Lectures on prediction of individual sequences

Presented at
2001, a Statistics Odyssey
Centre Émile Borel, Institut Henri Poincaré

Gábor Lugosi
Department of Economics,
Pompeu Fabra University
Ramon Trias Fargas 25-27,
08005 Barcelona, Spain,
lugosi@upf.es
http://www.econ.upf.es/~lugosi

May 30, 2001

Contents

1 Sequential prediction

1.1 Introduction

Imagine that we are asked to predict a hidden sequence $y_1, y_2, \ldots$, of zeroes and ones such that at each time instant $t \geq 1$, after making our guess $\widehat{p}_t \in [0,1]$, the true bit y_t is revealed so that in forming our prediction, we are allowed to look at the past $y_1^{t-1} = (y_1, \ldots, y_{t-1})$ (here y_1^0 denotes the empty string). Assume that our goal is to minimize the cumulative loss $\sum_{t=1}^n |\widehat{p}_t - y_t|$.

The traditional approach in statistics to such problems first assumes a probabilistic model for the generating mechanism of the bit sequence and then investigates the possibilities and limitations of the prediction of such random sequences. For example, in many applications the sequence is assumed to be a realization of some stationary process. This approach works in many cases when a tractable statistical model reasonably describes the underlying process.

However, there exist situations when any statistical model is doomed to failure and more robust prediction methods are required. Typical examples of hard-to-model processes emerge, for example, in financial processes or in internet data streams.

The purpose of this course is to describe some techniques of such robust prediction and discuss the possibilities, limitations, and difficulties.

1.2 Prediction using expert advice

Without assumptions on the sequence to be predicted, it is not immediately clear how the problem can be made meaningful. One possibility is to compare the predictor's performance to the best of a certain class of reference predictors, often called *experts*. The goal is to guarantee that the total loss of the predictor is not much more than the best loss among the reference predictors. Next we describe such a typical setup in sequential prediction problems.

The decision maker is a predictor whose goal is to forecast a hidden sequence $y_1, y_2, \ldots$ of elements in some *outcome space* $\mathcal{Y}$. At each time t, the predictor computes its guess $\widehat{p}_t \in \mathcal{X}$ for the next outcome y_t where $\mathcal{X}$ is a *decision space* which we assume to be a subset of a vector space. Before forming the prediction $\widehat{p}_t$, the predictor has access to $f_{1,t}, \ldots, f_{N,t} \in \mathcal{X}$ which we interpret as the predictions of y_t made by N reference predictors, or *experts*. The guesses of the predictor and the experts are then scored using a *loss function* $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. The predictor's goal is to keep as small as possible the *cumulative regret* $R_{i,n}$ with respect to each expert. This quantity is defined, for expert i, by the sum

$$R_{i,n} = \sum_{t=1}^{n} \left(\ell(\widehat{p}_t, y_t) - \ell(f_{i,t}, y_t)\right) = \widehat{L}(y_1^n) - L_i(y_1^n)$$

where $\widehat{L}(y_1^n) = \sum_{t=1}^n \ell(\widehat{p}_t, y_t)$ denotes the *cumulative loss* of the predictor and $L_i(y_1^n)$ is the cumulative loss of expert i. We also define the regret with respect to expert i at time t by $r_{i,t} = \ell(\widehat{p}_t, y_t) - \ell(f_{i,t}, y_t)$. We may think about $r_{i,t}$ as the regret the predictor feels of not having listened to the advice of expert i right after the t-th symbol y_t has been revealed. Thus, $R_{i,n} = \sum_{t=1}^n r_{i,t}$.

The goal of the predictor is to achieve that regardless of the outcome sequence,

$$\max_{i=1,\ldots,N} R_{i,n} = \widehat{L}(y_1^n) - \min_{i=1,\ldots,N} L_i(y_1^n) = o(n) \ .$$

In the next section we show that this ambitious goal may be achieved under very mild conditions by a simple predictor.

1.3 Weighted average prediction

The basic idea behind weighted average predictors is very simple: those experts that have suffered a smaller loss up to the moment of prediction are to be trusted more than those with large cumulative losses. Let $\phi : \mathbb{R} \to \mathbb{R}^+$ be a monotonically increasing twice differentiable convex function. The predictor uses ϕ' (the derivative of ϕ) to determine the weight assigned to the i-th expert and forms a *weighted average* of the experts' predictions

$$\widehat{p}_t = \frac{\sum_{i=1}^N \phi'(R_{i,t-1}) f_{i,t}}{\sum_{i=1}^N \phi'(R_{i,t-1})} \ . \tag{1}$$

First of all note that $\widehat{p}_t$ is a legitimate predictor since it only uses the experts' advices at time t and cumulative regrets based on the past. Since ϕ' is assumed to be nonnegative and nondecreasing, the weighted average predictor assigns more weight to those experts whose cumulative regret is large, that is, those that have performed better in the past.

