Online Prediction

Peter L. Bartlett

April 9, 2016

Abstract

We review game-theoretic models of prediction, in which the process generating the data is modelled as an adversary with whom the prediction method competes. We present a formulation that encompasses a wide variety of decision problems, and focus on the relationship between prediction in this game-theoretic setting and prediction in the more standard probabilistic setting. In particular, we present a view of standard prediction strategies as Bayesian decision methods, and we show how the regret of optimal strategies depends on complexity measures that are closely related to those that appear in probabilistic settings.

1 Prediction as a Repeated Game

Consider a repeated game between a decision strategy (which we call the learner) and its environment (the world): at round t of the game, the learner plays an *action* a_t from some set $\mathcal{A}$ of actions, and the world subsequently reveals a loss $\ell_t : \mathcal{A} \to \mathbb{R}$ from some set $\mathcal{L}$. The learner's choices over n rounds of the game determine $\hat{L}_n$, its *cumulative loss*,

$$\hat{L}_n = \sum_{t=1}^{n} \ell_t(a_t).$$

This is a zero-sum game: the learner's aim is to minimize its *regret*, that is, to perform well compared to L_n^*, the cumulative loss of the best single choice (in retrospect) from the action set $\mathcal{A}$:

$$\text{regret} = \underbrace{\sum_{t=1}^{n} \ell_t(a_t)}_{\hat{L}_n} - \underbrace{\min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a)}_{L_n^*}.$$

On the other hand, the world is acting adversarially: it chooses each loss ℓ_t with full knowledge of the learner, so as to maximize the learner's regret. We would like to develop decision strategies for which, for all sequences of losses, the regret is small. We can think of the learner as choosing the action a_t as some function of the sequence of losses $\ell_1, \ldots, \ell_{t-1}$ that it has seen so far. We define the *minimax regret* as the value of the game,

$$\mathcal{V}_n(\mathcal{A}, \mathcal{L}) = \min_{a_1} \max_{\ell_1} \cdots \min_{a_n} \max_{\ell_n} \left(\sum_{t=1}^{n} \ell_t(a_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) \right).$$

There are several strong motivations for studying these adversarial formulations of prediction problems. First, in many cases, an adversarial model of the process generating the data in a prediction problem is appropriate. For example, many prediction problems that arise in computer security or computational finance should be viewed as adversarial.

Consider, for instance, the problem of detecting spam email. A decision method has access to features of email messages (such as information about the header, words in the message, attachments), and its action a_t is a mapping from the space of these features to $[0, 1]$ (think of the value as an estimated probability that the message is spam). The sender of a spam email can determine if it is delivered (or detected as spam), and it could use that information to modify subsequent spam messages. In the adversarial model, at each round the adversary chooses a feature vector $x_t \in \mathcal{X}$ and a label $y_t \in \{0, 1\}$, and the loss is defined as

$$\ell_t(a_t) = (y_t - a_t(x_t))^2.$$

The regret is then the excess squared error, over the best achievable on the data sequence:

$$\sum_{t=1}^{n} \ell_t(a_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) = \sum_{t=1}^{n} (y_t - a_t(x_t))^2 - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} (y_t - a(x_t))^2.$$

Minimizing regret ensures that the spam detection accuracy is close to the best performance in retrospect on the particular email sequence.

Consider the problem of choosing a portfolio, that is, a distribution of capital over a set of financial instruments, so as to maximize utility. Other market players can profit from making our decisions bad ones. For example, if trades have a market impact, making similar trades ahead of us can be profitable at our expense. Here, the action a_t is a distribution on m instruments

$$a_t \in \Delta^m = \{a \in [0,1]^m : \sum_i a_i = 1\}.$$

At each round, the adversary chooses a vector of returns $r_t \in \mathbb{R}_+^m$; the ith component is the ratio of the price of instrument i at time t to its price at the previous time, and the loss is defined as

$$\ell_t(a_t) = -\log(a_t \cdot r_t).$$

The regret is then the log of the ratio of the maximum value the portfolio would have at the end (for the best mixture choice) to the final portfolio value:

$$\sum_{t=1}^{n} \ell_t(a_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell_t(a) = \max_{a \in \mathcal{A}} \sum_{t=1}^{n} \log(a \cdot r_t) - \sum_{t=1}^{n} \log(a_t \cdot r_t).$$

Adversarial formulations of prediction problems are also appealing because they make only weak assumptions about the process generating the data. It is often straightforward to convert a strategy for an adversarial environment to a method for a probabilistic environment, and to convert a regret bound to a bound on performance in an i.i.d. setting.

Finally, studying adversarial models can reveal the deterministic heart of a statistical problem, and can give insight into the behavior of statistical methods designed for probabilistic settings. Indeed, as we shall see, there are strong similarities between the performance guarantees in the two cases, and there are significant overlaps in the design of methods for the two problems (for instance, regularization plays a central role, and many online prediction strategies have a natural interpretation as a Bayesian method). In fact, many statistical methods, based on *probabilistic assumptions*, turn out to be effective in adversarial settings, and analyzing their performance in these settings provides perspective on their robustness to the probabilistic assumptions.

These lecture notes provide an introduction to the analysis of strategies for adversarial prediction problems. Section 2 gives an easy introduction to the case of a finite comparison class $\mathcal{A}$. Section 3 compares the adversarial and the probabilistic formulation of prediction problems. Section 4 considers the value of the online prediction game, revealing in particular the close relationship between performance bounds for adversarial and probabilistic problems. We save most bibliographic notes to Section 5.

2 A Finite Comparison Class

2.1 Prediction with expert advice

In this section, we consider the case of a finite comparison class $\mathcal{A} = \{1, 2, \ldots, m\}$.

Consider the problem of predicting whether it will rain tomorrow. Given access to a set of m experts, who each make a forecast of 0 or 1, is it possible to predict almost as well as the best of these experts? Suppose that the ith expert makes a sequence of predictions $f_1^i, f_2^i, \ldots$ from $\{0,1\}$. At round t, the adversary chooses an outcome $y_t \in \{0,1\}$, and sets

$$\ell_t(i) = 1[f_t^i \neq y_t] = \begin{cases} 1 & \text{if } f_t^i \neq y_t, \\ 0 & \text{otherwise.} \end{cases}$$

2.2 With a perfect expert

Consider an easier version of this game: suppose that the adversary is constrained to choose the sequence y_t so that some expert incurs no loss that is, there is an $i^* \in \{1, \dots, m\}$ such that for all t, $y_t = f_t^{i^*}$, and hence $L_n^* = 0$.

How should we predict?

Define C_t as the set of experts who have been *correct* so far:

$$C_t = \{i : \ell_1(i) = \cdots = \ell_{t-1}(i) = 0\}.$$

Define the *halving strategy* as one that chooses a_t as an element of the set of experts i for which f_t^i agrees with the majority of elements of the set $\{f_t^j : j \in C_t\}$ (whenever there is a clear majority).

Proposition 1. *The halving strategy has regret no more than* $\log_2 m$.

Proof. If the strategy makes a mistake (that is, $\ell_t(a_t) = 1$), then the minority of $\{f_t^j : j \in C_t\}$ is correct, so at least half of the experts are eliminated:

$$|C_{t+1}| \leq \frac{|C_t|}{2}.$$

And otherwise $|C_{t+1}| \leq |C_t|$ (because $|C_t|$ never increases). Thus,

$$\hat{L}_n = \sum_{t=1}^{n} \ell_t(a_t) \leq \log_2 \frac{|C_1|}{|C_{n+1}|} = \log_2 m - \log_2 |C_{n+1}| \leq \log_2 m.$$

□

It is a straightforward proof, but it follows a common pattern: it exploits a measure of progress (here, $|C_t|$) that both changes monotonically when an excess loss is incurred (here, it halves), and is somehow constrained (here, it cannot fall below 1, because there is an expert who predicts perfectly).

What if there is no perfect expert?

2.3 Without a perfect expert: mixed actions

Without constraints on the loss of the best expert, it turns out that the game of prediction with expert advice is too difficult. To see this, consider a specific strategy for the adversary: choose the losses so that $\ell_t(a_t) = 1$ but $\ell_t(a) = 0$ for $a \neq a_t$. In that case, the learner incurs a total loss of n, whereas the best expert in hindsight has a total loss of only

$$\min_i \sum_{t=1}^{n} 1[a_t = i] \leq \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{n} 1[a_t = i] = \frac{n}{m}.$$

That is, the regret can be as bad as $n(1 - 1/m)$; this game is too difficult.

Rather than requiring the learner to choose a single action, we allow the learner to play a *mixed action*: the action a_t is chosen from the simplex Δ^m—the set of distributions on $\{1, \dots, m\}$,

$$\Delta^m = \left\{ a \in [0,1]^m : \sum_{i=1}^{m} a^i = 1 \right\}.$$

We can think of this as choosing an element of $\{1, \dots, m\}$ randomly, according to a distribution a_t. Alternatively, we can think of it as playing an element a_t of Δ^m, and incurring the expected loss,

$$\ell_t(a_t) = \sum_{i=1}^{m} a_t^i \ell_t(e_i),$$

where $\ell_t(e_i) \in [0,1]$ is the *loss* incurred by expert i (here, $e_i \in \mathbb{R}^m$ denotes the vector with a single 1 in the ith coordinate, and the rest zeros.)

2.4 Exponential weights

The exponential weights strategy is a simple and effective approach to this problem. It proceeds as follows:

- Maintain a set of (unnormalized) weights over experts:

$$w_1^i = 1,$$
$$w_{t+1}^i = w_t^i \exp\left(-\eta \ell_t(e_i)\right),$$

  where $\eta > 0$ is a parameter of the strategy.

- Choose a_t as the normalized vector,

$$a_t = \frac{1}{\sum_{i=1}^m w_t^i} w_t.$$

Theorem 2. *The exponential weights strategy with parameter*

$$\eta = \sqrt{\frac{8 \log m}{n}}$$

has regret satisfying

$$\hat{L}_n - L_n^* \leq \sqrt{\frac{n \log m}{2}}.$$

(Here and throughout, log denotes the natural logarithm.)

Proof. We use a measure of progress:

$$W_t = \sum_{i=1}^m w_t^i,$$

and show that W_n shrinks no faster than $\exp(-\eta L_n^*)$, but shrinks at least as fast as $\exp(-\eta \hat{L}_n)$. Comparing these bounds gives the result.

