current projection direction from L_t to the set. This will work because points sufficiently close to L_{t-1} along the segment connecting L_{t-1} and the new payoff (on average on the other side of the hyperplane) are closer to the set S than L_{t-1}. Here is the explicit adaptive strategy we are going to use. At stage t, play that q_t which satisfies

$$\sup_{p} \langle a_{t-1}, \ell(q_t, p) \rangle \le c_{t-1}$$

where
$$a_{t-1} = \frac{L_{t-1} - \pi_S(L_{t-1})}{\|L_{t-1} - \pi_S(L_{t-1})\|}, \quad c_{t-1} = \langle a_{t-1}, \pi_S(L_{t-1}) \rangle.$$

Here π_S is the Euclidean projection onto S. Note that q_t exists by our assumption that we can put down an oblivious mixed strategy for any hyperplane. The rest is a bit of algebra:

$$
\begin{aligned}
d(L_t, S)^2 &= \|L_t - \pi_S(L_t)\|^2 \\
&\le \|L_t - \pi_S(L_{t-1})\|^2 \\
&= \left\| \frac{t-1}{t}(L_{t-1} - \pi_S(L_{t-1})) + \frac{1}{t}(\ell(f_t, x_t) - \pi_S(L_{t-1})) \right\|^2 \\
&= \left(\frac{t-1}{t} \right)^2 d(L_{t-1}, S)^2 + \frac{1}{t^2} \|\ell(f_t, x_t) - \pi_S(L_{t-1})\|^2 \\
&\quad + 2\frac{t-1}{t^2} \langle L_{t-1} - \pi_S(L_{t-1}), \ell(f_t, x_t) - \pi_S(L_{t-1}) \rangle
\end{aligned}
$$

We can rewrite this as

$$t^2 d(L_t, S)^2 - (t-1)^2 d(L_{t-1}, S)^2 \le const + 2(t-1) \langle L_{t-1} - \pi_S(L_{t-1}), \ell(f_t, x_t) - \pi_S(L_{t-1}) \rangle,$$

where we assumed that the payoffs are bounded in $\| \cdot \|$.

Summing the inequalities over all stages, and using the fact that

$$\langle L_{t-1} - \pi_S(L_{t-1}), \ell(q_t, p_t) - \pi_S(L_{t-1}) \rangle \le 0$$

we are left with an average of martingale differences

$$\langle L_{t-1} - \pi_S(L_{t-1}), \ell(f_t, x_t) - \ell(q_t, p_t) \rangle \le 0$$

which tends to zero almost surely. □

27.2 A non-constructive proof

We now give a very different proof of Blackwell's Approachability (Theorem 27.3). While it is non-constructive (that is, we do not exhibit a strategy for playing the game), the resulting statement is quite a bit more general and sharp (see [45] for more details).

The idea is to bite the bullet and write down the minimax value of the repeated game:

$$V_n = \inf_{q_1} \sup_{p_1} \mathop{\mathbb{E}}_{\substack{x_1 \sim p_1 \\ f_1 \sim q_1}} \dots \inf_{q_n} \sup_{p_n} \mathop{\mathbb{E}}_{\substack{x_n \sim p_n \\ f_n \sim q_n}} d\left(\frac{1}{n} \sum_{t=1}^{n} \ell(f_t, x_t), S \right)$$

An upper bound on this value means that there exists a sequential strategy (way of picking q_t's based on the past history) such that the expected distance to the set is at most this bound. On the other hand, a lower bound gives a guarantee to the opponent.

In view of Theorem 27.1, to prove Theorem 27.3 (in expectation) it is enough to show that the value V_n decays to zero under the assumption that for any strategy p of the opponent, there exists a $q^*(p)$ such that $\ell(q^*(p), p) \in S$.

Note that in each pair in V_n, infimum and supremum can be exchanged because what is inside of the pair is a bilinear function (expectation with respect to q_t and p_t). Importantly, p_t and q_t only appear in the respective tth expectation,

but nowhere else in the expression (for, in that case, it might not be a bilinear form). We arrive at

$$V_n = \sup_{p_1}\inf_{q_1}\mathop{\mathbb{E}}_{\substack{x_1\sim p_1\\f_1\sim q_1}}\ldots\sup_{p_n}\inf_{q_n}\mathop{\mathbb{E}}_{\substack{x_n\sim p_n\\f_n\sim q_n}}d\left(\frac{1}{n}\sum_{t=1}^n\ell(f_t,x_t),S\right)$$

We can now add and subtract a term involving the expected payoffs:

$$V_n = \sup_{p_1}\inf_{q_1}\mathop{\mathbb{E}}_{\substack{x_1\sim p_1\\f_1\sim q_1}}\ldots\sup_{p_n}\inf_{q_n}\mathop{\mathbb{E}}_{\substack{x_n\sim p_n\\f_n\sim q_n}}\left\{d\left(\frac{1}{n}\sum_{t=1}^n\ell(f_t,x_t),S\right)-d\left(\frac{1}{n}\sum_{t=1}^n\ell(q_t,p_t),S\right)\right.$$
$$\left.+d\left(\frac{1}{n}\sum_{t=1}^n\ell(q_t,p_t),S\right)\right\}$$

It is easy to check that

$$d\left(\frac{1}{n}\sum_{t=1}^n\ell(f_t,x_t),S\right)-d\left(\frac{1}{n}\sum_{t=1}^n\ell(q_t,p_t),S\right)\le\left\|\frac{1}{n}\sum_{t=1}^n(\ell(f_t,x_t)-\ell(q_t,p_t))\right\|$$

We now use the following properties: for functions $C_1(a)$ and $C_2(a)$,

$$\mathbb{E}(C_1(a)+C_2(a))=\mathbb{E}(C_1(a))+\mathbb{E}(C_2(a)),\quad \sup_a(C_1(a)+C_2(a))\le\sup_aC_1(a)+\sup_aC_2(a)$$

while for the infimum

$$\inf_a(C_1(a)+C_2(a))\le\sup_aC_1(a)+\inf_aC_2(a)\,.$$

The sequence minimax expression then splits (from inside out) as

$$V_n\le\sup_{p_1}\sup_{q_1}\mathop{\mathbb{E}}_{\substack{x_1\sim p_1\\f_1\sim q_1}}\ldots\sup_{p_n}\sup_{q_n}\mathop{\mathbb{E}}_{\substack{x_n\sim p_n\\f_n\sim q_n}}\left\|\frac{1}{n}\sum_{t=1}^n(\ell(f_t,x_t)-\ell(q_t,p_t))\right\|$$
$$+\sup_{p_1}\inf_{q_1}\mathop{\mathbb{E}}_{\substack{x_1\sim p_1\\f_1\sim q_1}}\ldots\sup_{p_n}\inf_{q_n}\mathop{\mathbb{E}}_{\substack{x_n\sim p_n\\f_n\sim q_n}}d\left(\frac{1}{n}\sum_{t=1}^n\ell(q_t,p_t),S\right)$$

For the second term, we use the assumption that we can respond to any strategy p_t by choosing a $q_t^*(p_t)$ such that $\ell(q_t^*(p_t),p_t)\in S$. This means that the second term

$$\sup_{p_1}\inf_{q_1}\mathop{\mathbb{E}}_{\substack{x_1\sim p_1\\f_1\sim q_1}}\ldots\sup_{p_n}\inf_{q_n}\mathop{\mathbb{E}}_{\substack{x_n\sim p_n\\f_n\sim q_n}}d\left(\frac{1}{n}\sum_{t=1}^n\ell(q_t,p_t),S\right)\le\sup_{p_1}\ldots\sup_{p_n}\left\{d\left(\frac{1}{n}\sum_{t=1}^n\ell(q_t^*(p_t),p_t),S\right)\right\}=0$$

where we used convexity of S in the first step.

The first term is upper bounded by the variation of the worst-case martingale difference sequence with values in $Im(\boldsymbol{\ell}) \subset \mathcal{B}$. We conclude that

$$V_n \le 2 \sup_M \mathbb{E} \left\| \frac{1}{n} \sum_{t=1}^{n} d_t \right\|$$

where the supremum is over all martingale difference sequences with $d_t \in Im(\boldsymbol{\ell})$.

27.3 Discussion

The second proof is non-constructive, yet yields the statement in Banach spaces where martingale convergence holds. In fact, it is easy to show that this bound is tight: there is a lower bound differing in only a constant. To the best of our knowledge, all previous proofs relied on the inner product structure.

As we mentioned, the Blackwell's approachability theorem has been used as a tool for repeated games. Typically, one take the problem at hand (e.g. minimization of *internal regret*), finds a potentially high-dimensional space and a set S, so that, if we can approach the set, we can solve the original problem. It is shown in [45] that one can analyze problems without appealing to Blackwell's approachability and without artificially embedding the problem at hand into a different space. The fact that Blackwell's approachability itself follows from the tools we have developed is a testament to the power of symmetrization. We refer to [45] for more details.

27.4 Algorithm Based on Relaxations: Potential-Based Approachability

28

From Sequential to Statistical Learning: Relationship Between Values and Online-to-Batch

28.1 Relating the Values

Now that we have a good understanding of complexities for both statistical learning and sequential prediction, it is time to draw some explicit connections between these two settings. First of all, the connection can be done at the level of the minimax values. We can think of a minimax value as a measure of complexity of a set $\mathcal{F}$, and it is natural to ask whether the minimax complexity of $\mathcal{F}$ is greater for sequential or for statistical learning. We have already seen that the sequential Rademacher complexity is an upper bound on the i.i.d. Rademacher complexity, so the following results should come at no surprise.

For concreteness, consider the setting of supervised learning with absolute loss. Recall the definition of the value $\mathcal{V}^{seq,ab}(\mathcal{F}, n)$ for the sequential problem (see Eq. (5.17)) and $\mathcal{V}^{iid,ab}(\mathcal{F}, n)$ for the statistical learning problem (defined in Eq. (5.9)).