We start the analysis of weighted average predictors by a simple technical observation.

Lemma 1 *Assume that the loss function ℓ is convex in its first argument. Then*

$$\sup_{y_t \in \mathcal{Y}} \sum_{i=1}^{N} r_{i,t} \phi'(R_{i,t-1}) \le 0 \ .$$

Proof. By convexity of the loss function, for all $y \in \mathcal{Y}$,

$$\ell(\widehat{p}_t, y) = \ell\left(\frac{\sum_{i=1}^N \phi'(R_{i,t-1}) f_{i,t}}{\sum_{i=1}^N \phi'(R_{i,t-1})}, y \right) \le \frac{\sum_{i=1}^N \phi'(R_{i,t-1}) \ell(f_{i,t}, y)}{\sum_{i=1}^N \phi'(R_{i,t-1})} \ .$$

Rearranging, we obtain the statement. ∎

The simple observation of the lemma above allows us to interpret the weighted average predictor in a different way. To do this, introduce the *regret vector*

$$\mathbf{r}_t = (r_{1,t}, \ldots, r_{N,t}) \in \mathbb{R}^N$$

and the corresponding *cumulative regret vector* $\mathbf{R}_n = \sum_{t=1}^n \mathbf{r}_t$. Then, introducing the *potential function* $\Phi : \mathbb{R}^n \to \mathbb{R}^+$ as

$$\Phi(\mathbf{u}) = \sum_{i=1}^{N} \phi(u_i)$$

we may rewrite the statement of the lemma as

$$\sup_{y_t \in \mathcal{Y}} \mathbf{r}_t \cdot \nabla \Phi(\mathbf{R}_{t-1}) \le 0 \ . \tag{2}$$

(The notation $\mathbf{u} \cdot \mathbf{v}$ stands for the the inner product of two vectors defined by $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \ldots + u_N v_N$.) Condition (2) shows that weighted average predictors tend to keep the point $\mathbf{R}_t$ as close as possible to the minimum of the potential by forcing the regret vector to point away from the gradient of the potential no matter what the outcome y_t may be.

This property, in fact, suggests a simple analysis since the increments of the potential function may now be easily bounded by Taylor's theorem. Actually, bounding the increments of $f(\Phi(\mathbf{R}_t))$ for a suitably chosen increasing twice differentiable function $f : \mathbb{R}^+ \to \mathbb{R}^+$ leads to better bounds. Note that f has a role in the analysis but not in the definition of the predictor.

The next general theorem will imply several interesting bounds for different versions of the weighted average predictor.

Theorem 1 *Assume that a predictor satisfies condition (2). Let* $f : \mathbb{R}^+ \to \mathbb{R}^+$ *be an increasing, concave, and twice differentiable auxiliary function such that*

$$\sup_{\mathbf{u} \in \mathbb{R}^N} f'(\Phi(\mathbf{u})) \sum_{i=1}^N \phi''(u_i) r_{i,t}^2 \le C(\mathbf{r}_t)$$

for some nonnegative function $C : \mathbb{R}^N \to \mathbb{R}^+$*. Then, for all* $n = 1, 2, \ldots,$

$$f(\Phi(\mathbf{R}_n)) \le f(\Phi(\mathbf{0})) + \frac{1}{2} \sum_{t=1}^n C(\mathbf{r}_t) \ .$$

Proof. We estimate $f(\Phi(\mathbf{R}_t))$ in terms of $f(\Phi(\mathbf{R}_{t-1}))$ using Taylor's theorem. Note that $\nabla f(\Phi(\mathbf{R}_{t-1})) = f'(\Phi(\mathbf{R}_{t-1})) \nabla \Phi(\mathbf{R}_{t-1})$. Thus, we obtain

$$\begin{aligned}
& f(\Phi(\mathbf{R}_t)) = f(\Phi(\mathbf{R}_{t-1} + \mathbf{r}_t)) \\
&= f(\Phi(\mathbf{R}_{t-1})) + f'(\Phi(\mathbf{R}_{t-1}))\, \nabla \Phi(\mathbf{R}_{t-1}) \cdot \mathbf{r}_t + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \left. \frac{\partial^2 f(\Phi)}{\partial u_i \partial u_j} \right|_{\xi} r_{i,t} r_{j,t} \\
& \quad \text{(where } \xi \text{ is some vector in } \mathbb{R}^N\text{)} \\
&\le f(\Phi(\mathbf{R}_{t-1})) + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \left. \frac{\partial^2 f(\Phi)}{\partial u_i \partial u_j} \right|_{\xi} r_{i,t} r_{j,t}
\end{aligned}$$