To see that W_t does not shrink too quickly when some expert has small loss, notice that

$$\begin{aligned}
\log \frac{W_{n+1}}{W_1} &= \log \left(\sum_{i=1}^m w_{n+1}^i \right) - \log m \\
&= \log \left(\sum_{i=1}^m \exp\left(-\eta \sum_{t=1}^n \ell_t(e_i) \right) \right) - \log m \\
&\geq \log \left(\max_i \exp\left(-\eta \sum_{t=1}^n \ell_t(e_i) \right) \right) - \log m \\
&= -\eta \min_i \left(\sum_{t=1}^n \ell_t(e_i) \right) - \log m \\
&= -\eta L_n^* - \log m.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\log \frac{W_{t+1}}{W_t} &= \log \left(\frac{\sum_{i=1}^m \exp(-\eta \ell_t(e_i)) w_t^i}{\sum_{i=1}^m w_t^i} \right) \\
&\leq -\eta \frac{\sum_{i=1}^m \ell_t(e_i) w_t^i}{\sum_{i=1}^m w_t^i} + \frac{\eta^2}{8} \\
&= -\eta \ell_t(a_t) + \frac{\eta^2}{8},
\end{aligned}$$

where we have used Hoeffding's inequality:

Lemma 3. *For a random variable $X \in [a, b]$ and $\lambda \in \mathbb{R}$,*

$$\log \left(\mathbb{E}\, e^{\lambda X} \right) \leq \lambda \, \mathbb{E}\, X + \frac{\lambda^2 (b-a)^2}{8}.$$

Proof. Define

$$A(\lambda) = \log\left(\mathbb{E}\, e^{\lambda X}\right) = \log\left(\int e^{\lambda x}\, dP(x)\right),$$

where $X \sim P$. Then for any $\lambda \in \mathbb{R}$, we can define a new random variable $X_\lambda \sim P_\lambda$ where

$$\frac{dP_\lambda}{dP}(x) = \exp\left(\lambda x - A(\lambda)\right).$$

(This is known as an *exponential family*: A is called the *log normalization*, λ the *natural parameter*, and we say that the family has *reference measure* P and *sufficient statistic* x.) Since P has bounded support, $A(\lambda) < \infty$ for all λ, and it is easy to check that

$$A'(\lambda) = \mathbb{E}\, X_\lambda,$$
$$A''(\lambda) = \operatorname{Var} X_\lambda.$$

Thus, we can write a Taylor expansion of A about $\lambda = 0$ (notice that $X_0 \sim P$):

$$A(\lambda) = \lambda\, \mathbb{E}\, X + \frac{\lambda^2}{2} \operatorname{Var} X_\xi$$

for some $\xi \in [0, \lambda]$. Since P has support in $[a, b]$, so does X_ξ, and hence $\operatorname{Var} X_\xi \le (b-a)^2/4$. This implies the result. □

Now, comparing the bounds on W_n, we have

$$-\eta L_n^* - \log m \le \log \frac{W_{n+1}}{W_1} \le -\eta \hat{L}_n + \frac{n\eta^2}{8}.$$

Thus,

$$\hat{L}_n - L_n^* \le \frac{\log m}{\eta} + \frac{\eta n}{8}.$$

Choosing the optimal η gives the result. □

With a perfect expert

We have seen that when there is an expert who incurs zero cumulative loss, there is a strategy (the halving algorithm) with per round regret of order $1/n$, whereas the exponential weights strategy gives a per round regret of order $1/\sqrt{n}$. It is natural to ask whether the exponential weights strategy gives the faster rate when $L^* = 0$.

In the proof, we can replace Hoeffding's inequality:

$$\log \mathbb{E}\, e^{\lambda X} \le \lambda\, \mathbb{E}\, X + \frac{\lambda^2}{8},$$

with:

$$\log \mathbb{E}\, e^{\lambda X} \le (e^\lambda - 1)\, \mathbb{E}\, X.$$

(for $X \in [0, 1]$, this is a linear upper bound on $e^{\lambda X}$). This gives

$$\begin{aligned} \log \frac{W_{t+1}}{W_t} &= \log\left(\frac{\sum_{i=1}^m \exp(-\eta \ell_t(e_i)) w_t^i}{\sum_i w_t^i}\right) \\ &\le \left(e^{-\eta} - 1\right) \ell_t(a_t). \end{aligned}$$

Thus

$$\hat{L}_n \le \frac{\eta}{1 - e^{-\eta}} L_n^* + \frac{\log m}{1 - e^{-\eta}}. \tag{1}$$

For example, if $L_n^* = 0$ and η is large, we again obtain a regret bound that is a constant times $\log m$. Notice that η large is rather similar to the halving algorithm (it puts equal weight on all experts that have zero loss so far, and roughly zero weight on the others).

With an unknown time horizon

One undesirable property of the exponential weights strategy is the fact that it uses n, the length of the game, to set an appropriate value for the gain constant η (in the proof above, we used the optimal setting $\eta = \sqrt{8 \log m/n}$). It is natural to ask if it is necessary to know n in advance. It turns out that using a time-varying value, such as $\eta_t = \sqrt{8 \log m/t}$ gives the same regret rate, but with worse constants. In fact, it is also possible to set η as a function of L_t^*, the best cumulative loss so far, to give the improved bound for small losses uniformly across time, again with worse constants. (See [3].)

With convex loss functions

We defined the loss for a point in the interior of the simplex by linearly extending its value at the vertices:

$$\ell_t(a) = \sum_i a^i \ell_t(e^i).$$

It's easy to see that we could replace this definition with any bounded convex function on Δ^m and retain the same regret bound; an equality becomes an inequality:

$$-\eta \frac{\sum_i \ell_t(e_i) w_t^i}{\sum_i w_t^i} \leq -\eta \ell_t(a_t).$$

But note that the exponential weights strategy only competes with the *corners* of the simplex:

Theorem 4. *For convex functions $\ell_t : \Delta^m \to [0,1]$, the exponential weights strategy with $\eta = \sqrt{8 \log m/n}$ satisfies*

$$\sum_{t=1}^n \ell_t(a_t) \leq \min_i \sum_{t=1}^n \ell_t(e^i) + \sqrt{\frac{n \log m}{2}}.$$

2.5 Exponential weights as Bayesian prediction

It turns out that we can interpret the exponential weights strategy as Bayesian prediction, and then the regret bound arises as a consequence of a straightforward regret bound for Bayesian prediction.

Recall that Bayesian prediction assumes a joint distribution over parameters $\theta \in \Theta$ and outcomes $y \in \mathcal{Y}$. We can factor that distribution into a *prior* distribution π on Θ and a *model*, or conditional distribution of y given θ. (We use notation that suggests Θ and $\mathcal{Y}$ are subsets of Euclidean space and the prior and likelihood are densities with respect to Lebesgue measure, but they can be defined as densities with respect to some other measure.) Given some data $y_1, \ldots, y_t \in \mathcal{Y}$, a Bayesian strategy computes the predictive distribution

$$\begin{aligned} \hat{p}_{t+1}(y) &= p(y|y_1, \ldots, y_t) \\ &= \int p(y|\theta) \underbrace{p(\theta|y_1, \ldots, y_t)}_{p_{t+1}(\theta)} \, d\theta, \end{aligned}$$

The *posterior distribution* $p_{t+1}(\theta) = p(\theta|y_1, \ldots, y_t)$ can be sequentially updated:

$$\begin{aligned} p_1(\theta) &= \pi(\theta) \\ p_{t+1}(\theta) &= \frac{p_t(\theta) p(y_t|\theta)}{\int p_t(\theta') p(y_t|\theta') d\theta'}. \end{aligned}$$

To see the relationship between exponential weights and a Bayesian strategy, set $\Theta = \{1, \ldots, m\}$ and define the prior as $\pi(j) = 1/m$. Set $\mathcal{Y} = [0,1]^m$ and define the model as

$$p(y|j) = \frac{1 - e^{-\eta}}{\eta} \exp(-\eta y^j) \tag{2}$$

for $j = 1, \ldots, m$ and $y = (y^1, \ldots, y^m) \in [0,1]^m$. (In fact, we can replace the constant $h_\eta := (1 - e^{-\eta})/\eta$ with $h(y)$ where $h : [0,1]^m \to \mathbb{R}$ is any function that ensures that $p(y|j)$ is a conditional probability density.) Then the Bayesian update to the distribution over Θ is

$$p_{t+1}(j) = \frac{1}{Z} p_t(j) \exp(-\eta y_t^j),$$

where Z is a normalization factor. If y_t^j is $\ell_t(e_j)$, the loss of expert j, this posterior distribution calculation is precisely the exponential weights update for the distribution over the set $\{1, \ldots, m\}$.

The regret bound for the exponential weights algorithm is a consequence of the following regret bound for a Bayesian strategy. Here, we consider another loss function: we define the loss incurred by a predictive distribution $\hat{p}_t$ with outcome y_t as the negative log of the predicted probability density,

$$\ell^{\log}(\hat{p}_t, y_t) = -\log \hat{p}_t(y_t).$$

The theorem shows that the regret for a parameter θ is small as long as the prior probability of the set of parameters that predict better (that is, parameters that incur a smaller cumulative loss) is large.

Theorem 5. *Fix a prior π on Θ and a model $p(y|\theta)$. Write*

$$L_n^{\log}(\theta) = -\sum_{t=1}^{n} \log p(y_t|\theta), \qquad \hat{L}_n^{\log} = -\sum_{t=1}^{n} \log \hat{p}_t(y_t) = -\sum_{t=1}^{n} \log p(y_t|y_1, \ldots, y_{t-1}).$$

For any sequence $y_1, \ldots, y_n$, and any $\theta \in \Theta$,

$$\hat{L}_n^{\log} \leq L_n^{\log}(\theta) - \log \pi\left(\{\theta' : L_n^{\log}(\theta') \leq L_n^{\log}(\theta)\}\right).$$

Proof. Fix a sequence $y_1, \ldots, y_n$ and a θ_0, and define the set of θ that have smaller cumulative loss than θ_0,

$$S = \{\theta \in \Theta : L_n^{\log}(\theta) \leq L_n^{\log}(\theta_0)\}.$$

Then

$$\begin{aligned}
\exp\left(-\hat{L}_n^{\log}\right) &= \hat{p}_1(y_1) \cdots \hat{p}_n(y_n) \\
&= p(y_1)p(y_2|y_1) \cdots p(y_n|y_1, \ldots, y_{n-1}) \\
&= p(y_1, \ldots, y_n) \\
&= \int_\Theta p(y_1|\theta) \cdots p(y_n|\theta)\, d\pi(\theta) \\
&= \int_\Theta \exp(-L_n^{\log}(\theta))\, d\pi(\theta) \\
&\geq \int_S \exp(-L_n^{\log}(\theta)) d\pi(\theta) \\
&\geq \exp(-L_n^{\log}(\theta_0))\pi(S).
\end{aligned}$$

□

We can use Theorem 5 to rederive the regret bound of Theorem 2 for the exponential weights strategy, by relating the log loss for the predicted distribution of y to the linear loss that we considered earlier: for a mixed action $a_t \in \Delta^m$ with outcome $y_t \in [0,1]^m$, we defined the loss as the linear function

$$\ell^{\text{dot}}(a_t, y_t) := a_t \cdot y_t = \mathbb{E}_{J \sim a_t}\, y_t^J.$$

Corollary 6. *The Bayesian strategy for the model (2) and a uniform prior on $\{1, \ldots, m\}$ achieves*

$$\hat{L}_n^{\text{dot}} = \min_j L_n^{\text{dot}}(e^j) + \frac{\log m}{\eta} + \frac{\eta n}{8}.$$

Proof. For the Bayesian strategy, the loss incurred by the predictive distribution $\hat{p}_t$ of y can be expressed in terms of a loss defined for the posterior distribution p_t on $\{1, \ldots, m\}$: Since

$$\hat{p}_t(y) = \mathbb{E}_{J \sim p_t}\, p(y|J) = h_\eta\, \mathbb{E}_{J \sim p_t} \exp(-\eta y^J),$$

we can write

$$\ell^{\log}(\hat{p}_t, y_t) = -\log \hat{p}_t(y_t) = -\log h_\eta - \log\left(\mathbb{E}_{J \sim p_t} \exp(-\eta y^J)\right) = -\log h_\eta + \eta \ell^{\text{Bayes}}(p_t, y_t),$$

where we have defined the *Bayes loss*,

$$\ell^{\text{Bayes}}(p_t, y_t) = -\frac{1}{\eta} \log \left(\mathbb{E}_{J \sim p_t} \exp(-\eta y_t^J) \right).$$

Because constant terms in the loss do not affect the regret, Theorem 5 shows that, for any prior on $\{1, \ldots, m\}$,

$$\hat{L}_n^{\text{Bayes}} \leq \min_j \left(L_n^{\text{Bayes}}(e^j) - \frac{\log \pi(j)}{\eta} \right), \tag{3}$$

where $\hat{L}_n^{\text{Bayes}}$ and L_n^{Bayes} are the cumulative Bayes losses of the Bayes strategy and the comparator, respectively.