By Corollary 7.15 and the fact that the absolute loss is 1-Lipschitz,

$$\mathcal{V}^{iid,ab}(\mathcal{F}, n) \leq 2 \sup_P \mathcal{R}^{iid}(\mathcal{F}) \tag{28.1}$$

and by (7.18),

$$\mathcal{R}^{iid}(\mathcal{F}) \leq \sup_{x_1,\ldots,x_n} \widehat{\mathcal{R}}^{iid}(\mathcal{F}, x_1,\ldots, x_n) \leq \mathcal{R}^{seq}(\mathcal{F}, n), \tag{28.2}$$

where the middle term is $\overline{\mathcal{R}}^{iid}(\mathcal{F}, n)$ by definition. On the other hand, by an argument very similar to that of Theorem 13.11,

$$\mathcal{R}^{seq}(\mathcal{F}, n) \leq \mathcal{V}^{seq,ab}(\mathcal{F}, n) \ .$$

We conclude that

$$\mathcal{V}^{iid,ab}(\mathcal{F}, n) \leq 2\mathcal{V}^{seq,ab}(\mathcal{F}, n). \tag{28.3}$$

As we suspected, the sequential prediction problem is harder (up to a constant factor) than the i.i.d. learning problem. What is interesting, by Lemma 12.13, for proper learning,

$$\mathcal{V}^{iid,ab}(\mathcal{F}, n) \geq \overline{\mathcal{R}}^{iid}(\mathcal{F}, 2n) - \frac{1}{2}\overline{\mathcal{R}}^{iid}(\mathcal{F}, n) \tag{28.4}$$

This has the following implication: if $\overline{\mathcal{R}}^{iid}(\mathcal{F}, 2n) - \frac{1}{2}\overline{\mathcal{R}}^{iid}(\mathcal{F}, n)$ is of the same order as $\mathcal{R}^{seq}(\mathcal{F}, n)$, then the difficulties of i.i.d. learning and sequential prediction (with adversarially chosen data) are the same. In view of (28.2) this happens when $\mathcal{R}^{seq}(\mathcal{F}, n)$ is not much larger than $\overline{\mathcal{R}}^{iid}(\mathcal{F}, 2n)$. For linear problems this happens when the expected length of a random walk with independent increments is of the same order as the worst possible random walk with martingale differences as increments. Such martingale and i.i.d. convergence is governed, respectively, by the M-type and type of the space. While there exist spaces where the two are different, such spaces are quite exotic and difficult to construct. Thus, for all practical purposes, for linear function classes the classical and sequential Rademacher complexities coincide up to a constant. Good news for linear function classes! Further, we already saw that kernel methods can be viewed as working with a linear class of functions in a certain Reproducing Kernel Hilbert Space. One of the methods of choice in practice, kernel methods are certainly powerful.

Finally, recall that we formalized the above statement (that sequential and i.i.d. Rademacher averages are close) for finite-dimensional linear classes in Assumption 23.1, and indeed the resulting randomized method yielded an upper bound of an i.i.d. (rather than sequential) Rademacher complexity.

More generally, for non-linear classes, we get a precise handle on the relative difficulty of statistical vs adversarial learning by studying the gap between the sequential Rademacher averages and the classical Rademacher averages.

28.2 Online to Batch Conversion

Given the close relationship between the values of the i.i.d. and sequential learning problems, we can ask whether algorithms developed in the sequential setting can be used for statistical learning. Since a regret bound holds for all sequences, it also holds for an i.i.d. sequence. What remains to be done is to construct a single estimator out of the sequence produced by the sequential method, and to convert the regret guarantee into a guarantee about excess loss. Such a process has been dubbed an "Online-to-Batch Conversion" [14]. We already performed such a conversion in Section 4 where a regret bound for regression with individual sequences was converted to a near-optimal bound on excess loss. The argument holds in more generality, and let us state it for convex loss functions.

Lemma 28.1. *Let $\ell : \mathcal{F} \times \mathcal{Z} \mapsto \mathbb{R}$ be convex in the first argument and suppose $\mathcal{F}$ is a convex set[1]. Suppose there exists a strategy for choosing $f_1, \dots, f_n$ in the sequential prediction problem against an oblivious adversary (who picks a sequence $z_1, \dots, z_n$) that guarantees a regret bound*

$$\frac{1}{n} \sum_{t=1}^{n} \ell(f_t, z_t) - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, z_t) \le R_n$$

where R_n may be a function of $z_1, \dots, z_n$. Then the estimator $\bar{f} = \frac{1}{n} \sum_{t=1}^{n} f_t$ enjoys the excess loss bound of $\mathbb{E} R_n$:

$$\mathbb{E} \boldsymbol{L}(\bar{f}) - \inf_{f \in \mathcal{F}} \boldsymbol{L}(f) \le \mathbb{E} R_n$$

In particular, this implies that the value of the statistical learning problem is upper bounded by the value of the corresponding sequential prediction problem (see Eq. (5.14)) against an oblivious (and, hence, non-oblivious) adversary:

$$\mathcal{V}^{iid}(\mathcal{F}, n) \le \mathcal{V}^{obliv}(\mathcal{F}, n)$$

Proof. Just as in the proof of Lemma 4.1, suppose the sequence of Nature's moves is i.i.d. and take an expectation on both sides of the regret expression:

$$\mathbb{E}\left\{ \frac{1}{n} \sum_{t=1}^{n} \ell(f_t, Z_t) \right\} \le \mathbb{E}\left\{ \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell(f, Z_t) \right\} + \mathbb{E} R_n \qquad (28.5)$$

---

[1]This assumption can be relaxed by considering mixed strategies.

Observe that by Jensen's inequality,

$$\mathbb{E}\boldsymbol{L}(\bar{f}) = \mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{\ell}(\bar{f}, Z) \mid Z_{1:n}\right\}\right\} \leq \mathbb{E}\left\{\mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(f_t, Z) \,\middle|\, Z_{1:n}\right\}\right\}$$

$$= \mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left\{\boldsymbol{\ell}(f_t, Z_t) \mid Z_{1:t-1}\right\}\right\} = \mathbb{E}\left\{\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(f_t, Z_t)\right\}$$

and

$$\mathbb{E}\left\{\inf_{f \in \mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\ell}(f, Z_t)\right\} \leq \inf_{f \in \mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\boldsymbol{\ell}(f, Z_t)\right\} = \inf_{f \in \mathcal{F}}\mathbb{E}\boldsymbol{\ell}(f, Z) = \inf_{f \in \mathcal{F}}\boldsymbol{L}(f)$$

Rearranging the terms, the proof is completed. $\qquad\square$

Since in practice we often aim to minimize a convex loss, the above lemma allows us to harness the power of sequential algorithms developed earlier in the course: the average of the outputs of any such method can be used as a predictor under the i.i.d. assumption. Much of the attractiveness of using sequential regret minimization methods for i.i.d. learning is in the computational properties of such algorithms. Processing one example at a time in a stream-like fashion, the methods are often preferable for large scale problems.

We mention that the idea of computing the average $\bar{f}$ (called the average of the trajectory) is well-known in stochastic optimization.

29

Sequential Prediction: Better Bounds for Predictable Sequences

Within the framework of sequential prediction, we developed methods that guarantee a certain level of performance irrespective of the sequence being presented. While such "protection" against the worst case is often attractive, the bounds are naturally pessimistic. It is, therefore, desirable to develop algorithms that yield tighter bounds for "more regular" sequences, while still providing protection against worst-case sequences.

There are a number of ways to model "more regular" sequences. Let us start with the following definition. Fix a sequence of functions $M_t : \mathcal{X}^{t-1} \mapsto \mathcal{X}$, for each $t \in \{1,\ldots,n\} \triangleq [n]$. These functions define a predictable process

$$M_1, \; M_2(x_1), \; \ldots, \; M_n(x_1,\ldots,x_{n-1}) \; .$$

If, in fact, $x_t = M_t(x_1,\ldots,x_{t-1})$ for all t, one may view the sequence $\{x_t\}$ as a (noiseless) time series, or as an oblivious strategy of Nature. If we knew that the sequence given by Nature follows exactly this evolution, we should suffer no regret.

Suppose that we have a hunch that the actual sequence $x_t \approx M_t(x_1,\ldots,x_{t-1})$ will be "roughly" given by this predictable process. In other words, we suspect that

$$\text{sequence} = \text{predictable process} + \text{adversarial noise}$$

Can we use this fact to incur smaller regret if our suspicion is correct? Ideally, we would like to "pay" only for the unpredictable part of the sequence.

Once we know how to use the predictable process M_t to incur small regret, we would like to choose the best one from a family of predictable processes. We will

address this issue of learning M_t later in the lecture. However, we'd like to make it clear that the two problems are separate: (a) using a particular M_t as a prior knowledge about the sequence in order to incur small regret, and (b) learning the best M_t from a family of models.

Let us focus on the setting of online linear optimization, with $\ell(f,x) = \langle f,x\rangle$. For notational convenience, assume for a moment that $\mathcal{F}$ and $\mathcal{X}$ are dual unit balls.