where the inequality follows by (2) and the fact that $f' \geq 0$. Now straightforward calculation shows that

$$
\begin{aligned}
& \sum_{i=1}^{N}\sum_{j=1}^{N} \left.\frac{\partial^2 f(\Phi)}{\partial u_i \partial u_j}\right|_{\xi} r_{i,t} r_{j,t} \\
= \; & f''\left(\Phi(\xi)\right) \sum_{i=1}^{N}\sum_{j=1}^{N} \phi'(\xi_i)\phi'(\xi_j) r_{i,t} r_{j,t} + f'\left(\Phi(\xi)\right) \sum_{i=1}^{N} \phi''(\xi_i) r_{i,t}^2 \\
= \; & f''\left(\Phi(\xi)\right) \left(\sum_{i=1}^{N} \phi'(\xi_i) r_{i,t} \right)^2 + f'\left(\Phi(\xi)\right) \sum_{i=1}^{N} \phi''(\xi_i) r_{i,t}^2 \\
\leq \; & f'\left(\Phi(\xi)\right) \sum_{i=1}^{N} \phi''(\xi_i) r_{i,t}^2 \qquad \text{(since } f \text{ is concave)} \\
\leq \; & C(\mathbf{r}_t)
\end{aligned}
$$

where at the last step we used the hypothesis of the theorem. Thus, we have obtained $f(\Phi(\mathbf{R}_t)) \leq f(\Phi(\mathbf{R}_{t-1})) + C(\mathbf{r}_t)/2$. The proof is finished by iterating the argument. ▮

Next, we review two simple applications of Theorem 1.

Corollary 1 *Consider the weighted average predictor (1) based on* $\phi(x) = x_+^p$ *where* $p \geq 2$*, that is,*

$$
\widehat{p}_t = \frac{\sum_{i=1}^{N} \left(\sum_{s=1}^{t-1} \left(\ell(\widehat{p}_s, y_s) - \ell(f_{i,s}, y_s) \right) \right)_+^{p-1} f_{i,t}}{\sum_{i=1}^{N} \left(\sum_{s=1}^{t-1} \left(\ell(\widehat{p}_s, y_s) - \ell(f_{i,s}, y_s) \right) \right)_+^{p-1}} .
$$

Assume that the loss function ℓ *is convex in its first argument and that it takes values in* $[0,1]$*. Then for any sequence* $y_1, y_2, \ldots \in \mathcal{Y}$ *of outcomes and for any* $n \geq 1$*,*

$$
\widehat{L}(y_1^n) - \min_{i=1,\ldots,N} L_i(y_1^n) \leq \sqrt{2n(p-1)N^{2/p}} .
$$

Corollary 1 shows that for any value of p,

$$
\lim_{n\to\infty} \sup_{y_1^n} \frac{1}{n} \left(\widehat{L}(y_1^n) - \min_{i=1,\ldots,N} L_i(y_1^n) \right) = 0
$$

and in all cases the left-hand side converges to zero at a speed of $n^{-1/2}$. The choice $p = 2$ yields a simple algorithm, but the dependence of the upper bound on the number of experts is suboptimal. An approximate minimization of the obtained upper bound suggests the value $p = 2\ln N$ (for $N > 2$) which leads to the bound

$$\widehat{L}(y_1^n) - \min_{i=1,\dots,N} L_i(y_1^n) \le \sqrt{2ne(2\ln N - 1)}\ .$$

Proof of Corollary 1. Apply Theorem 1 with $f(x) = x^{2/p}$, $x \ge 0$. Clearly,

$$f'(x) = \frac{2}{px^{(p-2)/p}} \quad \text{and} \quad \phi''(x) = p(p-1)x_+^{p-2}.$$

By Hölder's inequality,

$$\begin{aligned}\sum_{i=1}^N \phi''(u_i)r_{i,t}^2 &= p(p-1)\sum_{i=1}^N (u_i)_+^{p-2} r_{i,t}^2 \\ &\le p(p-1)\left(\sum_{i=1}^N \left((u_i)_+^{p-2}\right)^{p/(p-2)}\right)^{(p-2)/p} \left(\sum_{i=1}^N |r_{i,t}|^p\right)^{2/p}.\end{aligned}$$