Comparing the Bayes loss to the linear loss, we see that they coincide at the vertices of the simplex,

$$\ell^{\text{dot}}(e^j, y) = \ell^{\text{Bayes}}(e^j, y) = y_t^j.$$

The Bayes loss is convex, and Hoeffding's inequality shows how far it can be from the linear loss, that is, how much of a gap there is in Jensen's inequality: for $X \in [a, b]$:

$$-\log \left(\mathbb{E} \exp(-\eta X) \right) \geq \eta \, \mathbb{E} \, X - \frac{\eta^2}{8} (b - a)^2.$$

For $X = y_t^J \in [0, 1]$ with $J \sim p_t$, this is equivalent to the inequality

$$\ell^{\text{dot}}(p_t, y_t) \leq \ell^{\text{Bayes}}(p_t, y_t) + \frac{\eta}{8}.$$

Summing, applying the Bayes loss regret bound (3), and choosing the uniform prior $\pi(j) = 1/m$ gives

$$\begin{aligned}
\hat{L}_n^{\text{dot}} &\leq \hat{L}_n^{\text{Bayes}} + \frac{\eta n}{8} \\
&\leq \min_j \left(L_n^{\text{Bayes}}(e^j) - \frac{\log \pi(j)}{\eta} \right) + \frac{\eta n}{8} \\
&= \min_j \left(L_n^{\text{dot}}(e^j) - \frac{\log \pi(j)}{\eta} \right) + \frac{\eta n}{8} \\
&= \min_j L_n^{\text{dot}}(e^j) + \frac{\log m}{\eta} + \frac{\eta n}{8}.
\end{aligned}$$

□

3 Online and adversarial versus batch and probabilistic

In this section, we consider a batch, probabilistic formulation of a prediction problem. We are interested in how it compares to the online, game-theoretic formulation. We shall see that a successful method for the online formulation can be used to construct a successful method for the batch formulation. Later, we consider the relationship between regret bounds in the two formulations.

In the probabilistic formulation, a learning algorithm has access to a *sample* of size n,

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

drawn i.i.d. from an unknown probability distribution P on $\mathcal{X} \times \mathcal{Y}$. The algorithm chooses a *prediction rule* $\hat{f} : \mathcal{X} \to \mathcal{Y}$. It aims to minimize *risk*, or expected loss,

$$\mathbb{E} \, \ell(\hat{f}(X), Y),$$

and suffers *excess risk* of

$$\mathbb{E} \, \ell(\hat{f}(X), Y) - \min_{f \in F} \mathbb{E} \, \ell(f(X), Y),$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ is a *loss function* and F is a class of functions from $\mathcal{X}$ to $\mathcal{Y}$ (the *comparison class*). This is how much additional risk the prediction rule incurs compared to the best prediction rule in the comparison class F. Think of F as a finite class of experts, and we'd like to predict almost as well as the best in this class. To highlight the similarities between the online adversarial problem and the

batch probabilistic problem, we shall write the instantaneous loss in terms of an i.i.d. random function ℓ_t,

$$\ell_t(f) := \ell(f(X_t), Y_t),$$

so that the excess risk is

$$\mathbb{E}\,\ell_t(\hat{f}) - \min_{f \in F} \mathbb{E}\,\ell_t(f).$$

The halving strategy predicts according to the majority of experts that have incurred no loss. The following theorem shows that in the probabilistic setting, it suffices to follow any prediction rule that has incurred no loss so far. We use P to denote the distribution of ℓ_t and P_n to denote the empirical distribution that assigns mass $1/n$ at each ℓ_t, and write

$$Pf = \mathbb{E}_{\ell_t \sim P}\,\ell_t(f), \qquad P_n f = \frac{1}{n}\sum_{t=1}^{n} \ell_t(f)$$

for the risk and empirical risk, respectively.

Theorem 7. *If some $f^* \in F$ has $Pf^* = 0$, then choosing*

$$\hat{f} \in C_n = \{f \in F : P_n f = 0\}$$

leads to excess risk

$$\mathbb{E}\,\ell_t(\hat{f}) - \min_{f \in F} \mathbb{E}\,\ell_t(f) = O\left(\frac{\log |F|}{n}\right).$$

Proof.

$$\begin{aligned}
\Pr(P\hat{f} \geq \epsilon) &\leq \Pr(\exists f \in F : P_n f = 0,\, Pf \geq \epsilon) \\
&\leq |F|(1-\epsilon)^n \\
&\leq |F| e^{-n\epsilon}.
\end{aligned}$$

Integrating the tail bound $\Pr(nP\hat{f} \geq \log|F| + x) \leq e^{-x}$ gives $P\hat{f} \leq c \log|F|/n$. □

Notice the similarity to the regret rate for the halving strategy: there, if one of the m experts has zero cumulative loss, the per-trial regret decreases as $\log m/n$. Here, if some f in the comparison class F almost surely incurs zero loss, the excess risk decreases as $\log|F|/n$.

In the same way, the following theorem is analogous to the regret bound for the exponential weights strategy (Theorem 2). In the probabilistic setting, it suffices to choose a prediction rule that minimizes the empirical risk. In the online setting, an adversarial choice of data can easily make this simple approach fail; the argument in Section 2.3 shows that a learner that makes deterministic choices in the experts game will suffer linear regret.

Theorem 8. *Choosing $\hat{f}$ to minimize the empirical risk, $P_n f$, leads to excess risk*

$$\mathbb{E}\,\ell_t(\hat{f}) - \min_{f \in F} \mathbb{E}\,\ell_t(f) = O\left(\sqrt{\frac{\log |F|}{n}}\right).$$

The key step in proving the theorem is to bound the excess risk in terms of the supremum of an empirical process, the deviations between expectations and sample averages. Defining f^* as a minimizer in F of $\mathbb{E}\,\ell_t(f)$, we can write $\mathbb{E}\,\ell_t(\hat{f}) - \min_{f\in F}\mathbb{E}\,\ell_t(f) = \mathbb{E}\,P\hat{f} - Pf^*$ as the expectation of

$$\begin{aligned}
P\hat{f} - Pf^* &= P\hat{f} - P_n\hat{f} + P_n\hat{f} - Pf^* \\
&\leq P\hat{f} - P_n\hat{f} + P_n f^* - Pf^* \qquad \text{(since } P_n\hat{f} \text{ is minimal)} \\
&\leq 2\sup_{f \in F} |Pf - P_n f|\,.
\end{aligned}$$

That is, the excess risk can be bounded in terms of worst case deviations between sample averages and expectations. To prove the theorem, we shall take a short detour to investigate the behavior of the empirical process $f \mapsto Pf - P_n f$ and its largest values.

3.1 Analysis of probabilistic prediction: Rademacher averages

This section reviews the use of Rademacher averages to give excess risk bounds in probabilistic settings. We shall see in Section 4 that similar quantities and techniques give regret bounds in adversarial settings.

We have just seen that an empirical risk minimizer $\hat{f} \in \arg\min_{f \in F} P_n f$ has excess risk bounded by the supremum of an empirical process,

$$\|P - P_n\|_F := \sup_{f \in F} |Pf - P_n f| .$$

The following theorem shows that this quantity is concentrated about its expectation, and that its expectation is closely related to that of the *Rademacher process* indexed by F:

$$R_n f := \sum_{t=1}^{n} \epsilon_t \ell_t(f),$$

where the ℓ_t are i.i.d. and the ϵ_t are i.i.d. uniform on $\{-1, 1\}$.

Theorem 9. *Suppose the ℓ_t map to the interval $[0, 1]$. Then*

$$\frac{1}{2n} \mathbb{E} \|R_n\|_F - \sqrt{\frac{\log 2}{2n}} \leq \mathbb{E} \|P - P_n\|_F \leq \frac{2}{n} \mathbb{E} \|R_n\|_F ,$$

and with probability at least $1 - 2\exp(-2\epsilon^2 n)$,

$$\mathbb{E} \|P - P_n\|_F - \epsilon \leq \|P - P_n\|_F \leq \mathbb{E} \|P - P_n\|_F + \epsilon.$$

Thus, $\|P - P_n\|_F \xrightarrow{a.s.} 0$ iff $\mathbb{E} \|R_n\|_F / n \to 0$.

Notice that $\|R_n\|_F$ captures how well the losses $\ell_t(f)$ of functions in F can be aligned with a random vector $(\epsilon_1, \ldots, \epsilon_n)$. It is intuitive that if there is always a good alignment, then we should not expect expectations and sample averages to be close. The theorem shows that the two phenomena are very closely related.

Proof. To prove the upper bound on the expected supremum of $P - P_n$, define independent random variables ℓ'_t, with the same distribution as ℓ_t.

$$\begin{aligned}
\mathbb{E} \|P - P_n\|_F &= \mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E}\, \ell_t(f) - \ell_t(f) \right) \right| \\
&= \mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E}\, \ell'_t(f) - \ell_t(f) \right) \right| && \text{(because } \ell_t \stackrel{d}{=} \ell'_t\text{)} \\
&\leq \mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \left(\ell'_t(f) - \ell_t(f) \right) \right| \\
&= \mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \left(\ell'_t(f) - \ell_t(f) \right) \right| && \text{(because } \ell_t \stackrel{d}{=} \ell'_t\text{)} \\
&\leq 2\, \mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \ell_t(f) \right| && \text{(by the triangle inequality)} \\
&= \frac{2}{n} \mathbb{E} \|R_n\|_F .
\end{aligned}$$

The lower bound uses the assumption that the ℓ_t map to a bounded interval.

$$\begin{aligned}
\frac{1}{n}\,\mathbb{E}\,\|R_n\|_F &\le \mathbb{E}\left\|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(\ell_t(f)-\mathbb{E}\,\ell_t(f)\right)\right\|_F + \mathbb{E}\left\|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\,\mathbb{E}\,\ell_t(f)\right\|_F && \text{(by the triangle inequality)}\\
&\le \mathbb{E}\left\|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(\ell_t(f)-\ell_t'(f)\right)\right\|_F + \mathbb{E}\left\|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\,\mathbb{E}\,\ell_t(f)\right\|_F\\
&\le \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(X_i)-f(X_i')\right)\right\|_F + \|P\|_F\,\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right|\\
&\le \mathbb{E}\left\|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(\ell_t(f)-\ell_t'(f)\right)\right\|_F + \mathbb{E}\left|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\right| && \text{(because } |\ell_t|\le 1\text{)}\\
&\le \mathbb{E}\left\|\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\left(\ell_t(f)-\ell_t'(f)\right)\right\|_F + \sqrt{\frac{2\log 2}{n}} && \text{(by Lemma 11 below)}\\
&= \mathbb{E}\left\|\frac{1}{n}\sum_{t=1}^{n}\left(\ell_t(f)-\mathbb{E}\,\ell_t(f)+\mathbb{E}\,\ell_t'(f)-\ell_t'(f)\right)\right\|_F + \sqrt{\frac{2\log 2}{n}}\\
&\le 2\,\mathbb{E}\,\|P_n-P\|_F + \sqrt{\frac{2\log 2}{n}}.
\end{aligned}$$

To prove the concentration of the maximal deviations, we use the fact that the ℓ_t are independent and uniformly bounded, and apply the bounded differences inequality (Lemma 10 below).