In some sense, we would like to use $\{M_t\}$ as a "center" around which the actual sequence will be adversarially chosen. The key is the following observation, made in [46]. Let us go back to the symmetrization step in Equation (7.12), which we now state for the linear loss case:

$$\sup_{x_1,x_1'} \mathbb{E}_{\epsilon_1} \ldots \sup_{x_n,x_n'} \mathbb{E}_{\epsilon_n} \left\| \sum_{t=1}^{n} \epsilon_t \left(x_t' - x_t \right) \right\|_* \leq 2 \sup_{x_1} \mathbb{E}_{\epsilon_1} \ldots \sup_{x_n} \mathbb{E}_{\epsilon_n} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_*$$

If we instead only consider sequences such that at any time $t \in [n]$, x_t and x_t' have to be σ_t-close to the predictable process $M_t(x_1,\ldots,x_{t-1})$, we can add and subtract the "center" M_t on the left-hand side of the above equation and obtain tighter bounds **for free, irrespective of the form of** $M_t(x_1,\ldots,x_{t-1})$. To make this observation more precise, let

$$C_t = C_t(x_1,\ldots,x_{t-1}) = \left\{ x : \|x - M_t(x_1,\ldots,x_{t-1})\|_* \leq \sigma_t \right\}$$

be the set of allowed deviations from the predictable "trend". We then have a bound

$$\sup_{x_1,x_1' \in C_1} \mathbb{E}_{\epsilon_1} \ldots \sup_{x_n,x_n' \in C_n} \mathbb{E}_{\epsilon_n} \left\| \sum_{t=1}^{n} \epsilon_t \left(x_t' - M_t(x_1,\ldots,x_{t-1}) + M_t(x_1,\ldots,x_{t-1}) - x_t \right) \right\|_* \leq c\sqrt{\sum_{t=1}^{n} \sigma_t^2}$$

on the value of the game against such "constrained" sequences, where the constant c depends on the smoothness of the norm.

The development so far is a good example of how a purely theoretical observation can point to existence of better prediction methods. What is even more surprising, for most of the methods presented below, the individual σ_t's need not be known ahead of time except for their total sum $\sum_{t=1}^{n} \sigma_t^2$. The latter sum need not be known in advance either, thanks to the standard doubling trick, and one can obtain upper bounds of

$$\sum_{t=1}^{n} \langle f_t, x_t \rangle - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \langle f, x_t \rangle \leq c\sqrt{\sum_{t=1}^{n} \|x_t - M_t(x_1,\ldots,x_{t-1})\|_*^2}$$

on regret, for some problem-dependent constant c.

Let us now discuss several types of statistics M_t that could be of interest.

Example 24. Regret bounds in terms of

$$M_t(x_1,\ldots,x_{t-1}) = x_{t-1}$$

are known as *path length bounds* [46, 18]. Such bounds can be tighter than the pessimistic $O(\sqrt{n})$ bounds when the previous move of Nature is a good proxy for the next move.

Regret bounds in terms of

$$M_t(x_1,\ldots,x_{t-1}) = \frac{1}{t-1}\sum_{s=1}^{t-1} x_s$$

are known as *variance bounds* [17, 26, 27, 46]. One may also consider fading memory statistics

$$M_t(x_1,\ldots,x_{t-1}) = \sum_{s=1}^{t-1} \alpha_s x_s, \quad \sum_{s=1}^{t-1} \alpha_s = 1, \quad \alpha_s \geq 0$$

or even plug in an *auto-regressive model*.

If "phases" are expected in the data (e.g., stocks tend to go up in January), one may consider

$$M_t(x_1,\ldots,x_{t-1}) = x_{t-k}$$

for some phase length k. Alternatively, one may consider averaging of the past occurrences $n_j(t) \subset \{1,\ldots,t\}$ of the current phase j:

$$M_t(x_1,\ldots,x_{t-1}) = \sum_{s \in n_t} \alpha_s x_s.$$

29.1 Full Information Methods

We now exhibit a Mirror Descent type method which can be seen as a generalization of the recent algorithm of [18]. Let $\mathcal{R}$ be a 1-strongly convex function with respect to a norm $\|\cdot\|$, and let $D_{\mathcal{R}}(\cdot,\cdot)$ denote the Bregman divergence with respect to $\mathcal{R}$. Let $\nabla\mathcal{R}^*$ be the inverse of the gradient mapping $\nabla\mathcal{R}$. Let $\|\cdot\|_*$ be the norm dual to $\|\cdot\|$. We do not require $\mathcal{F}$ and $\mathcal{X}$ to be unit dual balls.

29.1 Full Information Methods

Optimistic Mirror Descent Algorithm (equivalent form)

Input: $\mathcal{R}$ 1-strongly convex w.r.t. $\|\cdot\|$, learning rate $\eta > 0$

Initialize $f_1 = g_1 = \operatorname{argmin}_g \mathcal{R}(g)$

At $t = 1, \ldots, n$, predict f_t and update

$$g_{t+1} = \operatorname*{argmin}_{g \in \mathcal{F}} \eta \langle g, x_t \rangle + D_{\mathcal{R}}(g, g_{t+1})$$

$$f_{t+1} = \operatorname*{argmin}_{f \in \mathcal{F}} \eta \langle f, M_{t+1} \rangle + D_{\mathcal{R}}(f, g_{t+1})$$

Lemma 29.1. *Let $\mathcal{F}$ be a convex set in a Banach space $\mathcal{B}$ and $\mathcal{X}$ be a convex set in the dual space $\mathcal{B}^*$. Let $\mathcal{R} : \mathcal{B} \mapsto \mathbb{R}$ be a 1-strongly convex function on $\mathcal{F}$ with respect to some norm $\|\cdot\|$. For any strategy of Nature, the Optimistic Mirror Descent Algorithm (with or without projection for g_{t+1}) yields, for any $f^* \in \mathcal{F}$,*

$$\sum_{t=1}^{n} \langle f_t, x_t \rangle - \sum_{t=1}^{n} \langle f^*, x_t \rangle \le \eta^{-1} R_{\max}^2 + \frac{\eta}{2} \sum_{t=1}^{n} \|x_t - M_t\|_*^2$$

where $R_{\max}^2 = \max_{f \in \mathcal{F}} \mathcal{R}(f) - \min_{f \in \mathcal{F}} \mathcal{R}(f)$.

Proof of Lemma 29.1. For any $f^* \in \mathcal{F}$,

$$\langle f_t - f^*, x_t \rangle = \langle f_t - g_{t+1}, x_t - M_t \rangle + \langle f_t - g_{t+1}, M_t \rangle + \langle g_{t+1} - f^*, x_t \rangle \qquad (29.1)$$

First observe that

$$\langle f_t - g_{t+1}, x_t - M_t \rangle \le \| f_t - g_{t+1} \| \, \| x_t - M_t \|_* \le \frac{\eta}{2} \| x_t - M_t \|_*^2 + \frac{1}{2\eta} \| f_t - g_{t+1} \|^2 . \qquad (29.2)$$

On the other hand, any update of the form $a^* = \operatorname{argmin}_{a \in A} \langle a, x \rangle + D_{\mathcal{R}}(a, c)$ satisfies (see e.g. [6, 42])

$$\langle a^* - d, x \rangle \le \langle d - a^*, \nabla \mathcal{R}(a^*) - \nabla \mathcal{R}(c) \rangle = D_{\mathcal{R}}(d, c) - D_{\mathcal{R}}(d, a^*) - D_{\mathcal{R}}(a^*, c) . \qquad (29.3)$$

This yields

$$\langle f_t - g_{t+1}, M_t \rangle \le \frac{1}{\eta} \left(D_{\mathcal{R}}(g_{t+1}, g_t) - D_{\mathcal{R}}(g_{t+1}, f_t) - D_{\mathcal{R}}(f_t, g_t) \right) . \qquad (29.4)$$

Next, note that by the form of update for g_{t+1},

$$\langle g_{t+1} - f^*, x_t \rangle = \frac{1}{\eta} \langle g_{t+1} - f^*, \nabla \mathcal{R}(g_t) - \nabla \mathcal{R}(g_{t+1}) \rangle$$

$$= \frac{1}{\eta} \left(D_\mathcal{R}(f^*, g_t) - D_\mathcal{R}(f^*, g_{t+1}) - D_\mathcal{R}(g_{t+1}, g_t) \right), \qquad (29.5)$$

and the same inequality holds by (29.3) if g_{t+1} is defined as in (**??**) with a projection. Using Equations (29.2), (29.5) and (29.4) in Equation (29.1) we conclude that

$$\langle f_t - f^*, x_t \rangle \leq \frac{\eta}{2} \|x_t - M_t\|_*^2 + \frac{1}{2\eta} \|f_t - g_{t+1}\|^2$$

$$+ \frac{1}{\eta} \left(D_\mathcal{R}(g_{t+1}, g_t) - D_\mathcal{R}(g_{t+1}, f_t) - D_\mathcal{R}(f_t, g_t) \right)$$

$$+ \frac{1}{\eta} \left(D_\mathcal{R}(f^*, g_t) - D_\mathcal{R}(f^*, g_{t+1}) - D_\mathcal{R}(g_{t+1}, g_t) \right)$$

$$\leq \frac{\eta}{2} \|x_t - M_t\|_*^2 + \frac{1}{2\eta} \|f_t - g_{t+1}\|^2 + \frac{1}{\eta} \left(D_\mathcal{R}(f^*, g_t) - D_\mathcal{R}(f^*, g_{t+1}) - D_\mathcal{R}(g_{t+1}, f_t) \right)$$

By strong convexity of $\mathcal{R}$, $D_\mathcal{R}(g_{t+1}, f_t) \geq \frac{1}{2} \|g_{t+1} - f_t\|^2$ and thus

$$\langle f_t - f^*, x_t \rangle \leq \frac{\eta}{2} \|x_t - M_t\|_*^2 + \frac{1}{\eta} \left(D_\mathcal{R}(f^*, g_t) - D_\mathcal{R}(f^*, g_{t+1}) \right)$$

Summing over $t = 1, \ldots, n$ yields, for any $f^* \in \mathcal{F}$,

$$\sum_{t=1}^n \langle f_t - f^*, x_t \rangle \leq \frac{\eta}{2} \sum_{t=1}^n \|x_t - M_t\|_*^2 + \frac{R_{\max}^2}{\eta}$$

where $R_{\max}^2 = \max_{f \in \mathcal{F}} \mathcal{R}(f) - \min_{f \in \mathcal{F}} \mathcal{R}(f)$. Choosing $\eta = \frac{\sqrt{2} R_{\max}}{\sqrt{\sum_{t=1}^n \|x_t - M_t\|_*^2}}$ balances the two terms. $\qquad \square$

We remark that the sum $\sum_{t=1}^n \|x_t - M_t\|_*^2$ need not be known in advance in order to set η, as the usual doubling trick can be employed. Moreover, the results carry over to online convex optimization, where x_t's are now gradients at the points chosen by the learner. Last but not least, notice that if the sequence is not following the trend M_t as we hoped it would, we still obtain the same bounds as for the Mirror Descent algorithm. In some sense, we got something for nothing!