Thus,

$$f'\left(\sum_{i=1}^N \phi(u_i)\right)\sum_{i=1}^N \phi''(u_i) \le 2(p-1)\left(\sum_{i=1}^N |r_{i,t}|^p\right)^{2/p}$$

and therefore the conditions of Theorem 1 are satisfied with $C(\mathbf{r}_t) = (p-1)\,\|\mathbf{r}_t\|_p^2$. Since $\Phi(\mathbf{0}) = 0$, Theorem 1, together with the boundedness of the loss function implies that

$$\|\mathbf{R}_n\|_p^2 = f(\Phi(\mathbf{R}_n)) \le 2(p-1)\sum_{t=1}^n \|\mathbf{r}_t\|_p^2 \le 2n(p-1)N^{2/p}\ .$$

Since

$$\widehat{L}(y_1^n) - \min_{i=1,\dots,N} L_i(y_1^n) = \max_{1\le i\le N} R_{i,n} \le \|\mathbf{R}_n\|_p\ ,$$

the result follows. ∎

Our second main example is the *exponentially weighted average* predictor which is predictor (1) based on $\phi(x) = e^{\eta x}$ where η is a positive parameter. In this case the weighted average predictor simplifies as

$$\widehat{p}_t = \frac{\sum_{i=1}^N e^{\eta \sum_{s=1}^{t-1}\left(\ell(\widehat{p}_s,y_s)-\ell(f_{i,s},y_s)\right)} f_{i,t}}{\sum_{i=1}^N e^{\eta \sum_{s=1}^{t-1}\left(\ell(\widehat{p}_s,y_s)-\ell(f_{i,s},y_s)\right)}} = \frac{\sum_{i=1}^N e^{-\eta \sum_{s=1}^{t-1}\ell(f_{i,s},y_s)} f_{i,t}}{\sum_{i=1}^N e^{-\eta \sum_{s=1}^{t-1}\ell(f_{i,s},y_s)}} \ . \qquad (3)$$

The beauty of the exponentially weighted average predictor is that it only depends on the past performance of the experts, whereas the predictions made using other general potentials depend on the past predictions $\widehat{p}_s$, $s < t$ as well. A simple application of Theorem 1 reveals the following performance bound for the exponentially weighted average predictor.

Corollary 2 *Assume that the loss function ℓ is convex in its first argument and that it takes values in $[0,1]$. For any n and $\eta > 0$, and for all $y_1, \ldots, y_n \in \mathcal{Y}$, the cumulative loss of the exponentially weighted average predictor satisfies*

$$\widehat{L}(y_1^n) - \min_{i=1,\ldots,N} L_i(y_1^n) \le \frac{\ln N}{\eta} + \frac{n\eta}{2} \ .$$

Optimizing the upper bound suggests the choice $\eta = \sqrt{2\ln N/n}$. In this case the upper bound becomes $\sqrt{2n\ln N}$, which is slightly better than the best bound we obtained using $\phi(x) = x_+^p$ with $p = 2\ln N$. The disadvantage of the exponential weighting is that optimal tuning of the parameter η requires knowledge of the horizon n in advance. In the next section we describe a version of the exponentially weighted average predictor which does not suffer from this drawback.

Proof of Corollary 2. We use Theorem 1 with $f(x) = (1/\eta)\ln x$. Then clearly

$$f'\left(\Phi(\mathbf{u})\right) \sum_{i=1}^N \phi''(u_i) r_{i,t}^2 \le \eta \max_{1\le i\le N} r_{i,t}^2 \le \eta \ .$$

Using $\Phi(\mathbf{0}) = N$, Theorem 1 implies

$$\max_{1\le i\le N} R_{i,n} \le \frac{1}{\eta}\ln \Phi(\mathbf{R}_n) \le \frac{\ln N}{\eta} + \frac{n\eta}{2}$$

as desired. ∎

1.4 The doubling trick

As we pointed out in the previous section, the exponentially weighted average predictor has the disadvantage that the optimal choice of the parameter η depends on the length of the sequence to be predicted and therefore it cannot be used efficiently in "unbounded horizon" prediction, that is, when the total length of the sequence to be predicted is not known in advance. The purpose of this section is to present a simple modification (using the so-called the "doubling trick") with similar performance guarantees which does not rely on previous knowledge of the total length n.

The idea is to partition time into periods of exponentially increasing lengths. Then in each period the exponentially weighted average predictor is used with a parameter η chosen optimally for the length of the interval.