Finally, almost-sure convergence follows from the Borel-Cantelli lemma. □

The proof used two lemmas. The first is a consequence of Hoeffding's inequality. See, for example, [22].

Lemma 10 (Bounded differences inequality)**.** *Suppose $\phi : \mathcal{X}^n \to \mathbb{R}$ has the following bounded differences property: for all i and $x_1, \ldots, x_n, x_i' \in \mathcal{X}$,*

$$\left|\phi(x_1,\ldots,x_n)-\phi(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)\right| \le B_i.$$

Then for independent $X_1, \ldots, X_n$,

$$\Pr\left(|\phi(X_1,\ldots,X_n)-\mathbb{E}\,\phi(X_1,\ldots,X_n)|\ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_i B_i^2}\right).$$

To bound the supremum of the Rademacher process for a finite class, we can appeal to the following lemma.

Lemma 11. *For $V \subseteq \mathbb{R}^n$,*

$$\mathbb{E}\max_{v\in V}\sum_{t=1}^{n}\epsilon_t v_t \le \sqrt{2\log|V|}\max_{v\in V}\|v\|.$$

Hence, if the loss functions ℓ_t map to the interval $[0,1]$ and F is finite,

$$\mathbb{E}\,\|R_n\|_F \le \sqrt{2n\log(2|F|)}.$$

Proof.

$$\begin{aligned}
\exp\left(\lambda\,\mathbb{E}\max_v\sum_t\epsilon_t v_t\right) &\le \mathbb{E}\exp\left(\lambda\max_v\sum_t\epsilon_t v_t\right) && \text{(Jensen's inequality)}\\
&\le \mathbb{E}\sum_v\exp\left(\lambda\sum_t\epsilon_t v_t\right)\\
&= \sum_v\prod_t\mathbb{E}\exp(\lambda\epsilon_t v_t)\\
&\le \sum_v\prod_t\exp(\lambda^2 v_t^2/2) && \text{(Hoeffding's inequality, Lemma 3)}\\
&= \sum_v\exp(\lambda^2\|v\|^2/2)\\
&\le |V|\exp\left(\lambda^2\max_v\|v\|^2/2\right).
\end{aligned}$$

The second part of the lemma follows: apply the inequality with $v_t = \pm\ell_t(f)$, so that $\|v\| \leq \sqrt{n}$. □

Now the proof of Theorem 8 is a straightforward combination of these ingredients: the argument after that theorem that excess risk is bounded by the supremum of $|P - P_n|$, Theorem 9, showing that this is bounded by the supremum of $|R_n|$, and Lemma 11:

$$\begin{aligned}\mathbb{E}\,\ell_t(\hat{f}) - \min_{f\in F}\mathbb{E}\,\ell_t(f) &\leq 2\,\mathbb{E}\,\|P - P_n\|_F \\ &\leq \frac{4}{n}\,\mathbb{E}\,\|R_n\|_F \\ &\leq \sqrt{\frac{32\log(2|F|)}{n}}.\end{aligned}$$

Rademacher averages are an elegant approach to the analysis of many prediction methods. For example, they can be bounded using covering numbers, or combinatorial dimensions, such as the Vapnik-Chervonenkis dimension, Pollard's pseudodimension, or the fat-shattering dimension (see, for example, [23]). Notice that we can write

$$\mathbb{E}\,\|R_n\|_F = \mathbb{E}\sup_{f\in F}\left|\sum_{t=1}^{n}\epsilon_t\ell_t(f)\right| \leq \sup_{\ell_1,\ldots,\ell_t}\mathbb{E}\sup_{f\in F}\left|\sum_{t=1}^{n}\epsilon_t\ell_t(f)\right| = \sup_{\ell_1,\ldots,\ell_t}\mathbb{E}\,\|R_n\|_F\,.$$

We call $\sup_{\ell_1,\ldots,\ell_t}\mathbb{E}\,\|R_n\|_F$ the *maximal Rademacher averages*. In the case of finite F,

$$\Pi_F(n) := \sup_{\ell_1,\ldots,\ell_t}|\{(\ell_1(f),\ldots,\ell_n(f)) : f\in F\}| \leq |F|,$$

so Lemma 11 also gives a bound on the maximal Rademacher averages. If $\{\{\ell(f) : \ell\in\mathcal{L}\} : f\in F\}$ is a Vapnik-Chervonenkis class, then $\Pi_F(n)$ grows only polynomially with n, and this appears inside the logarithm in Lemma 11, giving a bound on the maximal Rademacher averages that is only a log factor worse than the finite case.

In Section 4, we shall see quantities closely related to the maximal Rademacher averages arising in the adversarial setting.

3.2 Online to Batch Conversion

We have seen similar regret bounds for probabilistic and adversarial prediction problems. In this section, we see that we can use a strategy that performs well in the online, adversarial setting to obtain a strategy that performs well in the batch, probabilistic setting. The regret per trial in the probabilistic setting is bounded by the regret per trial in the adversarial setting.

Suppose that an online strategy, given observations $\ell_1,\ldots,\ell_{t-1}$, produces $a_t = S(\ell_1,\ldots,\ell_{t-1})$. To convert this to a method that is suitable for a probabilistic setting, we wish to exploit S in a method that uses i.i.d. observations $\ell_1,\ldots,\ell_n$ to choose an $\hat{a}\in\mathcal{A}$ so that

$$\mathbb{E}\,\ell_1(\hat{a}) - \min_{a\in\mathcal{A}}\mathbb{E}\,\ell_1\,(a)$$

is small whenever S has a good regret bound. One issue is that having a good regret bound implies that the predictions that S makes are accurate on average, but might not be uniformly accurate. We consider the following simple randomized method, which we call the *random tail* method:

1. Pick T uniformly from $\{0,\ldots,n\}$.
2. Let $\hat{a} = S(\ell_{T+1},\ldots,\ell_n)$.

It turns out that this approach gives small expected loss under a slightly milder assumption than i.i.d. observations: we only need that the process generating the loss sequence is stationary, that is, the joint distribution of any contiguous subsequence is independent of the time index.

Theorem 12. *If an online strategy S has a regret bound of C_{n+1} for sequences of length $n+1$, then for any stationary process generating the $\ell_1,\ldots,\ell_{n+1}$, the random tail method returns a prediction $\hat{a}$ that satisfies*

$$\mathbb{E}\,\ell_{n+1}(\hat{a}) - \min_{a\in\mathcal{A}}\mathbb{E}\,\ell_n\,(a) \leq \frac{C_{n+1}}{n+1}.$$

Proof.

$$\begin{aligned}
\mathbb{E}\,\ell_{n+1}(\hat{a}) &= \mathbb{E}\,\ell_{n+1}(A(\ell_{T+1},\ldots,\ell_n)) \\
&= \mathbb{E}\,\frac{1}{n+1}\sum_{t=0}^{n}\ell_{n+1}(A(\ell_{t+1},\ldots,\ell_n)) \\
&= \mathbb{E}\,\frac{1}{n+1}\sum_{t=0}^{n}\ell_{n-t+1}(A(\ell_1,\ldots,\ell_{n-t})) \\
&= \mathbb{E}\,\frac{1}{n+1}\sum_{t=1}^{n+1}\ell_t(A(\ell_1,\ldots,\ell_{t-1})) \\
&\leq \mathbb{E}\,\frac{1}{n+1}\left(\min_a\sum_{t=1}^{n+1}\ell_t(a)+C_{n+1}\right) \\
&\leq \min_a \mathbb{E}\,\ell_t(a) + \frac{C_{n+1}}{n+1}.
\end{aligned}$$

□

Notice that the random tail method is a randomized algorithm, and the expectation in the theorem averages also over its randomness. Thus, the theorem only shows how the method performs on average. If we wanted a method to have small excess risk with high probability (for instance, with independent, identically distributed data), this method is not suitable, but there are several alternatives. For instance, we could

1. Choose $\hat{a} = \frac{1}{n}\sum_{t=1}^{n} a_t$ (provided $\mathcal{A}$ is convex and the ℓ_t are all convex).
2. Choose $\hat{a}$ to minimize a high probability upper bound on expected loss,

$$\hat{a} = \arg\min_{a_t}\left(\frac{1}{n-t}\sum_{s=t+1}^{n}\ell_s(a_t) + c\sqrt{\frac{\log(n/\delta)}{n-t}}\right).$$

In both cases, the analysis involves concentration of martingales. See, for example, [9].

4 Optimal Regret

In this section, we consider the minimax regret, that is the regret incurred by a strategy that plays optimally against an optimal adversary:

$$V_n(\mathcal{A},\mathcal{L}) := \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_n}\sup_{\ell_n}\left(\sum_{t=1}^{n}\ell_t(a_t) - \inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right). \tag{4}$$

Here, the player chooses an action a_t at each step from the set $\mathcal{A}$ and the adversary chooses a loss $\ell_t : \mathcal{A} \to \mathbb{R}$ at each step from the set $\mathcal{L}$.

We saw in the previous section (in proving Theorem 8) that we can relate the performance of an empirical risk minimization algorithm on a probabilistic prediction problem to uniform bounds on the deviation between sample averages and expectations. The following theorem shows that the minimax regret in the adversarial setting is equal to a similar quantity: a deviation between expectations and averages.

Theorem 13. *If $\mathcal{A}$ is compact and all $\ell \in \mathcal{L}$ are convex, lower semi-continuous functions, then*

$$V_n(\mathcal{A},\mathcal{L}) = \sup_P \mathbb{E}\left(\sum_{t=1}^{n}\inf_{a_t\in\mathcal{A}}\mathbb{E}\left[\ell_t(a_t)|\ell_1,\ldots,\ell_{t-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^{n}\ell_t(a)\right),$$

where the supremum is over joint distributions P over sequences $\ell_1,\ldots,\ell_n$ in $\mathcal{L}^n$.

We can view the expression for the minimax regret as the value of a dual game, in which the adversary plays first by choosing the joint distribution P. The value of the game is the difference between the minimal conditional expected loss and the minimal empirical loss. If P were i.i.d., this would again be the difference between an expectation and a sample average.

Instead of constraining the ℓ_t uniformly, so that they are all chosen from a fixed set $\mathcal{L}$, we could replace the set $\mathcal{L}^n$ by a set of length n sequences of loss functions, and obtain a similar result.