29.2 Learning The Predictable Processes

So far we have seen that the learner can pick any arbitrary predictable process $(M_t)_{t \geq 1}$ and suffer a regret of at most $O\left(\sqrt{\sum_{t=1}^n \|x_t - M_t\|_*^2}\right)$. Now if the predictable process we chose was good then our regret will be low. This raises the question as to how the learner can choose a good predictable process $(M_t)_{t \geq 1}$? Is it possible to learn it online as we go, and if so, what does it mean to learn?

To formalize the concept of learning the predictable process, let us consider the case where we have a set Π indexing a set of predictable processes (strategies) we are interested in. That is each $\pi \in \Pi$ corresponds to predictable process given by $(M_t^\pi)_{t \geq 1}$. Now if we had an oracle which in the start of the game told us which $\pi^* \in \Pi$ predicts the sequence optimally (in hindsight) then we could use the predictable process given by $(M_t^{\pi^*})_{t \geq 1}$ and enjoy a regret bound of $\sqrt{\inf_{\pi \in \Pi} \sum_{t=1}^n \|x_t - M_t^\pi\|_*^2}$. However we cannot expect to know which $\pi \in \Pi$ is the optimal one at the start of the game. In this scenario one would like to learn a predictable process that in turn can be used with algorithms proposed thus far to get a regret bound comparable regret bound one could have obtained knowing the optimal $\pi^* \in \Pi$.

To motivate this setting better let us consider an example. Say there are n stock options we can choose to invest in. On each day t, associated with each stock option one has a loss/payoff that occurs upon investing in a single share of that stock. Our goal in the long run is to have a low regret with respect to the single best stock in hindsight. Up to this point, the problem just corresponds to the simple experts setting where each of the n stocks is one expert and on each day we split our investment according to a probability distribution over the n options. However now additionally we allow the learner/investor access to *prediction models* from the set Π. These could be human strategists making forecasts, or outcomes of some hedge-fund model. At each time step the learner can query prediction made by each $\pi \in \Pi$ as to what the loss on the n stocks would be on that day. Now we would like to have a regret comparable to the regret we can achieve knowing the best model $\pi^* \in \Pi$ that in hind-sight predicted the losses of each stock optimally. We shall now see how to achieve this.

Optimistic Mirror Descent Algorithm with Learning the Predictable Process

Input: $\mathcal{R}$ 1-strongly convex w.r.t. $\|\cdot\|$, learning rate $\eta > 0$

Initialize $f_1 = g_1 = \arg\min_g \mathcal{R}(g)$ and initialize $q_1 \in \Delta(\Pi)$ as, $\forall \pi \in \Pi, q_1(\pi) = \frac{1}{|\Pi|}$

Set $M_1 = \sum_{\pi \in \Pi} q_1(\pi) M_1^\pi$

At $t = 1, \ldots, n$, predict f_t and update

$$\forall \pi \in \Pi, \ q_{t+1}(\pi) \propto q_t(\pi) e^{-\|M_t^\pi - x_t\|_*^2} \ \text{ and } \ M_{t+1} = \sum_{\pi \in \Pi} q_{t+1}(\pi) M_{t+1}^\pi$$

$$g_{t+1} = \nabla \mathcal{R}^* \left(\nabla \mathcal{R}(g_t) - \eta x_t \right)$$

$$f_{t+1} = \underset{f \in \mathcal{F}}{\arg\min} \ \eta \left\langle f, M_{t+1} \right\rangle + D_{\mathcal{R}}(f, g_{t+1})$$

Lemma 29.2. *Let $\mathcal{F}$ be a convex subset of a unit ball in a Banach space $\mathcal{B}$ and $\mathcal{X}$ be a convex subset of the dual unit ball. Let $\mathcal{R} : \mathcal{B} \mapsto \mathbb{R}$ be a 1-strongly convex function on $\mathcal{F}$ with respect to some norm $\|\cdot\|$. For any strategy of Nature, the Optimistic Mirror Descent Algorithm yields, for any $f^* \in \mathcal{F}$,*

$$\sum_{t=1}^{n} \left\langle f_t, x_t \right\rangle - \sum_{t=1}^{n} \left\langle f^*, x_t \right\rangle \leq \eta^{-1} R_{\max}^2 + 3.2 \eta \left(\inf_{\pi \in \Pi} \sum_{t=1}^{n} \|x_t - M_t^\pi\|_*^2 + \log |\Pi| \right)$$

where $R_{\max}^2 = \max_{f \in \mathcal{F}} \mathcal{R}(f) - \min_{f \in \mathcal{F}} \mathcal{R}(f)$.

Once again, let us discuss what makes this setting different from the usual setting of experts: the forecast given by the prediction models is in the form of a vector, one for each stock. If we treat each prediction model as an expert with the loss $\|x_t - M_t^\pi\|_*^2$, the experts algorithm would guarantee that we achieve the best cumulative loss of this kind. However, this is not the object of interest to us, as we are after the best allocation of our money among the stocks, as measured by $\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \left\langle f, x_t \right\rangle$.

Proof. First note that by Lemma 29.1 we have that for the M_t chosen in the algo-

rithm,

$$\sum_{t=1}^{n} \langle f_t, x_t \rangle - \sum_{t=1}^{n} \langle f^*, x_t \rangle \leq \eta^{-1} R_{\max}^2 + \frac{\eta}{2} \sum_{t=1}^{n} \| x_t - M_t \|_*^2$$

$$\leq \eta^{-1} R_{\max}^2 + \frac{\eta}{2} \sum_{t=1}^{n} \sum_{\pi \in \Pi} q_t(\pi) \| x_t - M_t^{\pi} \|_*^2 \qquad \text{(Jensen's Inequality)}$$

$$\leq \eta^{-1} R_{\max}^2 + \frac{\eta}{2} \left(\frac{4e}{e-1} \right) \left(\inf_{\pi \in \Pi} \sum_{t=1}^{n} \| x_t - M_t^{\pi} \|_*^2 + \log |\Pi| \right)$$

where the last step is due to Corollary 2.3 of [16]. Indeed, the updates for q_t's are exactly the experts algorithm with point-wise loss at each round t for expert $\pi \in \Pi$ given by $\| M_t^{\pi} - x_t \|_*^2$. Also as each $M_t^{\pi} \in \mathcal{X}$ the unit ball of dual norm, we can conclude that $\| M_t^{\pi} - x_t \|_*^2 \leq 4$ which is why we have a scaling by factor 4. Simplifying leads to the bound in the lemma. $\qquad \square$

Notice that the general idea used above is to get M_t's for the algorithm by running another (secondary) regret minimizing strategy where loss per round is simply $\| M_t - x_t \|_*^2$ and regret is considered with respect to the best $\pi \in \Pi$. That is, regret of the secondary regret minimizing game is given by

$$\sum_{t=1}^{n} \| x_t - M_t \|_*^2 - \inf_{\pi \in \Pi} \sum_{t=1}^{n} \| x_t - M_t^{\pi} \|_*^2$$

In general, the experts algorithm for minimizing secondary regret can be replaced by any other online learning algorithm.

29.3   Follow the Perturbed Leader Method

29.4   A General Framework of Stochastic, Smoothed, and Constrained Adversaries

30

Sequential Prediction: Competing With Strategies

One of the most frequent criticisms of the regret definition is the fact that the comparator term is a static decision is hindsight, not able to change its predictions as the sequence evolves. What good is showing small regret if we only perform as well as the best single action for the whole sequence?

We now show that one can obtain regret bounds with respect to a possibly infinite set of *strategies* which look at the actual sequence and produce a prediction. In fact, this is not an increase in generality of our results so far, but rather the opposite. Observe that in the setting of supervised learning, the side information x_t is presented at the beginning of the round. This information is treated as a worst-case move in our framework, but can instead be restricted to contain the history so far. In this case, an element $f \in \mathcal{F}$ can be viewed as a strategy which produces a prediction $f(x_t)$ based on the history. At the end of the day, we will be competing with the best strategy $f \in \mathcal{F}$ for the given sequence of $(x_1, y_1), \ldots, (x_n, y_n)$.

So, if the setting of supervised learning subsumes competing with strategies, can we detail the bounds for the latter case? The key is that the covering numbers, the combinatorial parameters, and the other relevant complexities are now defined not for the worst-case tree $\mathbf{x}$ but for the worst tree with some "history" structure. This makes the problem easier. In fact, we will easily derive bounds for regret defined in terms of a set of bounded-lookback strategies, or strategies limited in some other way (of course, we cannot compete with absolutely all possible strategies). Further, we will show that we can achieve the correct rate for regret against the set of all monotonic strategies, a problem that was considered in [15].

We do not know whether such a result is possible without the tools developed in this course.