Formally, for each time instance t with $t \in \{2^m, \ldots, 2^{m+1}-1\}$ (where $m = 0, 1, 2, \ldots$) the predictor uses the exponentially weighted average predictor initiated at time 2^m with parameter $\eta_m = \sqrt{2^{m-1} \ln N}$. Thus, the algorithm is reset at each time instance which is an integer power of two and is restarted with a new value of η. Then for each n, denoting $\overline{n} = 2^{\lfloor \log_2 n \rfloor + 1}$, for the cumulative loss $\widehat{L}(y_1^n)$ of this modified predictor, we have

$$
\begin{aligned}
\widehat{L}(y_1^n) &= \sum_{m=0}^{\lfloor \log_2 n \rfloor - 1} \widehat{L}(y_{2^m}^{2^{m+1}-1}) + \widehat{L}(y_{\overline{n}/2}^{n}) \\
&\le \sum_{m=0}^{\lfloor \log_2 n \rfloor - 1} \left(\min_{i=1,\ldots,N} L_i(y_{2^m}^{2^{m+1}-1}) + \sqrt{2^{m+1} \ln N} \right) \\
&\qquad + \left(\min_{i=1,\ldots,N} L_i(y_{\overline{n}/2}^{n}) + \sqrt{\overline{n} \ln N} \right) \quad \text{(by Corollary 2)} \\
&\le \min_{i=1,\ldots,N} L_i(y_1^n) + \sum_{m=0}^{\lfloor \log_2 n \rfloor} \sqrt{2^{m+1} \ln N} \\
&\le \min_{i=1,\ldots,N} L_i(y_1^n) + \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{n \ln N} \, .
\end{aligned}
$$

Thus, we have derived that the modified weighted average predictor is such that for any sequence $y_1, y_2, \ldots \in \mathcal{Y}$ of outcomes and for any $n \geq 1$,

$$\widehat{L}(y_1^n) - \min_{i=1,\ldots,N} L_i(y_1^n) \leq \frac{\sqrt{2}}{\sqrt{2}-1}\sqrt{n \ln N}\ .$$

1.5 Randomized prediction

The results of the previous sections crucially build on the assumption that the loss function ℓ is convex in its first argument. While this assumption is natural in many applications, it is not satisfied in some important examples.

One such example is the case when both the decision space and the outcome space are $\{0,1\}$ and the loss function is $\ell(x,y) = \mathbb{I}_{[x \neq y]}$. In this case it is clear that for any deterministic strategy of the predictor, there exists an outcome sequence y_1^n such that the predictor errs at every single time instant, that is, $\widehat{L}(y_1^n) = n$. Thus, except for trivial cases, it is impossible to achieve that $\sup_{y_1^n}\left(\widehat{L}(y_1^n) - \min_{i \leq N} L_i(y_1^n)\right) = o(n)$. For example, if $N = 2$ and the two experts' predictions are $f_{1,t} = 0$ and $f_{2,t} = 1$ for all t, then $\min_{i \leq N} L_i(y_1^n) \leq n/2$ and no matter what the predictor does,

$$\sup_{y_1^n \in \mathcal{Y}^n}\left(\widehat{L}(y_1^n) - \min_{i \leq N} L_i(y_1^n)\right) \geq \frac{n}{2}.$$

In such a situation randomization may help. Next we describe a version of the sequential prediction problem in which the predictor has access, at each time instant, to an independent random variable uniformly distributed on the interval $[0,1]$.

Assume that there are N experts, and that at time t, the prediction of the i-th expert is $f_{i,t}$. The predictor computes a probability distribution $\mathbf{w}_t = (w_{1,t}, \ldots, w_{N,t})$ and predicts as expert i with probability $w_{i,t}$, independently of past predictions. Formally,

$$\widehat{p}_t = \sum_{i=1}^{N} f_{i,t}\mathbb{I}_{[Z_t = i]},$$

where $\mathbb{P}[Z_t = i] = w_{i,t}$. After the t-th outcome symbol is revealed, the expert suffers loss $\ell(f_{i,t}, y_t)$ and the predictor suffers the (random) loss

$\tilde{\ell}(\widehat{p}_t, y_t) = \sum_{i=1}^N \ell(f_{i,t}, y_t)\mathbb{I}_{[Z_t=i]}$. Here, for simplicity, we only consider the *expected loss*

$$\overline{\ell}(\widehat{p}_t, y_t) = \mathbb{E}\tilde{\ell}(\widehat{p}_t, y_t) = \sum_{i=1}^N \ell(f_{i,t}, y_t)\mathbb{P}[Z_t = i] \ .$$