The assumption of convexity of the ℓ_t is easily satisfied by allowing mixed strategies: if we replace $\mathcal{A}$ by the set $\mathcal{P}(\mathcal{A})$ of probability distributions on $\mathcal{A}$ and replace $a \mapsto \ell(a)$ by the linear, hence convex, $P \mapsto \mathbb{E}_{a\sim P}\,\ell(a)$, then we can replace convexity with a milder condition on the loss, as the following theorem shows.

Theorem 14. *Suppose $\mathcal{A}$ is a compact, separable metric space. Let $\mathcal{P}(\mathcal{A})$ denote the set of Borel probability measures on $\mathcal{A}$. Suppose all $\ell \in \mathcal{L}$ are lower semi-continuous and bounded from below. Then the value of the game with mixed strategies $a_t \sim P_t \in \mathcal{P}(\mathcal{A})$ is*

$$\begin{aligned} V_n(\mathcal{P}(\mathcal{A}),\mathcal{L}) &:= \inf_{P_1}\sup_{\ell_1}\cdots\inf_{P_n}\sup_{\ell_n}\mathbb{E}\left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a\in\mathcal{A}}\sum_{t=1}^n \ell_t(a)\right) \\ &= \sup_P \mathbb{E}\left(\sum_{t=1}^n \inf_{a_t\in\mathcal{A}}\mathbb{E}\left[\ell_t(a_t)|\ell_1,\ldots,\ell_{t-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^n \ell_t(a)\right), \end{aligned}$$

where the supremum is over joint distributions P on sequences $\ell_1,\ldots,\ell_n$ in $\mathcal{L}^n$.

4.1 Dual Game: Proof Idea

The central ingredient in the proof is Sion's generalization [27] of von Neumann's minimax theorem.

Lemma 15. *If $\mathcal{X}$ is compact and for every $y \in \mathcal{Y}$, $f(\cdot, y)$ is a convex, lower semi-continuous function, and for every $x \in \mathcal{X}$, $f(x,\cdot)$ is concave, then*

$$\inf_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}} f(x,y) = \sup_{y\in\mathcal{Y}}\inf_{x\in\mathcal{X}} f(x,y).$$

To apply the lemma to prove Theorem 13, we define $\mathcal{X}$ as $\mathcal{A}$, $\mathcal{Y}$ as the set of probability distributions on $\mathcal{L}$, and $f(a,Q) = c + \mathbb{E}_{\ell\sim Q}[\ell(a) + \phi(\ell)]$ for some c and ϕ.

Then, because allowing mixed strategies does not help the adversary, we have

$$\begin{aligned} V_n(\mathcal{A},\mathcal{L}) &= \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_n}\sup_{\ell_n}\left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a\in\mathcal{A}}\sum_{t=1}^n \ell_t(a)\right) \\ &= \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_n}\sup_{P_n}\mathbb{E}_{\ell_n}\left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a\in\mathcal{A}}\sum_{t=1}^n \ell_t(a)\right) \\ &= \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_{n-1}}\sup_{\ell_{n-1}}\sup_{P_n}\inf_{a_n}\mathbb{E}_{\ell_n}\left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a\in\mathcal{A}}\sum_{t=1}^n \ell_t(a)\right) \qquad \text{(by Lemma 15)} \\ &= \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_{n-1}}\sup_{\ell_{n-1}}\left(\sum_{t=1}^{n-1}\ell_t(a_t) + \sup_{P_n}\left(\inf_{a_n}\mathbb{E}\left[\ell_n(a_n)|\ell_1,\ldots,\ell_{n-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^n \ell_t(a)\right)\right) \\ &= \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_{n-1}}\sup_{P_{n-1}}\mathbb{E}_{\ell_{n-1}}\left(\sum_{t=1}^{n-1}\ell_t(a_t) + \sup_{P_n}\left(\inf_{a_n}\mathbb{E}\left[\ell_n(a_n)|\ell_1,\ldots,\ell_{n-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^n \ell_t(a)\right)\right) \\ &= \inf_{a_1}\sup_{\ell_1}\cdots\sup_{P_{n-1}}\inf_{a_{n-1}}\mathbb{E}_{\ell_{n-1}}\left(\sum_{t=1}^{n-1}\ell_t(a_t) + \sup_{P_n}\left(\inf_{a_n}\mathbb{E}\left[\ell_n(a_n)|\ell_1,\ldots,\ell_{n-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^n \ell_t(a)\right)\right) \\ &= \inf_{a_1}\sup_{\ell_1}\cdots\inf_{a_{n-2}}\sup_{\ell_{n-2}}\left(\sum_{t=1}^{n-2}\ell_t(a_t) + \sup_{P_{n-1}^n}\mathbb{E}\left(\sum_{t=n-1}^n\inf_{a_t}\mathbb{E}\left[\ell_t(a_t)|\ell_1,\ldots,\ell_{t-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^n \ell_t(a)\right)\right) \\ &\vdots \\ &= \sup_P \mathbb{E}\left(\sum_{t=1}^n \inf_{a_t}\mathbb{E}\left[\ell_t(a_t)|\ell_1,\ldots,\ell_{t-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^n \ell_t(a)\right). \end{aligned}$$

To prove Theorem 14, we apply Lemma 15, defining $\mathcal{X}$ as $\mathcal{P}(\mathcal{A})$, $\mathcal{Y}$ as the set of probability distributions on $\mathcal{L}$, and $f(P,Q) = c + \mathbb{E}_{\ell\sim Q}[\mathbb{E}_{a\sim P}\,\ell(a) + \phi(\ell)]$ for some c and ϕ. Then the following lemma shows that the conditions of the theorem ensure that $\mathcal{A}$ is compact and $P \mapsto \mathbb{E}_{a\sim P}\,\ell(a)$ is lower semi-continuous, as required. (The result follows from properties of Borel measures and the portmanteau theorem. See, for example, [13].)

Lemma 16. *Suppose $\mathcal{A}$ is a compact, separable metric space. Let $\mathcal{P}(\mathcal{A})$ denote the set of Borel probability measures on $\mathcal{A}$. Then*

1. *$\mathcal{P}(\mathcal{A})$ is metrizable (by the Lévy-Prokhorov metric), separable, and compact, and*
2. *For any $\ell : \mathcal{A} \to \mathbb{R}$ that is lower semi-continuous and bounded from below, the function $E_\ell : \mathcal{P}(\mathcal{A}) \to \mathbb{R}$ defined by $E_\ell(P) = \mathbb{E}_{a\sim P}\, \ell(a)$ is lower semi-continuous.*

The rest of the proof proceeds as before.

4.2 Optimal Regret and Sequential Rademacher Averages

We can obtain an upper bound on the optimal regret in terms of a quantity that is closely related to the maximal Rademacher averages that we encountered in Section 3.1. Recall that deviations between sample averages and expectations in the probabilistic setting was bounded above and below by the expected supremum of the Rademacher process $R_n = \sum_{t=1}^n \epsilon_t \ell_t$, where the ℓ_t are i.i.d. and the ϵ_t are independent, uniform, $\{-1, 1\}$-valued. In the adversarial setting, the upper bound depends on a similar stochastic process, but with dependent ℓ_t and ϵ_t.

Definition 17. *Define the* sequential Rademacher averages *of action set $\mathcal{A}$ and loss set $\mathcal{L}$ as*

$$S_n(\mathcal{A}, \mathcal{L}) = \sup_{\ell_1 \in \mathcal{L}} \mathbb{E}_{\epsilon_1} \cdots \sup_{\ell_n \in \mathcal{L}} \mathbb{E}_{\epsilon_n} \sup_{a \in \mathcal{A}} \sum_{t=1}^n \epsilon_t \ell_t(a),$$

where the ϵ_t are i.i.d. uniform $\{-1, 1\}$-valued random variables.

In what follows, when we write $V_n(\mathcal{A}, \mathcal{L})$, we assume either

1. that $\mathcal{A}$ and $\mathcal{L}$ satisfy the conditions of Theorem 13 (that is, $\mathcal{A}$ is compact, all $\ell \in \mathcal{L}$ are convex and l.s.c.), in which case $V_n(\mathcal{A}, \mathcal{L})$ is the value of the game, defined in (4):

$$V_n(\mathcal{A}, \mathcal{L}) := \inf_{a_1} \sup_{\ell_1} \cdots \inf_{a_n} \sup_{\ell_n} \left(\sum_{t=1}^n \ell_t(a_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a) \right),$$

   or

2. that $\mathcal{A}$ and $\mathcal{L}$ satisfy the conditions of Theorem 14 (that is, $\mathcal{A}$ is a compact, separable metric space, all $\ell \in \mathcal{L}$ are l.s.c. and bounded from below), in which case we can replace $V_n(\mathcal{A}, \mathcal{L})$ with $V_n(\mathcal{P}(\mathcal{A}), \mathcal{L})$, where $\mathcal{P}(\mathcal{A})$ is the set of Borel probability measures on $\mathcal{A}$.

Theorem 18. $V_n(\mathcal{A}, \mathcal{L}) \le 2S_n(\mathcal{A}, \mathcal{L})$.

Compare this result to the Rademacher bounds on excess risk in the probabilistic setting:

$$n \left(\mathbb{E}\, \ell_t(\hat{f}) - \min_{f \in F} \mathbb{E}\, \ell_t(f) \right) \le 4\, \mathbb{E}\, \|R_n\|_F = 4\, \mathbb{E} \sup_{f \in F} \left| \sum_{t=1}^n \epsilon_t \ell_t(f) \right| \le 4 \sup_{\ell_1, \ldots, \ell_n \in \mathcal{L}} \mathbb{E} \left\| \sum_{t=1}^n \epsilon_t \ell_t \right\|_F .$$

In the probabilistic setting, the ℓ_t that appear in the definition of the Rademacher process R_n are i.i.d., but we can obtain an upper bound by a deterministic choice of values that maximize the expected supremum over F. In the adversarial setting, the choice of the ℓ_t that appear in the definition of the Rademacher process is deterministic, but crucially it can depend on the preceding $\epsilon_1, \ldots, \epsilon_{t-1}$. We can also think of the alternating supremum-expectation operation as a supremum over mappings $L : \bigcup_{t=0}^{n-1} \{-1, 1\}^t \to \mathcal{L}$:

$$S_n(\mathcal{A}, \mathcal{L}) = \sup_L \mathbb{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^n \epsilon_t L(\epsilon_1, \ldots, \epsilon_{t-1})(a).$$

Of course, requiring that L only depend on the length of its input (so that the ℓ_t form a fixed sequence, independent of the ϵ_t) gives a smaller quantity:

$$\sup_{\ell_1, \ldots, \ell_n \in \mathcal{L}} \mathbb{E} \sup_{a \in \mathcal{A}} \sum_{t=1}^n \epsilon_t \ell_t(a) \le S_n(\mathcal{A}, \mathcal{L}).$$

Proof. We start with the dual form of the optimal regret.

$$\begin{aligned} V_n(\mathcal{A},\mathcal{L}) &= \sup_P \mathbb{E}\left(\sum_{t=1}^n \inf_{a_t\in\mathcal{A}} \mathbb{E}\left[\ell_t(a_t)|\ell_1,\dots,\ell_{t-1}\right] - \inf_{a\in\mathcal{A}}\sum_{t=1}^n \ell_t(a)\right) \\ &= \sup_P \mathbb{E}\sup_{a\in\mathcal{A}}\sum_{t=1}^n \left(\inf_{a_t\in\mathcal{A}} \mathbb{E}\left[\ell_t(a_t)|\ell_1,\dots,\ell_{t-1}\right] - \ell_t(a)\right) \\ &\le \sup_P \mathbb{E}\sup_{a\in\mathcal{A}}\sum_{t=1}^n \left(\mathbb{E}\left[\ell_t(a)|\ell_1,\dots,\ell_{t-1}\right] - \ell_t(a)\right). \end{aligned} \tag{5}$$