Let us fix some notation. At each round, we choose an action $f_t \in \mathcal{F}$, Nature responds with z_t, and the cost associated with the two decisions is $\ell(f_t, z_t)$. Let

$$\Pi = \left\{\pi = (\pi_t)_{t=1}^n \; : \; \pi_t : \mathcal{Z}^{t-1} \to \mathcal{F}\right\}$$

be a set of strategies π, each consisting of a sequence of mappings π_t from history to the next action. We may then formulate the value of the game with regret defined in terms of these strategies.

$$\mathcal{V}(\Pi) = \inf_{q_1} \sup_{z_1} \mathop{\mathbb{E}}_{f_1 \sim q_1} \ldots \inf_{q_n} \sup_{z_n} \mathop{\mathbb{E}}_{f_n \sim q_n} \left[\frac{1}{n} \sum_{t=1}^n \ell(f_t, z_t) - \inf_{\pi \in \Pi} \frac{1}{n} \sum_{t=1}^n \ell(\pi_t(z_{1:t-1}), z_t)\right]$$

30.1 Bounding the Value with History Trees

To proceed, we need a definition of a history tree.

Definition 30.1. Let $\mathcal{H} = \cup_{t \geq 0} \mathcal{Z}^t$ be the space of histories. We say a tree $\mathbf{h}$ is a *history tree* if it is $\mathcal{H}$-valued and has the following consistency properties for any t:

- $\mathbf{h}_i : \{\pm 1\}^{i-1} \mapsto \mathcal{Z}^{i-1}$

- for any ϵ, if $\mathbf{h}_t(\epsilon_1, \ldots, \epsilon_{t-2}, +1) = (z_1, \ldots, z_{t-1})$ and $\mathbf{h}_t(\epsilon_1, \ldots, \epsilon_{t-2}, -1) = (z_1', \ldots, z_{t-1}')$, then $z_s = z_s'$ for all $1 \leq s \leq t-2$.

As an example, a valid history tree would have

$\mathbf{h}_1 = \emptyset, \mathbf{h}_2(1) = z_1, \mathbf{h}_3(1, -1) = (z_1, z_2), \mathbf{h}_3(1, 1) = (z_1, z_3), \mathbf{h}_2(-1) = z_4, \mathbf{h}_3(-1, -1) = (z_4, z_5), \mathbf{h}_3(-1, 1) = (z_4, z_6).$

Theorem 30.2. *The value of the prediction problem with a set Π of strategies is upper bounded as*

$$\mathcal{V}(\Pi) \leq 2 \sup_{\mathbf{w}, \mathbf{z}} \mathbb{E}_\epsilon \sup_{\pi \in \Pi} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(\pi_t(\mathbf{w}_1(\epsilon), \ldots, \mathbf{w}_{t-1}(\epsilon)), \mathbf{z}_t(\epsilon))\right]$$

where the supremum is over two $\mathcal{Z}$-valued trees $\mathbf{z}$ and $\mathbf{w}$ of depth n. Equivalently, we may rewrite the upper bound as

$$\mathcal{V}(\Pi) \leq 2 \sup_{\mathbf{g}} \mathbb{E}_\epsilon \sup_{\pi \in \Pi} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_t \ell'(\pi, \mathbf{g}_t(\epsilon))\right] \triangleq 2\mathcal{R}^{seq}(\Pi)$$

for the new loss function $\ell'(\pi, (\mathbf{h}_t, \mathbf{z}_t)(\epsilon)) = \ell(\pi_t(\mathbf{h}_t(\epsilon)), \mathbf{z}_t(\epsilon))$ where $\mathbf{g}$ ranges over all history-outcome *trees* $(\mathbf{h}, \mathbf{z})$, *with $\mathbf{z}$ being an arbitrary $\mathcal{Z}$-valued tree, and $\mathbf{h}$ being an arbitrary history tree (not necessarily consistent with $\mathbf{z}$).*

Proof. The proof of Section 7.1.2 goes through and we obtain

$$
n \cdot \mathcal{V}(\Pi) = \left\langle\!\!\left\langle \sup_{p_t} \mathbb{E}_{z_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\pi \in \Pi} \left[\sum_{t=1}^{n} \inf_{f_t \in \mathcal{F}} \mathbb{E}_{z'_t} \ell(f_t, z'_t) - \ell(\pi_t(z_{1:t-1}), z_t) \right]
$$

$$
\leq \left\langle\!\!\left\langle \sup_{p_t} \mathbb{E}_{z_t \sim p_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\pi \in \Pi} \left[\sum_{t=1}^{n} \mathbb{E}_{z'_t} \ell(\pi_t(z_{1:t-1}), z'_t) - \ell(\pi_t(z_{1:t-1}), z_t) \right]
$$

$$
\leq \left\langle\!\!\left\langle \sup_{p_t} \mathbb{E}_{z_t, z'_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\pi \in \Pi} \left[\sum_{t=1}^{n} \ell(\pi_t(z_{1:t-1}), z'_t) - \ell(\pi_t(z_{1:t-1}), z_t) \right]
$$

Define the "selector function" $\chi : \mathcal{Z} \times \mathcal{Z} \times \{\pm 1\} \mapsto \mathcal{Z}$ by

$$
\chi(z, z', \epsilon) = \left\{ \begin{array}{ll} z' & \text{if } \epsilon = -1 \\ z & \text{if } \epsilon = 1 \end{array} \right.
$$

When z_t and z'_t are understood from the context, we will use the shorthand $\chi_t(\epsilon) := \chi(z_t, z'_t, \epsilon)$. In other words, χ_t selects between z_t and z'_t depending on the sign of ϵ.

With the notation of the selector function, we may write

$$
\left\langle\!\!\left\langle \sup_{p_t} \mathbb{E}_{z_t, z'_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\pi \in \Pi} \left[\sum_{t=1}^{n} \ell(\pi_t(z_{1:t-1}), z'_t) - \ell(\pi_t(z_{1:t-1}), z_t) \right]
$$

$$
= \left\langle\!\!\left\langle \sup_{p_t} \mathbb{E}_{z_t, z'_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\pi \in \Pi} \left[\sum_{t=1}^{n} \epsilon_t \big(\ell(\pi_t(\chi_1(z_1, z'_1, \epsilon_1), \ldots, \chi_{t-1}(z_{t-1}, z'_{t-1}, \epsilon_{t-1})), z'_t) \right.
$$

$$
\left. - \ell(\pi_t(\chi_1(z_1, z'_1, \epsilon_1), \ldots, \chi_{t-1}(z_{t-1}, z'_{t-1}, \epsilon_{t-1})), z_t)) \big) \right]
$$

To be convinced of this equality, we should think of $\epsilon_t = -1$ as "z_t and z'_t getting renamed". The process of renaming these random variables has two effects: the sign on the t-th term gets flipped, and all the conditioning that was previously done on z_t now has z'_t instead. This is exactly accomplished by multiplying the t-th term by ϵ_t and introducing the selector throughout the history.

We now split the above terms into two, yielding an upper bound of

$$
\left\langle\!\!\left\langle \sup_{p_t} \mathbb{E}_{z_t, z'_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\pi \in \Pi} \left[\sum_{t=1}^{n} \epsilon_t \ell(\pi_t(\chi_1(z_1, z'_1, \epsilon_1), \ldots, \chi_{t-1}(z_{t-1}, z'_{t-1}, \epsilon_{t-1})), z'_t) \right]
$$

$$
+ \left\langle\!\!\left\langle \sup_{p_t} \mathbb{E}_{z_t, z'_t} \mathbb{E}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^{n} \sup_{\pi \in \Pi} \left[\sum_{t=1}^{n} -\epsilon_t \ell(\pi_t(\chi_1(z_1, z'_1, \epsilon_1), \ldots, \chi_{t-1}(z_{t-1}, z'_{t-1}, \epsilon_{t-1})), z_t)) \right]
$$

Now the claim is that the second term is equal to the first. Let us replace ϵ_t with $-\epsilon_t$ for all t:

$$\left\langle\!\!\left\langle \sup_{p_t} \mathop{\mathbb{E}}_{z_t,z_t'} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^n \sup_{\pi\in\Pi} \left[\sum_{t=1}^n -\epsilon_t \ell(\pi_t(\chi_1(z_1,z_1',\epsilon_1),\ldots,\chi_{t-1}(z_{t-1},z_{t-1}',\epsilon_{t-1})),z_t)) \right]$$

$$= \left\langle\!\!\left\langle \sup_{p_t} \mathop{\mathbb{E}}_{z_t,z_t'} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^n \sup_{\pi\in\Pi} \left[\sum_{t=1}^n \epsilon_t \ell(\pi_t(\chi_1(z_1',z_1,\epsilon_1),\ldots,\chi_{t-1}(z_{t-1}',z_{t-1},\epsilon_{t-1})),z_t)) \right]$$

Renaming z_t' and z_t concludes the claim. Therefore,

$$n\cdot\mathcal{V}(\Pi) \le 2\left\langle\!\!\left\langle \sup_{p_t} \mathop{\mathbb{E}}_{z_t,z_t'} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^n \sup_{\pi\in\Pi} \left[\sum_{t=1}^n \epsilon_t \ell(\pi_t(\chi_1(z_1,z_1',\epsilon_1),\ldots,\chi_{t-1}(z_{t-1},z_{t-1}',\epsilon_{t-1})),z_t')) \right]$$

$$\le 2\left\langle\!\!\left\langle \sup_{z_t,z_t'} \mathop{\mathbb{E}}_{\epsilon_t} \right\rangle\!\!\right\rangle_{t=1}^n \sup_{\pi\in\Pi} \left[\sum_{t=1}^n \epsilon_t \ell(\pi_t(\chi_1(z_1,z_1',\epsilon_1),\ldots,\chi_{t-1}(z_{t-1},z_{t-1}',\epsilon_{t-1})),z_t')) \right]$$

$$= 2\sup_{\mathbf{z},\mathbf{z}'} \mathbb{E}_\epsilon \sup_{\pi\in\Pi} \left[\sum_{t=1}^n \epsilon_t \ell(\pi_t(\chi_1(\mathbf{z}_1,\mathbf{z}_1',\epsilon_1),\ldots,\chi_{t-1}(\mathbf{z}_{t-1}(\epsilon),\mathbf{z}_{t-1}'(\epsilon),\epsilon_{t-1})),\mathbf{z}_t'(\epsilon)) \right]$$

We now replace the dependence of π_t on two trees and a selector between them by a dependence on only one tree, resulting in an upper bound of

$$2\sup_{\mathbf{w},\mathbf{z}'} \mathbb{E}_\epsilon \sup_{\pi\in\Pi} \left[\sum_{t=1}^n \epsilon_t \ell(\pi_t(\mathbf{w}_1,\ldots,\mathbf{w}_{t-1}(\epsilon)),\mathbf{z}_t'(\epsilon)) \right]$$

While this might be a loose upper bound, we do not expect this step to be too bad: for sequences ϵ with the majority of +1's, the arguments to π_t anyway come from the tree $\mathbf{z}$, unrelated to $\mathbf{z}'$.