Thus, the predictor's goal is to minimize the excess expected loss with respect to the best expert, that is, to make

$$\sup_{y_1^n \in \mathcal{Y}} \left(\overline{L}(y_1^n) - \min_{i=1\dots,N} L_i(y_1^n) \right) = \sup_{y_1^n \in \mathcal{Y}} \left(\sum_{t=1}^n \overline{\ell}(\widehat{p}_t, y_t) - \min_{i=1\dots,N} \sum_{t=1}^n \ell(f_{i,t}, y_t) \right)$$

as small as possible. Denote the regret with respect to the i-th expert by $r_{i,t} = \overline{\ell}(\widehat{p}_t, y_t) - \ell(f_{i,t}, y_t)$ and, as before, let $\mathbf{r}_t = (r_{1,t}, \dots, r_{N,t})$ and $\mathbf{R}_t = \sum_{s=1}^t \mathbf{r}_t$.

The analogue of the weighted average predictor defined in Section 1.3 is a predictor which assigns a larger probability to those experts whose cumulative loss, up to the time of prediction, has been smaller. Thus, given a nonnegative increasing and convex twice differentiable function $\phi : \mathbb{R} \to \mathbb{R}^+$, we may define the prediction rule

$$w_{i,t} = \frac{\phi'(R_{i,t-1})}{\sum_{j=1}^N \phi'(R_{j,t-1})} \ .$$

Then, defining once again

$$\Phi(\mathbf{u}) = \sum_{i=1}^N \phi(u_i)$$

it is immediate to see that

$$\mathbf{r}_t \cdot \nabla\Phi(\mathbf{R}_{t-1}) = 0 \ .$$

Hence, condition (2) is satisfied with equality and therefore the analysis of Section 1.3 may be repeated. In particular, if the loss function is bounded between zero and one then choosing $\phi(x) = x_+^p$ (with $p \geq 2$) and $\phi(x) = e^{\eta x}$, the respective performance bounds of Corollaries 1 and 2 hold in this new setup as well. Note that in the case of randomized prediction, apart from its boundedness, there is no need to assume anything about the loss function.

1.6 Exp-concave loss functions

Corollaries 1 and 2 show that, under very general assumptions—only convexity and boundedness of the loss function is required—it is possible to achieve that the excess cumulative loss, with respect to the best expert $\widehat{L}(y_1^n) - \min_{i\le N} L_i(y_1^n)$ be of the order of $\sqrt{n \ln N}$ for all possible outcome sequences. However, there is an important class of loss functions for which even better bounds can be achieved. This is summarized in the next result.

Theorem 2 *Consider the exponentially weighted majority predictor (3). Assume that the loss function* ℓ *is such that the function* $F(z) = e^{-\eta\ell(z,y)}$ *is concave for all* $y \in \mathcal{Y}$*, then for all* $n \ge 1$*,*

$$\sup_{y_1^n} \left(\widehat{L}(y_1^n) - \min_{i\le N} L_i(y_1^n) \right) \le \frac{\ln N}{\eta} .$$

For the loss functions satisfying the assumption of the theorem, the excess loss of the exponentially weighted average predictor is bounded by a quantity which is independent on the length n of the sequence of outcomes. Note that apart from the assumption of "exp-concavity" we do not assume anything else about the loss function. In particular, we do not explicitly assume that the loss function is convex and that it is bounded. (Note, however, that the concavity of $F(z)$ implies that ℓ is convex in its first argument, see the exercises.)

Proof. It suffices to show that $\Phi(\mathbf{R}_t) \le \Phi(\mathbf{0})$ where Φ is the exponential potential function $\Phi(\mathbf{u}) = \sum_{i=1}^N e^{\eta u_i}$ The result then follows immediately from the fact that $\Phi(\mathbf{0}) = N$. Thus, it suffices to show that the value of the potential function can never increase, that is, $\Phi(\mathbf{R}_t) \le \Phi(\mathbf{R}_{t-1})$ or, equivalently, that

$$\sum_{i=1}^N e^{-\eta\sum_{s=1}^{t-1}\ell(f_{i,s},y_s)} e^{\eta(\ell(\widehat{p}_t,y_t)-\ell(f_{i,t},y_t))} \le \sum_{i=1}^N e^{-\eta\sum_{s=1}^{t-1}\ell(f_{i,s},y_s)} .$$