Now define a *tangent sequence*, ℓ'_t, that is conditionally independent of $\ell_t,\dots,\ell_n$ given $\ell_1,\dots,\ell_{t-1}$, and has the same conditional distribution as ℓ_t.

$$\begin{aligned} V_n(\mathcal{A},\mathcal{L}) &\le \sup_P \mathbb{E}\sup_{a\in\mathcal{A}}\sum_{t=1}^n \left(\mathbb{E}\left[\ell_t(a)|\ell_1,\dots,\ell_{t-1}\right] - \ell_t(a)\right) \\ &= \sup_P \mathbb{E}\sup_{a\in\mathcal{A}}\sum_{t=1}^n \left(\mathbb{E}\left[\ell'_t(a)|\ell_1,\dots,\ell_n\right] - \ell_t(a)\right) && (\ell'_t \text{ has the same conditional}) \\ &\le \sup_P \mathbb{E}\sup_{a\in\mathcal{A}}\sum_{t=1}^n \left(\ell'_t(a) - \ell_t(a)\right) && (\text{moving the sup inside the } \mathbb{E}) \\ &= \sup_P \mathbb{E}\sup_{a\in\mathcal{A}}\left(\sum_{t=1}^{n-1}\left(\ell'_t(a)-\ell_t(a)\right) + \epsilon_n\left(\ell'_n(a)-\ell_n(a)\right)\right). && (\text{same conditional}) \end{aligned}$$

At this point, we cannot use the same argument to split off a term of the form $\epsilon_{n-1}(\ell'_{n-1}(a)-\ell_{n-1}(a))$. The reason is that the supremum does not have the same distribution if we interchange ℓ'_{n-1} and ℓ_{n-1}. For instance, given $\ell_1,\dots,\ell_{n-2}$, ℓ_n is conditionally independent of ℓ'_{n-1} but not of ℓ_{n-1}. For this reason, we replace the expectation over ℓ_n and ℓ'_n with a supremum:

$$\begin{aligned} V_n(\mathcal{A},\mathcal{L}) &\le \sup_P \mathbb{E}_{\ell_1,\dots,\ell_{n-1}}\mathbb{E}_{\ell_n,\ell'_n}\mathbb{E}_{\epsilon_n}\sup_{a\in\mathcal{A}}\left(\sum_{t=1}^{n-1}\left(\ell'_t(a)-\ell_t(a)\right)+\epsilon_n\left(\ell'_n(a)-\ell_n(a)\right)\right) \\ &\le \sup_P \mathbb{E}_{\ell_1,\dots,\ell_{n-1}}\sup_{\ell_n,\ell'_n}\mathbb{E}_{\epsilon_n}\sup_{a\in\mathcal{A}}\left(\sum_{t=1}^{n-1}\left(\ell'_t(a)-\ell_t(a)\right)+\epsilon_n\left(\ell'_n(a)-\ell_n(a)\right)\right) \\ &= \sup_P \mathbb{E}_{\ell_1,\dots,\ell_{n-1}}\mathbb{E}_{\epsilon_{n-1}}\sup_{\ell_n,\ell'_n}\mathbb{E}_{\epsilon_n}\sup_{a\in\mathcal{A}}\left(\sum_{t=1}^{n-2}\left(\ell'_t(a)-\ell_t(a)\right)+\sum_{t=n-1}^{n}\epsilon_t\left(\ell'_t(a)-\ell_t(a)\right)\right) \\ &\le \sup_P \mathbb{E}_{\ell_1,\dots,\ell_{n-2}}\sup_{\ell_{n-1},\ell'_{n-1}}\mathbb{E}_{\epsilon_{n-1}}\sup_{\ell_n,\ell'_n}\mathbb{E}_{\epsilon_n}\sup_{a\in\mathcal{A}}\left(\sum_{t=1}^{n-2}\left(\ell'_t(a)-\ell_t(a)\right)+\sum_{t=n-1}^{n}\epsilon_t\left(\ell'_t(a)-\ell_t(a)\right)\right) \\ &\vdots \\ &\le \sup_{\ell_1,\ell'_1}\mathbb{E}_{\epsilon_1}\cdots\sup_{\ell_n,\ell'_n}\mathbb{E}_{\epsilon_n}\sup_{a\in\mathcal{A}}\left(\sum_{t=1}^{n}\epsilon_t\left(\ell'_t(a)-\ell_t(a)\right)\right) \\ &\le 2\sup_{\ell_1}\mathbb{E}_{\epsilon_1}\cdots\sup_{\ell_n}\mathbb{E}_{\epsilon_n}\sup_{a\in\mathcal{A}}\left(\sum_{t=1}^{n}\epsilon_t\ell_t(a)\right) \\ &\le 2S_n(\mathcal{A},\mathcal{L}). \end{aligned}$$

□

4.2.1 Sequential versus maximal Rademacher averages: an example

Consider step functions defined on the real line:

$$f_a : x \mapsto 1[x \ge a],$$

define $\mathcal{A}=\mathbb{R}$, and define losses $\ell_{x,y}\in\mathcal{L}\subset\{0,1\}^{\mathcal{A}}$ in terms of $x\in\mathbb{R}$ and $y\in\{0,1\}$ via

$$\ell_{x,y}(a) = 1[f_a(x)\neq y].$$

The following proposition shows that, for this example, the maximal and sequential Rademacher averages are very different: rescaled so that they are comparable, the maximal Rademacher averages are a factor of roughly $\sqrt{n}$ smaller than the sequential Rademacher averages.

Proposition 19. *For the step functions, the Rademacher averages and the sequential Rademacher averages satisfy*

$$\frac{1}{n}\,\mathbb{E}\,\|R_n\|_{\mathcal{A}} = \mathbb{E}\sup_a \left|\frac{1}{n}\sum_{t=1}^n \epsilon_t \ell_t(a)\right| \leq \sqrt{\frac{2\log(n+1)}{n}},$$
$$\frac{1}{n}\sup_{\ell_1}\mathbb{E}_{\epsilon_1}\cdots\sup_{\ell_n}\mathbb{E}_{\epsilon_n}\sup_a \sum_{t=1}^n \epsilon_t \ell_t(a) = \frac{1}{2}.$$

Proof. First, we calculate an upper bound on the maximal Rademacher averages.

$$\begin{aligned}\mathbb{E}\,\|R_n\|_{\mathcal{A}} &\leq \sup_{(x_1,y_1),\ldots,(x_n,y_n)} \mathbb{E}\,\|R_n\|_{\mathcal{A}} \\ &= \sup_{(x_1,y_1),\ldots,(x_n,y_n)} \mathbb{E}\sup_{a\in\mathbb{R}}\left|\sum_{t=1}^n \epsilon_t \ell_{x_t,y_t}(a)\right|.\end{aligned}$$

Consider the set of loss vectors

$$V := \{(\ell_{x_1,y_1}(a),\ldots,\ell_{x_n,y_n}(a)) : a\in\mathbb{R}\}.$$

The cardinality of this set is not affected if we set $y_1 = \cdots = y_n = 0$ and reorder the x_ts so that $x_1 \leq \cdots \leq x_n$. But then V is a set of monotone sequences of binary vectors, so $|V| \leq n+1$. Lemma 11 shows that

$$\mathbb{E}\,\|R_n\|_{\mathcal{A}} \leq \sqrt{2n\log(n+1)}.$$

(It turns out that the log factor is unnecessary in this bound, but this needs a more refined argument.)

Second, for the sequential Rademacher averages, we have

$$\begin{aligned}&\sup_{x_1,y_1}\mathbb{E}_{\epsilon_1}\cdots\sup_{x_n,y_n}\mathbb{E}_{\epsilon_n}\sup_a\sum_{t=1}^n \epsilon_t\ell_{x_t,y_t}(a) \\ &\geq \sup_{x_1}\mathbb{E}_{\epsilon_1}\cdots\sup_{x_n}\mathbb{E}_{\epsilon_n}\sup_a\sum_{t=1}^n \epsilon_t 1[x_t < a].\end{aligned}$$

To maximize this quantity, we should aim to choose a so that $x_t < a$ when $\epsilon_t = 1$ and $x_t \geq a$ when $\epsilon_t = -1$. To ensure that this is possible, we can choose

$$\begin{aligned}x_1 &= 0,\\ x_2 &= \epsilon_1/2,\\ x_3 &= \epsilon_1/2 + \epsilon_2/4,\\ &\vdots\\ x_t &= \sum_{i=1}^{t-1} 2^{-i}\epsilon_i.\end{aligned}$$

Then if we set $a = x_n + 2^{-n}\epsilon_n$, it is easy to see that, for all t,

$$\epsilon_t 1[x_t < a] = \begin{cases}1 & \text{if } \epsilon_t = 1,\\ 0 & \text{otherwise,}\end{cases}$$

which is maximal. So the sequential Rademacher averages are

$$\sup_{\ell_1}\mathbb{E}_{\epsilon_1}\cdots\sup_{\ell_n}\mathbb{E}_{\epsilon_n}\sup_a\sum_{t=1}^n \epsilon_t\ell_t(a) = \mathbb{E}\sum_{t=1}^n 1[\epsilon_t = 1] = \frac{n}{2}.$$

□

4.3 Martingales and Sequential Rademacher Averages

In this section, we see that there is a close relationship between sequential Rademacher averages of $\mathcal{L}$ and the size of martingales with differences in $\mathcal{L}$. In particular, if $\ell \in \mathcal{L}$ implies $-\ell \in \mathcal{L}$, this result shows that the value of the game is within a factor of 2 of the sequential Rademacher averages.

An $\mathbb{R}^{\mathcal{A}}$-*valued martingale* is a sequence $Z_1, Z_2, \ldots$ of stochastic processes, each indexed by $\mathcal{A}$, for which

$$\mathbb{E} \sup_{a \in \mathcal{A}} |Z_t(a)| < \infty \qquad \text{and} \qquad \mathbb{E}\left[Z_t | Z_1, \ldots, Z_{t-1}\right] = Z_{t-1}.$$

The sequence $Z_1, Z_2 - Z_1, \ldots, Z_n - Z_{n-1}$ of increments is called the *difference sequence*.