We may now re-write the upper bound as

$$2\sup_{\mathbf{h},\mathbf{z}} \mathbb{E}_\epsilon \sup_{\pi\in\Pi} \left[\sum_{t=1}^n \epsilon_t \ell(\pi_t(\mathbf{h}_t(\epsilon)),\mathbf{z}_t(\epsilon)) \right]$$

where $\mathbf{h}$ ranges over history trees and $\mathbf{z}$ ranges over $\mathcal{Z}$-valued trees. For the new loss function $\ell'(\pi,(\mathbf{h}_t,\mathbf{z}_t)(\epsilon)) = \ell(\pi_t(\mathbf{h}_t(\epsilon)),\mathbf{z}_t(\epsilon))$ we can rewrite the (normalized by n) bound as

$$2\sup_{\mathbf{g}} \mathbb{E}_\epsilon \sup_{\pi\in\Pi} \left[\frac{1}{n}\sum_{t=1}^n \epsilon_t \ell'(\pi,\mathbf{g}_t(\epsilon)) \right] \triangleq 2\mathcal{R}^{seq}(\Pi) \,.$$

$\square$

We remark that in the case $\pi \in \Pi$ are constant history-independent strategies $\pi_1^f = \ldots = \pi_n^f = f \in \mathcal{F}$, we recover the sequential Rademacher complexity. In fact, for all the notions we present below, taking Π to be the set of all constant strategies should recover the notions introduced earlier.

Example 25. Theorem 30.2 has the following immediate implication. Let Π^k be a set of strategies which only look back at k past outcomes to determine the next move. The bound then becomes

$$\mathcal{V}(\Pi^k) \leq 2 \sup_{\mathbf{w},\mathbf{z}} \mathbb{E}_\epsilon \sup_{\pi \in \Pi^k} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_t \boldsymbol{\ell}(\pi_t(\mathbf{w}_{t-k}(\epsilon),\ldots,\mathbf{w}_{t-1}(\epsilon)), \mathbf{z}_t(\epsilon)) \right]$$

Now, suppose that $\mathcal{Z} = \mathcal{F}$ is a finite set, of cardinality s. Then there are effectively s^{s^k} strategies π. The bound on the sequential Rademacher complexity will then scale as $\sqrt{\frac{2s^k \log s}{n}}$. However, we could have used the Exponential Weights algorithm here, treating each π as an expert, arriving at the same bound on regret. The next example presents a situation where one cannot simply use the exponential weights algorithm.

Example 26. Let Π^k be the set of strategies which only depend on the past k moves, and let us parametrize the strategies by a set $\Theta \subset \mathbb{R}^k$ as:

$$\pi_t^\theta(z_1,\ldots,z_{t-1}) = \pi_t^\theta(z_{t-k},\ldots,z_{t-1}) = \sum_{i=1}^k \theta_i z_{t-i} \quad \theta = (\theta_1,\ldots,\theta_k) \in \Theta$$

Note that in this example the parameters θ of a given strategy π^θ are fixed throughout time. We may view this as an auto-regressive strategy. For illustration purposes, suppose $\mathcal{Z} = \mathcal{F} \subset \mathbb{R}^d$ are ℓ_2 unit balls, the loss is $\boldsymbol{\ell}(f,z) = \langle f,z \rangle$, and $\Theta \subset \mathbb{R}^k$ is also a unit ℓ_2 ball. Then

$$\begin{aligned}
\mathcal{R}(\Pi) &= \sup_{\mathbf{w},\mathbf{z}} \mathbb{E}_\epsilon \sup_{\theta \in \Theta} \left[\sum_{t=1}^n \epsilon_t \left\langle \pi^\theta(\mathbf{w}_{t-k}(\epsilon),\ldots,\mathbf{w}_{t-1}(\epsilon)), \mathbf{z}_t(\epsilon) \right\rangle \right] \\
&= \sup_{\mathbf{w},\mathbf{z}} \mathbb{E}_\epsilon \sup_{\theta \in \Theta} \left[\sum_{t=1}^n \epsilon_t \mathbf{z}_t(\epsilon)^\top [\mathbf{w}_{t-k}(\epsilon),\ldots,\mathbf{w}_{t-1}(\epsilon)] \cdot \theta \right] \\
&= \sup_{\mathbf{w},\mathbf{z}} \mathbb{E}_\epsilon \left\| \sum_{t=1}^n \epsilon_t \mathbf{z}_t(\epsilon)^\top [\mathbf{w}_{t-k}(\epsilon),\ldots,\mathbf{w}_{t-1}(\epsilon)] \right\| \\
&\leq \sqrt{kn}
\end{aligned}$$

This bound against all strategies parametrized by Θ is achieved by the gradient-descent method. However, one can see that it is possible to consider slowly-changing

strategies, and one may arrive at an upper bound on the value in a non-constructive way.

Example 27. As another example, consider the binary prediction problem, $\mathcal{Z} = \{0, 1\}$, and the indicator loss. Suppose for now, strategies have potentially dependence on the full history. Then,

$$
\sup_{\mathbf{w},\mathbf{z}} \mathbb{E}_\epsilon \sup_{\pi \in \Pi^k} \left[\sum_{t=1}^n \epsilon_t \mathbf{I}\{\pi_t(\mathbf{w}_1(\epsilon), \ldots, \mathbf{w}_{t-1}(\epsilon)) \neq \mathbf{z}_t(\epsilon)\} \right]
$$

$$
= \sup_{\mathbf{w},\mathbf{z}} \mathbb{E}_\epsilon \sup_{\pi \in \Pi^k} \left[\sum_{t=1}^n \epsilon_t \big(\pi_t(\mathbf{w}_1(\epsilon), \ldots, \mathbf{w}_{t-1}(\epsilon))(1 - 2\mathbf{z}_t(\epsilon)) + \mathbf{z}_t(\epsilon) \big) \right]
$$

$$
= \sup_{\mathbf{w}} \mathbb{E}_\epsilon \sup_{\pi \in \Pi^k} \left[\sum_{t=1}^n \epsilon_t \pi_t(\mathbf{w}_1(\epsilon), \ldots, \mathbf{w}_{t-1}(\epsilon)) \right]
$$

30.2 Static Experts

It is now easy to recover the results of [15] for static experts. These are experts that do not depend on the history! That is, each strategy π is a predetermined sequence of outcomes, and we may therefore associate each π with a vector in $\mathcal{Z}^n$. Since there is not dependence on the history, the sequential Rademacher complexity of Π simplifies:

$$
\mathcal{V}(\Pi) \leq 2\mathcal{R}^{seq}(\Pi) = 2 \sup_{\mathbf{g}} \mathbb{E}_\epsilon \sup_{\pi \in \Pi} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_t \ell'(\pi, \mathbf{g}_t(\epsilon)) \right] = 2 \sup_{\mathbf{z}} \mathbb{E}_\epsilon \sup_{\pi \in \Pi} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(\pi_t, \mathbf{z}_t(\epsilon)) \right]
$$

Suppose $\mathcal{Z} = \{0, 1\}$, the loss is the indicator of a mistake, and $\pi \in \mathcal{Z}^n$ for each $\pi \in \Pi$. Then we may erase the tree $\mathbf{z}$ as in the above example:

$$
\mathcal{V}(\Pi) \leq 2\mathbb{E}_\epsilon \sup_{\pi \in \Pi} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_t \pi_t \right] = \mathcal{R}^{iid}(\Pi)
$$

which is simply the classical i.i.d. Rademacher averages. One may, of course, further upper bound this complexity by covering numbers, as done in the earlier part of the course. This example has been worked out in [15], and it was shown that the minimax value has a closed form. The authors were intrigued by the fact that empirical process theory came into play in the problem of sequential prediction. In fact, the development of the general theory we presented was also largely influenced by this surprising fact. It is now clear how the classical empirical process

theory arises due to the static nature of experts, yet in the non-static case a more general theory is required.

After introducing the combinatorial parameters and covering numbers which are specific to history trees, we will turn to the case of monotonic experts and resolve an open issue left by [15].

30.3   Covering Numbers and Combinatorial Parameters

We may now define covering numbers, combinatorial parameters, and other measures of complexity for the set Π of strategies over the history-outcome trees **g**. If we assume that the loss function is Lipschitz, then we may only need to define covering numbers on history trees **h**. The development is a straightforward modification of the notions we developed earlier, where we replace "any **x**" with a "history tree" **h**.

Definition 30.3. A set V of $\mathbb{R}$-valued history trees is a α-*cover* (with respect to ℓ_p) of a set of strategies Π on an $\mathcal{X}$-valued history tree **h** if

$$\forall \pi \in \Pi, \ \forall \epsilon \in \{\pm 1\}^n, \ \exists \mathbf{v} \in V \quad \text{s.t.} \quad \left(\frac{1}{n} \sum_{t=1}^{n} |\pi_t(\mathbf{h}_t(\epsilon)) - \mathbf{v}_t(\epsilon)|^p \right)^{1/p} \leq \alpha . \tag{30.1}$$

An α-*covering number* is defined as

$$\mathcal{N}_p(\Pi, \mathbf{h}, \alpha) = \min \left\{ \text{card}(V) \ : \ V \text{ is an } \alpha\text{-cover} \right\}.$$

If Π is a set of all constant strategies π^f with $\pi_1^f = \ldots = \pi_n^f = f \in \mathcal{F}$, we recover the notion of a covering number defined in Section 14.1.