Denoting $w_{i,t-1} = e^{-\eta\sum_{s=1}^{t-1}\ell(f_{i,s},y_s)}$, this may be re-written as

$$e^{-\eta\ell(\widehat{p}_t,y_t)} \ge \frac{\sum_{i=1}^N w_{i,t-1} e^{-\eta\ell(f_{i,t},y_t)}}{\sum_{i=1}^N w_{i,t-1}} .$$

But since $\widehat{p}_t = \sum_{i=1}^N w_{i,t-1} f_{i,t} \Big/ \sum_{i=1}^N w_{i,t-1}$, this follows by the concavity of $F(z)$ and Jensen's inequality. ∎

Some simple and important loss functions satisfy the exp-concavity assumption of Theorem 2:

Example. Let $\mathcal{Y} = \{0,1\}$ and $\mathcal{X} = [0,1]$ and consider the *relative entropy loss*

$$\ell(z,y) = \mathbb{I}_{[y=1]} \ln \frac{1}{z} + \mathbb{I}_{[y=0]} \ln \frac{1}{1-z} \ .$$

If we take $\eta = 1$, then

$$F(z) = e^{-\eta \ell(z,y)} = \mathbb{I}_{[y=1]} z + \mathbb{I}_{[y=0]} (1-z)$$

which is a linear function of z for both values of y and therefore it is concave. Note that this is an unbounded loss function.

Example. Consider next the quadratic loss function

$$\ell(z,y) = (z-y)^2$$

where $\mathcal{X} = \mathcal{Y} = [0,1]$. Then straightforward calculation shows that $F(z)$ is concave if and only if for all z, $(z-y)^2 \le 1/(2\eta)$. This is clearly guaranteed if $\eta \le 1/2$.

Example. Let $\mathcal{Y} = \{0,1\}$, $\mathcal{X} = [0,1]$, and consider the *absolute loss* $\ell(z,y) = |z-y|$. It is easy to see that $F(z)$ is not concave for any value of η. In fact, as we will see later, for this loss function there is no predictor whose cumulative excess loss can be bounded independently of n.

The exercises provide further examples.

1.7 An optimal bound

The purpose of this section is to show that, even for general convex loss functions, the bound of Corollary 2 may be improved for the exponentially weighted average predictor (3). The result below improves Corollary 2 by a constant factor. Later we will see that the obtained bound cannot be improved further.

Theorem 3 *Assume that the loss function ℓ is convex in its first argument and that it takes values in $[0,1]$. For any n and $\eta > 0$, and for all $y_1, \ldots, y_n \in \mathcal{Y}$, the cumulative loss of the exponentially weighted average predictor satisfies*

$$\widehat{L}(y_1^n) - \min_{i=1,\ldots,N} L_i(y_1^n) \le \frac{\ln N}{\eta} + \frac{n\eta}{8} .$$

In particular, with $\eta = \sqrt{8 \ln N / n}$, the upper bound becomes $\sqrt{(n/2) \ln N}$.

The proof is similar, in spirit to that of Corollary 2, but now, instead of bounding the evolution of $\ln\left(\sum_i e^{\eta R_{i,t}}\right)$, we consider the sum of weights

$$W_t = \sum_{i=1}^N w_{i,t} = \sum_{i=1}^N e^{-\eta \sum_{s=1}^t \ell(f_{i,s}, y_s)}$$

for $t \ge 1$, and $W_0 = N$. In the proof we use the following classical inequality due to Hoeffding [31]:

Lemma 2 *Let X be a random variable with $a \le X \le b$. Then for any $s \in \mathbb{R}$,*

$$\ln \mathbb{E}\left[e^{sX}\right] \le s\mathbb{E}X + \frac{s^2(b-a)^2}{8} .$$

The proof is in the Appendix.

Proof of Theorem 3 First observe that

$$\begin{aligned}
\ln \frac{W_n}{W_0} &= \ln\left(\sum_{i=1}^N e^{-\eta L_i(y_1^n)}\right) - \ln N \\
&\ge \ln\left(\max_{i=1,\ldots,N} e^{-\eta L_i(y_1^n)}\right) - \ln N \\
&= -\eta \min_{i=1,\ldots,N} L_i(y_1^n) - \ln N . \qquad (4)
\end{aligned}$$

On the other hand, for each $t = 1, \ldots, n$

$$\begin{aligned}
\ln \frac{W_t}{W_{t-1}} &= \ln \frac{\sum_{i=1}^N e^{-\eta \ell(f_{i,t}, y_t)} e^{-\eta L_i(y_1^{t-1})}}{\sum_{i=1}^N e^{-\eta L_i(y_1^{t-1})}} \\
&= \ln \frac{\sum_{i=1}^N w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{i=1}^N w_{i,t-1}}
\end{aligned}$$