Theorem 20. *Let $\mathcal{M}_{\mathcal{L}}$ be the set of $\mathbb{R}^{\mathcal{A}}$-valued martingales with difference sequences in $\mathcal{L} \cup -\mathcal{L}$, where*

$$-\mathcal{L} = \left\{-\ell : \ell \in \mathcal{L}\right\}.$$

Then

$$\sup_{\ell_1} \mathbb{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbb{E}_{\epsilon_n} \sup_a \sum_{t=1}^n \epsilon_t \ell_t(a) \leq \sup_{\{Z_t\} \in \mathcal{M}_{\mathcal{L}}} \mathbb{E} \sup_{a \in \mathcal{A}} Z_t(a).$$

If $\mathcal{L}$ is symmetric, that is, $-\mathcal{L} = \mathcal{L}$, then

$$\sup_{\{Z_t\} \in \mathcal{M}_{\mathcal{L}}} \mathbb{E} \sup_{a \in \mathcal{A}} Z_t(a) \leq V_n(\mathcal{A}, \mathcal{L}) \leq 2 \sup_{\ell_1} \mathbb{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbb{E}_{\epsilon_n} \sup_a \sum_{t=1}^n \epsilon_t \ell_t(a) \leq 2 \sup_{\{Z_t\} \in \mathcal{M}_{\mathcal{L}}} \mathbb{E} \sup_{a \in \mathcal{A}} Z_t(a).$$

Proof. Fix a strategy $L : \bigcup_{t=0}^{n-1} \{-1, 1\}^t \to \mathcal{L}$. Define $\ell_t = L(\epsilon_1, \ldots, \epsilon_{t-1})$ and $Z_t = \sum_{i=1}^t \epsilon_i \ell_i$. Then the sequential Rademacher averages can be written

$$\sup_{\ell_1} \mathbb{E}_{\epsilon_1} \cdots \sup_{\ell_n} \mathbb{E}_{\epsilon_n} \sup_a \sum_{t=1}^n \epsilon_t \ell_t(a) = \sup_L \mathbb{E} \sup_{a \in \mathcal{A}} Z_t(a) \leq \sup_{\{Z_t\} \in \mathcal{M}_{\mathcal{L}}} \mathbb{E} \sup_{a \in \mathcal{A}} Z_t(a),$$

because $\{Z_t\}$ is an $\mathbb{R}^{\mathcal{A}}$-valued martingale with difference sequence in $\mathcal{L} \cup -\mathcal{L}$.

For the lower bound, consider the duality result:

$$V_n(\mathcal{A}, \mathcal{L}) = \sup_P \mathbb{E}\left(\sum_{t=1}^n \inf_{a_t \in \mathcal{A}} \mathbb{E}\left[\ell_t(a_t) | \ell_1, \ldots, \ell_{t-1}\right] - \inf_{a \in \mathcal{A}} \sum_{t=1}^n \ell_t(a)\right),$$

where the supremum is over joint distributions of sequences $\ell_1, \ldots, \ell_n$. When $\mathcal{L}$ is symmetric, a martingale $\{Z_t\}$ in $\mathcal{M}_{\mathcal{L}}$ has a difference sequence $\{\ell_t\}$ in $\mathcal{L}^n$. If we restrict P to the subset of joint distributions on $\mathcal{L}^n$ corresponding to difference sequences of martingales $\{Z_t\}$ in $\mathcal{M}_{\mathcal{L}}$, then

$$\mathbb{E}\left[\ell_t | \ell_1, \ldots, \ell_{t-1}\right] = 0$$

and

$$\sum_{t=1}^n \ell_t = Z_t,$$

and so we have

$$V_n(\mathcal{A}, \mathcal{L}) \geq \sup_{\{Z_t\} \in \mathcal{M}_{\mathcal{L}}} \mathbb{E} - \inf_{a \in \mathcal{A}} Z_n(a) = \sup_{\{Z_t\} \in \mathcal{M}_{\mathcal{L}}} \mathbb{E} \sup_{a \in \mathcal{A}} Z_n(a).$$

□

Although the symmetry condition in this theorem seems to require that we allow losses to be negative, it is clear that an offset does not change the value of the game or the sequential Rademacher averages.

Theorem 21. *For an action set $\mathcal{A}$, loss class $\mathcal{L} \subseteq \mathbb{R}^{\mathcal{A}}$, and constants $b, c \in \mathbb{R}$, define*

$$b\mathcal{L} = \left\{a \mapsto b\ell(a) : \ell \in \mathcal{L}\right\},$$
$$\mathcal{L} + c = \left\{a \mapsto \ell(a) + c : \ell \in \mathcal{L}\right\}.$$

Then

$$V_n\left(\mathcal{A}, |b|\mathcal{L} + c\right) = |b| V_n\left(\mathcal{A}, \mathcal{L}\right), \qquad S_n\left(\mathcal{A}, b\mathcal{L} + c\right) = |b| S_n\left(\mathcal{A}, \mathcal{L}\right).$$

Furthermore, if $\mathcal{A}_1 \subseteq \mathcal{A}$, and $\mathcal{L}_1, \mathcal{L}_2 \subseteq \mathcal{L}$,

$$S_n\left(\mathcal{A}_1, \mathcal{L}\right) \leq S_n\left(\mathcal{A}, \mathcal{L}\right), \qquad S_n\left(\mathcal{A}, \mathcal{L}_1\right) \leq S_n\left(\mathcal{A}, \mathcal{L}\right).$$

Proof. The equality for $V_n(\mathcal{A}, |b|\mathcal{L} + c)$ is immediate from the definition of V_n. For the sequential Rademacher averages, we have

$$\begin{aligned} S_n(\mathcal{A}, b\mathcal{L} + c) &= \sup_{\ell_1 \in \mathcal{L}} \mathbb{E}_{\epsilon_1} \cdots \sup_{\ell_n \in \mathcal{L}} \mathbb{E}_{\epsilon_n} \sup_a \sum_{t=1}^n \epsilon_t (b\ell_t(a) + c) \\ &= |b| \sup_{\ell_1 \in \mathcal{L}} \mathbb{E}_{\epsilon_1} \cdots \sup_{\ell_n \in \mathcal{L}} \mathbb{E}_{\epsilon_n} \sup_a \sum_{t=1}^n \operatorname{sign}(b)\epsilon_t \ell_t(a) + c\,\mathbb{E} \sum_{t=1}^n \epsilon_t \\ &= |b| S_n(\mathcal{A}, \mathcal{L}). \end{aligned}$$

The other two inequalities are immediate from the definition. □

4.4 Linear Games

As an application of the martingale characterization of the previous section, consider prediction games with linear losses,

$$\ell_c(a) = \langle c, a \rangle.$$

Define the loss class

$$\mathcal{L} = \{\ell_c : c \in \mathcal{C}\},$$

where $\mathcal{C} \subset \mathbb{R}^d$. Let $\mathcal{A}$ be a compact subset of $\mathbb{R}^d$.

Example 22 (Prediction with expert advice)**.** *Define $\mathcal{C} = [-1, 1]^m$ and*

$$\mathcal{A} = \Delta^m := \left\{ (w_1, \ldots, w_m) : w_i \geq 0, \sum_{i=1}^m w_i = 1 \right\}.$$

This is the game of prediction with expert advice: at each round, the strategy must choose a probability distribution $a \in \mathcal{A}$ over the m experts, and it subsequently sees the vector $c \in \mathcal{C}$ of experts' losses and itself incurs a loss $\langle c, a \rangle$.

Example 23 (Online shortest path)**.** *Let $G = (V, E)$ be a graph. Define $\mathcal{A} \subseteq \{0, 1\}^E$ so that for each path p between a source s and a sink t in G, there is an a in $\mathcal{A}$ with $a_e = 1[p \text{ contains } e]$ for all $e \in E$. Each element of the set $\mathcal{C} \subseteq \mathbb{R}^E$ corresponds to a vector of delays on the edges. The aim of the prediction strategy is to choose a path, that is, a vector $a \in \mathcal{A}$, at each time t, so that the total delay incurred over n rounds is not much worse than the best single path.*

Theorem 24. *For the linear loss class $\mathcal{L} = \{\ell_c(a) = \langle c, a \rangle : c \in \mathcal{C}\}$, with $\mathcal{C}$ a symmetric subset of $\mathbb{R}^d$ and $\mathcal{A}$ a compact subset of $\mathbb{R}^d$, the regret satisfies*

$$\sup_{\{Z_t\} \in \mathcal{M}_\mathcal{C}} \mathbb{E} \sup_{a \in \mathcal{A}} \langle Z_n, a \rangle \leq V_n(\mathcal{A}, \mathcal{L}) \leq 2 \sup_{\{Z_t\} \in \mathcal{M}_\mathcal{C}} \mathbb{E} \sup_{a \in \mathcal{A}} \langle Z_n, a \rangle,$$

where $\mathcal{M}_\mathcal{C}$ is the set of martingales with differences in $\mathcal{C}$.

It is natural that the supremum of these linear functions should depend on the sizes of elements of $\mathcal{C}$ and $\mathcal{A}$. Suppose that we have a norm $\|\cdot\|$ defined on $\mathcal{C}$, and we measure the size of elements of $\mathcal{C}$ using this norm. Then we can view $\mathcal{A}$ as a subset of the dual space of linear functions on $\mathcal{C}$, equipped with the dual norm

$$\|a\|_* = \sup \{\langle c, a \rangle : c \in \mathcal{C},\ \|c\| \leq 1\}.$$

To allow a little more flexibility, rather than measuring the size of elements of $\mathcal{A}$ using this dual norm, we will use a strongly convex function defined on $\mathcal{A}$. Think of this as a generalization of the squared dual norm.

Definition 25. *We say that $\Phi : \mathcal{A} \to \mathbb{R}$ is* σ-strongly convex wrt $\|\cdot\|_*$ *if for all $a, b \in \mathcal{A}$ and $\alpha \in [0, 1]$,*

$$\Phi(\alpha a + (1 - \alpha)b) \leq \alpha\Phi(a) + (1 - \alpha)\Phi(b) - \frac{\sigma}{2}\alpha(1 - \alpha)\|a - b\|_*^2.$$

Example 26. *Consider the p-norm*

$$\|c\|_p := \left(\sum_i |c_i|^p \right)^{1/p},$$

for $1 < p < \infty$, with dual

$$\|a\|_q := \left(\sum_i |c_i|^q \right)^{1/q},$$

where $1/p + 1/q = 1$. Then $\Phi(a) := \|a\|_q^2$ is $2(q-1)$-strongly convex wrt $\|\cdot\|_q$.

Theorem 27. *Consider the class of linear losses $\{\ell(a) = \langle c, a\rangle : c \in \mathcal{C}\}$ defined on $\mathcal{A}$. Fix a norm $\|\cdot\|$ on $\mathcal{C}$ and its dual norm $\|\cdot\|_*$ on $\mathcal{A}$. Suppose $\Phi : \mathcal{A} \to \mathbb{R}$ is σ-strongly convex wrt $\|\cdot\|_*$. Define*

$$C := \sup_{c \in \mathcal{C}} \|c\|, \qquad A^2 := \sup_{a \in \mathcal{A}} \Phi(a) - \inf_{a \in \mathcal{A}} \Phi(a).$$

Then the optimal regret satisfies

$$V_n(\mathcal{A}, \mathcal{L}) \le 2 \sup_{\{Z_t\} \in \mathcal{M}_{\mathcal{C} \cup -\mathcal{C}}} \mathbb{E} \sup_{a \in \mathcal{A}} \langle Z_n, a\rangle \le 2AC\sqrt{\frac{2n}{\sigma}}.$$

Example 28. *If $\mathcal{C}$ is the p-norm ball and $\mathcal{A}$ is the q-norm ball,*

$$\mathcal{C} = B_p := \{a : \|a\|_p \le 1\},$$
$$\mathcal{A} = B_q := \{a : \|a\|_q \le 1\}$$

for $1 < p < \infty$ and $1/p + 1/q = 1$, then

$$V_n(\mathcal{A}, \mathcal{L}) \le 2\sqrt{\frac{n}{q-1}} = 2\sqrt{n(p-1)}.$$

Notice that we cannot apply this result to bound the optimal regret for the game of prediction with expert advice, where $\mathcal{C}$ is an ∞-ball and $\mathcal{A}$ is contained in a 1-ball. Instead, we will use a suitable strongly convex function Φ.