For any history tree **h**, sequential Rademacher averages of a class of strategies Π, where each $\pi \in \Pi$ is a sequence $(\pi_t)_{t=1}^n$ with $\pi_t : \mathcal{Z}^{t-1} \mapsto \mathbb{R}$ satisfy

$$\mathcal{R}^{seq}(\Pi, \mathbf{h}) \leq \inf_{\alpha \geq 0} \left\{ \alpha + \sqrt{\frac{2 \log \mathcal{N}_1(\Pi, \alpha, \mathbf{h})}{n}} \right\}$$

and the Dudley entropy integral type bound also holds:

$$\mathcal{R}^{seq}(\Pi, \mathbf{h}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_2(\Pi, \mathbf{h}, \delta)} \, d\delta \right\} \tag{30.2}$$

For completeness, let us define history-based fat-shattering dimension.

Definition 30.4. A history tree **h** of depth n is α-*shattered* by a set of strategies Π, if there exists an $\mathbb{R}$-valued tree **s** of depth n such that

$$\forall \epsilon \in \{\pm 1\}^n, \ \exists \pi \in \Pi \ \text{ s.t. }, \ \epsilon_t(\pi_t(\mathbf{h}_t(\epsilon)) - \mathbf{s}_t(\epsilon)) > \alpha/2 \quad \forall t \in \{1,\ldots,n\}$$

The *(sequential) fat-shattering dimension for history trees* $\text{fat}(\mathcal{F}, \alpha)$ at scale α is the largest n such that $\mathcal{F}$ α-shatters some history tree of depth n.

In general, the combinatorial results of Section 14.3 do not immediately extend to history trees, due to the fact that, technically, two subtrees of a history tree are not (according to our definition) valid history trees. There might be modified definitions such that the combinatorial results extend more naturally. However, we now present a case where the upper bound on the sequential covering does hold. Together with the Dudley entropy bound for the sequential cover, this resolves a problem of [15].

30.4 Monotonic Experts

Consider a class Π of strategies (non-static experts), each predicting a value $\pi_t(z_1,\ldots,z_{t-1}) \in [0,1] = \mathcal{F}$ which we can treat as the probability of predicting a 1. The outcome sequence is $\{0,1\}$-valued, and the loss is $\ell(f,z) = |f - z|$. This is the probabilistic version of the indicator loss considered in the previous example, but one can pass easily from one setting to the other.

The requirement we place on the experts in the collection Π is that their predictions either never decrease or never increase. That is, for every $\pi \in \Pi$, either

$$\forall t, \forall (z_1,\ldots,z_n): \quad \pi_t(z_1,\ldots,z_{t-1}) \le \pi_{t+1}(z_1,\ldots,z_t)$$

or

$$\forall t, \forall (z_1,\ldots,z_n): \quad \pi_t(z_1,\ldots,z_{t-1}) \ge \pi_{t+1}(z_1,\ldots,z_t)$$

Following [15], we will call these experts *monotonic*. Observe that they are non-static, as their actual predictions may depend on the history. Also, observe that this class of experts is huge! We will calculate an upper bound on the sequential covering number and will see that it does not behave as a "parametric" class.

Recall that a related class of nondecreasing (or, nonincreasing) functions on the real line played an important role in Section 12.3 for the problem of Statistical Learning. The i.i.d. ℓ_∞-covering numbers at scale α were shown to be bounded

by $n^{2/\alpha}$. It was then shown that a bound of Proposition 12.3 at a single scale gives a suboptimal rate of $\tilde{O}(n^{-1/3})$ while the Dudley entropy integral bound of Theorem 12.4 yields the correct (up to logarithmic factors) rate of $\tilde{O}(n^{-1/2})$. Further, it was also shown that the i.i.d. fat-shattering dimension of this class is $2/\alpha$, which gave us a direct estimate on the ℓ_∞ covering numbers via the combinatorial result. This was all done in the setting of Statistical Learning, and one may wonder if the same gap exists in the sequential case.

In fact, we have all the tools necessary to prove the $\tilde{O}(n^{-1/2})$ bound for the sequential case with monotonic (in time) experts, resolving the question of [15]. Consider the bound given by Theorem 30.2:

$$\mathcal{V}(\Pi) \le 2 \sup_{\mathbf{w},\mathbf{z}} \mathbb{E}_\epsilon \sup_{\pi \in \Pi} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_t |\pi_t(\mathbf{w}_1(\epsilon),\ldots,\mathbf{w}_{t-1}(\epsilon)) - \mathbf{z}_t(\epsilon)| \right]$$

$$= 2 \sup_{\mathbf{w}} \mathbb{E}_\epsilon \sup_{\pi \in \Pi} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_t \pi_t(\mathbf{w}_1(\epsilon),\ldots,\mathbf{w}_{t-1}(\epsilon)) \right]$$

$$= 2 \sup_{\mathbf{h}} \mathbb{E}_\epsilon \sup_{\pi \in \Pi} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_t \pi_t(\mathbf{h}_t(\epsilon)) \right]$$

where we employed the same contraction technique as in Example 27 and then passed to a history tree. We claim that the sequential fat-shattering dimension for history trees is at most $2/\alpha$ for exactly the same reason as in the i.i.d. case: by taking the alternating sequence of signs (tracing the left-right-left-right path) the shattered tree needs to furnish a function that passes above and below the witness levels. Exactly as argued in Example 13, monotonicity prevents us from having a shattered tree of depth more than $2/\alpha$.

To conclude the argument, it remains to show that the covering numbers on any history tree are at most $O((2en/\alpha)^{\text{fat}_\alpha})$, similarly to the combinatorial result of Theorem 14.5, and then plug this bound into (30.2), exactly as in Section 12.3. The combinatorial proof does hold in this particular case, yet there are several subtleties: at certain steps in the proof of Theorem 14.6, we need to guarantee existence of a shattered tree, but it, unfortunately, the tree might not respect our restriction of being a history tree. Still, a simple modification works: we need to simply extend the notion of fat shattering from history trees to a larger set. Instead of getting involved in these modifications, we now give a separate proof of the fact that the covering numbers are polynomial for the class of nonincreasing experts (the non-decreasing case is analogous). The proof is also a good illustration of a combinatorial argument that underlies Theorem 14.6.

Lemma 30.5. *For the class* Π *of nonincreasing experts satisfying*

$$\forall t, \forall (z_1, \ldots, z_n): \quad \pi_t(z_1, \ldots, z_{t-1}) \geq \pi_{t+1}(z_1, \ldots, z_t),$$

the size of the sequential ℓ_∞ *cover is* $O((2en/\alpha)^{1/\alpha})$.

Proof. Define $g_k(d, n) = \sum_{i=0}^{d} \binom{n}{i} k^i$, as in the proof of Theorem 14.6. The key recursion for g_k is:

$$g_k(d, n) = g_k(d, n-1) + k g_k(d-1, n-1).$$

Throughout the proof we will refer to a subtree of a history tree simply as a history tree. Let $k = \frac{1}{\alpha}$, and assume for simplicity that k is an integer. Fix a history tree $\mathbf{h}$. Let $t = |\mathbf{h}_1|$ stand for the length of the history at the root $\mathbf{h}_1$ (that is, $\mathbf{h}_1 \in \mathcal{Z}^t$). Since $\mathbf{h}$ can be a subtree of the history tree, t may not be 1. Let Π' be a set of nonincreasing experts, subset of the original class Π. Define

$$d(\Pi', \mathbf{h}) = \max \{ \lceil \pi_t(\mathbf{h}_1)/\alpha \rceil : \pi \in \Pi' \}$$

It is not difficult to see that d in fact corresponds to an upper bound on the fat-shattering dimension of the class Π' on $\mathbf{h}$. We now proceed by induction. Fix d and n. We suppose by the way of induction that for any Π', the size of the ℓ_∞ cover of Π' on any history tree $\mathbf{h}$ of size $n-1$ is upper bounded by $g_k(d(\Pi', \mathbf{h}), n-1)$ under the assumption $d(\Pi', \mathbf{h}) \leq d-1$. The basis of the induction is trivial, so let's proceed to the induction step. Suppose we have a history tree $\mathbf{h}$ of size n and a subset Π' of nonincreasing experts such that $d(\Pi', \mathbf{h}) = d$. Define $1/\alpha$ subsets $\Pi_i = \{\pi \in \Pi' : \lceil \pi_t(\mathbf{h}_1)/\alpha \rceil = i\}$ according to their (rounded up) value at the root $\mathbf{h}_1$, where $t = |\mathbf{h}_1|$. Consider any $j \neq d(\Pi', \mathbf{h})$. By definition, $d(\Pi_j, \mathbf{h}) < d(\Pi', \mathbf{h})$. Let $\mathbf{h}^\ell$ and $\mathbf{h}^r$ be the left and the right subtrees of $\mathbf{h}$ at the root. By the nonincreasing property of the experts, $d(\Pi_j, \mathbf{h}^\ell) \leq d(\Pi_j, \mathbf{h}) < d(\Pi, \mathbf{h})$ and similarly $d(\Pi_j, \mathbf{h}^r) < d(\Pi, \mathbf{h})$. By the induction assumption, there is an ℓ_∞-cover of Π_j on $\mathbf{h}^r$ of size at most $g_k(d(\Pi', \mathbf{h}) - 1, n-1)$, and the same holds on the left subtree. The covering trees for the left subtree and for the right subtree can be joined together, with the root being the value $j\alpha$. It is easy to see that this preserves the property of an α-cover in the ℓ_∞ sense. In this process, the size of the cover stays the same, so the ℓ_∞-cover for Π_j is of size at most $g_k(d(\Pi', \mathbf{h}) - 1, n-1)$. Note that there are a total of at most $k = 1/\alpha$ such sets Π_j, resulting in the total size of the cover of $k \cdot g_k(d(\Pi', \mathbf{h}) - 1, n - 1)$. It remains to provide a cover for the case $j = d(\Pi', \mathbf{h})$. We perform the same

joining operation, yet we can only guarantee the size $g_k(d(\Pi', \mathbf{h}), n-1)$ rather than $g_k(d(\Pi', \mathbf{h}) - 1, n-1)$. But this is exactly what is needed to prove the inductive step, thanks to the recursive form of g_k. □

Remark 30.6. *Clearly, it is impossible to compete with the set of all possible strategies. It is quite interesting that the seemingly innocuous assumption of monotonicity of the non-static experts yields the rate of $\tilde{O}(n^{-1/2})$. Such a rate is expected from a small class. Yet, we see that the covering numbers are of the type $O((n/\alpha)^{1/\alpha})$, which is larger than the VC type $O((n/\alpha)^{dim})$ size of the ℓ_∞ cover. It is then imperative to use the Dudley type bound in order to obtain the correct rate. Thanks to the extension of the empirical process theory to the sequential prediction problem, this is now possible.*

30.5   Compression and Sufficient Statistics

We now assume that the strategies in Π have a particular form: they all work with a "sufficient statistic", or, more loosely, *compression* of the past data. Suppose "sufficient statistics" can take values in some set Γ. Fix a set $\bar{\Pi}$ of mappings $\bar{\pi} : \Gamma \mapsto \mathcal{F}$. We assume that all the strategies in Π are of the form

$$\pi_t(z_1, \ldots, z_{t-1}) = \bar{\pi}(\gamma(z_1, \ldots, z_{t-1})), \quad \text{for some} \quad \bar{\pi} \in \bar{\Pi}$$

and $\gamma : \mathcal{Z}^* \mapsto \Gamma$ is a fixed function (we can relax this assumption). Such a bottleneck Γ can arise due to a finite memory or finite precision, but can also arise if the strategies in Π are actually solutions to a statistical problem. The latter case is of a particular interest to us. If we assume a certain stochastic source for the data, we may estimate the parameters of the model, and there is often a natural set of sufficient statistics associated with it. If we collect all such solutions to stochastic models in a set Π, we may compete with all these strategies as long as the set where the sufficient statistics take values is not too large.

In the setting described above, the sequential Rademacher complexity for strategies Π can be upper bounded by the complexity of $\bar{\Pi}$ on Γ-valued trees:

$$\mathcal{R}^{seq}(\Pi) \le \sup_{\mathbf{g}} \mathbb{E}_{\epsilon} \sup_{\bar{\pi} \in \bar{\Pi}} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t \ell(\bar{\pi}, \mathbf{g}_t(\epsilon)) \right]$$

We refer to [23, 45] for more details on these types of bounds.

[TO BE EXPANDED]

31

Localized Analysis and Fast Rates.
Local Rademacher Complexities

Appendix

Lemma A.1 (McDiarmid's Inequality). *Let $\phi : \{0,1\}^n \to \mathbb{R}$ satisfy*

$$\forall j, \quad \forall z_1, \ldots, z_n, z_j', \quad |\phi(z_1, \ldots, z_j, \ldots, z_j) - \phi(z_1, \ldots, z_j', \ldots, z_n)| \le c_j$$

for some $c_1, \ldots, c_n \ge 0$. Suppose $Z_1, \ldots, Z_n$ are i.i.d. random variables. Then for any $\gamma > 0$,

$$\mathbb{P}\left(|\phi(Z_1, \ldots, Z_n) - \mathbb{E}\phi| \ge \gamma\right) \le 2\exp\left\{-\frac{2\gamma^2}{\sum_{t=1}^n c_i^2}\right\}$$

and the factor 2 in front of $\exp$ can be removed for the one-sided statements.

Lemma A.2 (Integrating out the Tails). *Suppose for a nonnegative random variable X we can prove*

$$\mathbb{P}\left(X \ge A + B\sqrt{\frac{\gamma}{n}}\right) \le e^{-\gamma}$$

Then $\mathbb{E}[X] \le A + \frac{C}{\sqrt{n}}$ for some C that depends only on B.

Proof.

$$\mathbb{E}[X] = \int_0^A \mathbb{P}(X > \delta)\, d\delta + \int_A^\infty \mathbb{P}(X > \delta)\, d\delta \le A + \int_0^\infty \mathbb{P}(X > A + \delta)\, d\delta$$

Setting $\delta = B\sqrt{\frac{\gamma}{n}}$, we get $\gamma = \frac{n\delta^2}{B^2}$ and

$$\mathbb{E}[X] \le A + \int_0^\infty \exp\left\{-\frac{n\delta^2}{B^2}\right\} d\delta = A + \frac{C}{\sqrt{n}}.$$

$\square$

Bibliography

[1] J. Abernethy, P. L. Bartlett, and E. Hazan. Blackwell approachability and low-regret learning are equivalent. *CoRR*, abs/1011.1936, 2010.

[2] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44:615–631, 1997.

[3] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[4] P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48(1):85–113, 2002.

[5] P.L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[6] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[7] J.O. Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.

[8] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

[9] D. Blackwell. Minimax vs. bayes prediction. *Probability in the Engineering and Informational Sciences*, 9:pp 53–58, 1995.

[10] D. Blackwell and Meyer A. Girshick. *Theory of games and statistical decisions*. Wiley publications in statistics. Dover Publications, 1979.

[11] V.I. Bogachev. *Measure Theory*, volume 2. Springer, 2007.

[12] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to Statistical Learning Theory. In O. Bousquet, U.v. Luxburg, and G. RâĂřtsch, editors, *Advanced Lectures in Machine Learning*, pages 169–207. springer, 2004.

[13] L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[14] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *Information Theory, IEEE Transactions on*, 50(9):2050–2057, 2004.

[15] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *The Annals of Statistics*, 27(6):1865–1895, 1999.

[16] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[17] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007.

[18] C.K. Chiang, T. Yang, C.J. Lee, M. Mahdavi, C.J. Lu, R. Jin, and S. Zhu. Online optimization with gradual variations. *COLT*, 2012.

[19] A.S. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Proceedings of the 20th annual conference on Learning theory*, pages 97–111. Springer-Verlag, 2007.

[20] P. A. Dawid and V.G. Vovk. Prequential probability: Principles and properties. *Bernoulli*, 5(1):125–162, 1999.

[21] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Verlag, 1996.

BIBLIOGRAPHY

[22] R.M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6(6):899–929, 1978.

[23] D. Foster, A. Rakhlin, K. Sridharan, and A. Tewari. Complexity-based approach to calibration with checking rules. In *COLT*, 2011.

[24] L. Györfi, M. Kohler, A. Krzyżack, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Verlag, 2002.

[25] D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

[26] E. Hazan and S. Kale. Better algorithms for benign bandits. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 38–47. Society for Industrial and Applied Mathematics, 2009.

[27] E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2):165–188, 2010.

[28] D. Hsu, S.M. Kakade, and T. Zhang. An analysis of random design linear regression. *Arxiv preprint arXiv:1106.2363*, 2011.

[29] A. Juditsky, P. Rigollet, and A.B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

[30] A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *J. Comput. Syst. Sci.*, 71(3):291–307, 2005.

[31] M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts*. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

[32] M.J. Kearns, R.E. Schapire, and L.M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2):115–141, 1994.

[33] V. Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems. *Saint-Flour Lectures Notes*, 2008.

[34] T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[35] M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer-Verlag, New York, 1991.

[36] HR Lerche and J. Sarkar. The blackwell prediction algorithm for infinite 0-1 sequences, and a generalization. *Statistical Decision Theory and Related Topies V, Ed.: SS Gupta, JO Berger, Springer Verlag*, pages 503–511, 1994.

[37] P. Massart. Concentration inequalities and model selection. 2007.

[38] S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures in Machine Learning, LNCS 2600, Machine Learning Summer School 2002, Canberra, Australia, February 11-22*, pages 1–40. Springer, 2003.

[39] S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152(1):37–55, 2003.

[40] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147, 1998.

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

[42] A. Rakhlin. Lecture notes on online learning, 2008. Available at http://www-stat.wharton.upenn.edu/~rakhlin/papers/online_learning.pdf.

[43] A. Rakhlin, O. Shamir, and K. Sridharan. Relax and localize: From value to algorithms, 2012. Available at http://arxiv.org/abs/1204.0870.

[44] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In *NIPS*, 2010. Available at http://arxiv.org/abs/1006.1138.

[45] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Beyond regret. In *COLT*, 2011. Available at http://arxiv.org/abs/1011.3168.

[46] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *NIPS*, 2011. Available at http://arxiv.org/abs/1104.5070.

[47] H. Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535, 1952.

[48] M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *The Annals of Mathematics*, 164(2):603–648.

[49] S. Simons. You cannot generalize the minimax theorem too much. *Milan Journal of Mathematics*, 59(1):59–64, 1989.

[50] K. Sridharan. *Learning From An Optimization Viewpoint*. PhD thesis, 2011.

[51] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[52] A.B. Tsybakov. *Introduction to nonparametric estimation*. Springer Verlag, 2009.

[53] L. G. Valiant. A theory of learnable. *Proc. of the 1984 STOC*, pages 436–445, 1984.

[54] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York, 1996.

[55] V. N. Vapnik. *Statistical learning theory*. J. Wiley, 1998.

[56] A. Wald. *Statistical decision functions*. John Wiley and Sons, 1950.

[57] L. Wasserman. *All of nonparametric statistics*. Springer-Verlag New York Inc, 2006.

[58] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27 (5):1564–1599, 1999.