Example 29. *Suppose $\mathcal{C}$ is the ∞-norm ball and $\mathcal{A}$ is the simplex,*

$$\mathcal{C} = B_\infty := [-1, 1]^d,$$
$$\mathcal{A} = \Delta^d := \left\{ a \in [0,1]^d : \sum_{i=1}^d a_i = 1 \right\}.$$

The infinity-norm $\|\cdot\|_\infty$,

$$\|c\|_\infty := \max_i |c_i|,$$

has the 1-norm,

$$\|a\|_1 := \sum_{i=1}^d |c_i|,$$

as its dual. If we define

$$\Phi(a) := \log d + \sum_{i=1}^d a_i \log a_i$$

(where we set $0 \log 0 = 0$), then it is easy to see that $0 \le \Phi(a) \le \log d$ for $a \in \mathcal{A}$, and straightforward to show that Φ is 1-strongly convex wrt $\|\cdot\|_1$. Thus,

$$V_n(\mathcal{A}, \mathcal{L}) \le \sqrt{2n \log d}.$$

4.4.1 Proof of Theorem 27

The proof uses ideas from convex analysis. The *convex conjugate* of Φ is

$$\Phi^*(c) = \sup_{a \in \mathcal{A}} \left(\langle c, a\rangle - \Phi(a) \right).$$

Without loss of generality, assume that $\inf_{a \in \mathcal{A}} \Phi(a) = 0$, so that $\sup_{a \in \mathcal{A}} \Phi(a) = A^2$. Then $\Phi^*(0) = 0$.

The *Fenchel-Young inequality* follows from the definition of the convex conjugate: for any $a \in \mathcal{A}$,

$$\Phi^*(c) + \Phi(a) \ge \langle c, a\rangle.$$

Thus, for any $\lambda > 0$,

$$\mathbb{E} \sup_{a \in \mathcal{A}} \langle a, Z_n \rangle \leq \mathbb{E} \sup_{a \in \mathcal{A}} \frac{1}{\lambda} \left(\Phi(a) + \Phi^*(\lambda Z_n) \right) \leq \frac{A^2}{\lambda} + \frac{\mathbb{E}\, \Phi^*(\lambda Z_n)}{\lambda}. \tag{6}$$

Consider the evolution of $\Phi^*(\lambda Z_t)$. It is straightforward to show that when Φ is σ-strongly convex wrt $\|\cdot\|_*$, then Φ^* is differentiable and σ-smooth wrt $\|\cdot\|$, that is, for all $c, d \in \mathcal{C}$,

$$\Phi^*(c + d) \leq \Phi^*(c) + \langle \nabla \Phi^*(c), d \rangle + \frac{1}{2\sigma} \|d\|^2.$$

Because of this,

$$\begin{aligned} \mathbb{E}\left[\Phi^*(\lambda Z_t) | Z_1, \ldots, Z_{t-1}\right] &\leq \Phi^*(\lambda Z_{t-1}) + \mathbb{E}\left[\langle \nabla \Phi^*(\lambda Z_{t-1}), Z_t - Z_{t-1} \rangle \, | Z_1, \ldots, Z_{t-1}\right] \\ &\quad + \frac{1}{2\sigma} \mathbb{E}\left[\lambda^2 \|Z_t - Z_{t-1}\|^2 | Z_1, \ldots, , Z_{t-1}\right] \\ &\leq \Phi^*(\lambda Z_{t-1}) + \frac{\lambda^2 C^2}{2\sigma}. \end{aligned}$$

Since $\Phi^*(0) = 0$, we have $\Phi^*(\lambda Z_n) \leq n\lambda^2 C^2/(2\sigma)$.

Substituting into Equation (6),

$$\begin{aligned} \mathbb{E} \sup_{a \in \mathcal{A}} \langle a, Z_n \rangle &\leq \frac{A^2}{\lambda} + \frac{n\lambda C^2}{2\sigma} \\ &= 2AC\sqrt{\frac{n}{2\sigma}}, \end{aligned}$$

for $\lambda = \sqrt{2A^2\sigma/(nC^2)}$.

5 Bibliographic notes

Prediction in adversarial settings has a long history across many communities, including game theory [15], information theory [11, 29], computer science [21, 28], and statistics [14]. The classic text of Cesa-Bianchi and Lugosi [10] gives an excellent coverage of much of the material in Section 2, as well as many other aspects of prediction in game-theoretic settings. Lemma 3 is due to Hoeffding [16]; the elegant proof in terms of properties of exponential families was communicated by David Pollard. Inequality (1), which improves on Theorem 2 when L_n^* is small, is from [21].

The analysis techniques for probabilistic settings are classical. For text-book treatments in terms of combinatorial dimensions and covering numbers, see, for example, [12, 2]. The treatment given here, working with Rademacher averages and bounds in terms of other complexity parameters, follows the approach in [4]. Rademacher averages were introduced as complexity parameters in [18] and [5] independently; see also [8]. Theorem 12 uses the same ideas as a corresponding result (for the case of $L_n^* = 0$) from [6]. For more refined conversions from online to batch problems, see, for example, [9].

The ideas and results of Section 4 are from [1], including the application (Theorems 13 and 14) of the minmax theorem to express the value of the game in terms of the difference between conditional risk minimizers and empirical risk minimizers. The proof of Theorem 18 was also in [1] (although the conference version contained a mistake: Rademacher averages in place of sequential Rademacher averages). The name "sequential Rademacher averages" was introduced in [25]. The example of linear games in Section 4.4 uses ideas due to Kakade *et. al.* [17], together with the observation that their elegant analysis of Rademacher averages of linear classes extends immediately to the case of sequential Rademacher averages. Since these lectures were given, there have been several advances. The supremum of the martingale process 5 that is central to the proof of Theorem 18 has been further studied in [26], and its uniform convergence related to sequential versions of combinatorial dimensions and covering numbers. All of these results give information about the value of the game, but do not immediately lead to algorithms. However, by decomposing this value into the current contribution and the value-to-go, dynamic programming immediately gives a strategy, albeit one that is typically computationally intractable. By considering bounds on the value-to-go, [24] show how to derive many algorithms for regret minimization. In some cases, the value-to-go can be calculated efficiently as a function of a succinct summary of the history; examples include prediction in a Euclidean space [20], fixed design linear regression [7], and online time series prediction [19].

References

[1] J. Abernethy, A. Agarwal, P. L. Bartlett, and A. Rakhlin, *A stochastic view of optimal regret through minimax duality*, in Proceedings of the 22nd Annual Conference on Learning Theory – COLT 2009, June 2009, pp. 257–266.

[2] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, 1999.

[3] P. Auer, N. Cesa-Bianchi, and C. Gentile, *Adaptive and self-confident on-line learning algorithms*, Journal of Computer and System Sciences, 64 (2002), pp. 48–75.

[4] P. L. Bartlett, *CS281B/Stat241B: Statistical learning theory lectures*, 2003. http://www.cs.berkeley.edu/∼bartlett/281B.

[5] P. L. Bartlett, S. Boucheron, and G. Lugosi, *Model selection and error estimation*, in Proceedings of the Thirteenth Annual Conference on Computational Learning Theory, 2000, pp. 286–297.

[6] P. L. Bartlett, P. Fischer, and K.-U. Höffgen, *Exploiting random walks for learning*, in Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory, ACM Press, 1994, pp. 318–327.

[7] P. L. Bartlett, W. Koolen, A. Malek, E. Takimoto, and M. Warmuth, *Minimax fixed-design linear regression*, in Proceedings of the Conference on Learning Theory (COLT2015), vol. 40, June 2015.

[8] P. L. Bartlett and S. Mendelson, *Rademacher and Gaussian complexities: Risk bounds and structural results*, Journal of Machine Learning Research, 3 (2002), pp. 463–482.

[9] N. Cesa-Bianchi, A. Conconi, and C. Gentile, *On the generalization ability of on-line learning algorithms*, IEEE Transactions on Information Theory, 50 (2004), pp. 2050–2057.

[10] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*, Cambridge University Press, 2006.

[11] T. Cover, *Behavior of sequential predictors of binary sequences*, in Proceedings of the 4th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, 1965, pp. 263–272.

[12] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*, Applications of Mathematics: Stochastic Modelling and Applied Probability (31), Springer, 1996.

[13] R. M. Dudley, *Real Analysis and Probability*, Wadsworth & Brooks/Cole, California, 1989.

[14] D. P. Foster, *Prediction in the worst case*, Annals of Statistics, 19 (1991), pp. 1084–1090.

[15] J. Hannan, *Approximation to Bayes risk in repeated play*, Contributions to the Theory of Games, 3 (1957), pp. 97–139.

[16] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association, 58 (1963), pp. 13–30.

[17] S. M. Kakade, K. Sridharan, and A. Tewari, *On the complexity of linear prediction: Risk bounds, margin bounds, and regularization*, in Advances in Neural Information Processing Systems 21, MIT Press, 2009, pp. 793–800.

[18] V. Koltchinskii, *Rademacher penalties and structural risk minimization*, IEEE Transactions on Information Theory, 47 (2001), pp. 1902–1914.

[19] W. Koolen, A. Malek, P. L. Bartlett, and Y. Abbasi-Yadkori, *Minimax time series prediction*, in Advances in Neural Information Processing Systems 28, 2015. To appear.

[20] W. M. Koolen, A. Malek, and P. L. Bartlett, *Efficient minimax strategies for square loss games*, in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds., 2014, pp. 3230–3238.

[21] N. Littlestone and M. Warmuth, *Weighted majority algorithm*, in IEEE Symposium on Foundations of Computer Science, 1989, pp. 256–261.

[22] C. McDiarmid, *On the method of bounded differences*, in Surveys in Combinatorics 1989, Cambridge University Press, 1989, pp. 148–188.

[23] S. Mendelson, *A few notes on statistical learning theory*, in Advanced Lectures in Machine Learning, S. Mendelson and A. J. Smola, eds., vol. 2600 of Lecture Notes in Computer Science, Springer, 2003, pp. 1–40.

[24] A. Rakhlin, O. Shamir, and K. Sridharan, *Relax and randomize : From value to algorithms*, in Advances in Neural Information Processing Systems 25, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds., 2012, pp. 2141–2149.

[25] A. Rakhlin, K. Sridharan, and A. Tewari, *Online learning: Random averages, combinatorial parameters, and learnability*, in Advances in Neural Information Processing Systems 23, 2010.

[26] ———, *Sequential complexities and uniform martingale laws of large numbers*, Probability Theory and Related Fields, 161 (2015), pp. 111–153.

[27] M. Sion, *On general minimax theorems.*, Pacific Journal of Mathematics, 8 (1958), pp. 171–176.

[28] V. Vovk, *Aggregating strategies*, in Proceedings of the Third Annual Workshop on Computational Learning Theory, Morgan Kaufmann, 1990, pp. 372–383.

[29] J. Ziv, *Coding theorems for individual sequences*, IEEE Transactions on Information Theory, 24 (1978), pp. 405–412.