Now using Lemma 2 we observe that the quantity above may be upper bounded by

$$
\begin{aligned}
-\eta \frac{\sum_{i=1}^N w_{i,t-1} \ell(f_{i,t}, y_t)}{\sum_{i=1}^N w_{i,t-1}} + \frac{\eta^2}{8} &\leq -\eta \ell\left(\frac{\sum_{i=1}^N w_{i,t-1} f_{i,t}}{\sum_{i=1}^N w_{i,t-1}}, y_t\right) + \frac{\eta^2}{8} \\
&= -\eta \ell(\widehat{p}_t, y, t) + \frac{\eta^2}{8}
\end{aligned}
$$

where we used the convexity of the loss function in its first argument and the definition (3) of the exponentially weighted average predictor. Summing over $t = 1, \ldots, n$ we get

$$
\ln \frac{W_n}{W_0} \leq -\eta \widehat{L}(y_1^n) + \frac{\eta^2}{8} n \ .
$$

Combining this with the lower bound (4) and solving for $\widehat{L}(y_1^n)$ we find that

$$
\widehat{L}(y_1^n) \leq \min_{i=1,\ldots,N} L_i(y_1^n) + \frac{\ln N}{\eta} + \frac{\eta}{8} n
$$

as desired. ∎

1.8 Improvement for small losses

Next we show that the worst-case bound of Theorem 3 may be significantly improved if the cumulative loss of the best expert is small. This may be achieved even without prior knowledge of the fact that one of the experts will have a small cumulative loss.

To understand why one can hope for better bounds for the cumulative regret when the cumulative loss of the best expert is small, consider the extreme case when $\mathcal{Y} = \mathcal{X} = \{0, 1\}$, $\ell(\widehat{p}, y) = |\widehat{p} - y|$, and the predictor knows in advance that one of the N experts "knows the truth", that is, $\min_{i=1,\ldots,N} L_i(y_1^n) = 0$. Then the predictor may use the following simple "vote-with-the majority" rule: at time $t = 1$ predict $\widehat{p}_1 = 0$ if and only if at least half of the N experts predict 0. After the first bit y_1 is revealed, discard all experts with $f_{i,1} \neq y_1$. At time $t = 2$ predict $\widehat{p}_2 = 0$ if and only if at least half of the N surviving experts predict 0 and discard all incorrectly predicting expert. Continuing this way, each time the predictor makes a mistake, at least half of the surviving experts are discarded (since the predictor always votes according to the majority of the surviving experts) until only one expert stays alive after which the predictor will not make any further mistake. Thus, the total number of mistakes of the predictor is at most $\log_2 N$, that is, for any $n = 1, 2, \ldots,$

$$\widehat{L}(y_1^n) = \widehat{L}(y_1^n) - \min_{i=1,\ldots,N} L_i(y_1^n) \le \log_2 N \ ,$$

a bound significantly better than the worst-case bound $\sqrt{(n/2)\ln N}$.

The purpose of this section is to "interpolate" between the extreme example shown above and the general bound of Theorem 3. Our main tool is the following result. For simplicity, we write $L^*(y_1^n) = \min_{i=1,\ldots,N} L_i(y_1^n)$.

Theorem 4 *Assume that the loss function ℓ is convex in its first argument and that it takes values in $[0,1]$. Then for any $\eta \le 1$ the cumulative loss $\widehat{L}(y_1^n)$ of the exponentially weighted average predictor (3) satisfies*

$$\widehat{L}(y_1^n) - L^*(y_1^n) \le \eta L^*(y_1^n) + \frac{1+\eta}{\eta}\ln N \ .$$

An immediate corollary is the following:

Corollary 3 *Assume that the exponentially weighted average predictor is used with $\eta = 1$. Then, under the conditions of Theorem 4,*

$$\widehat{L}(y_1^n) \le 2L^*(y_1^n) + 2\ln N \ .$$

The obtained bound is much better than the general bound of Theorem 3 if $L^*(y_1^n) \ll \sqrt{n}$, but may be much worse otherwise. Optimizing the upper bound in Theorem 4 suggests the choice $\eta = \min\left(\sqrt{\ln N / L^*(y_1^n)}, 1\right)$. This choice, however, requires the prior knowledge of the best possible loss $L^*(y_1^n)$:

Corollary 4 *Assume that the value of* $L^*(y_1^n)$ *is known in advance and the exponentially weighted average predictor is used with* $\eta = \min\left(\sqrt{\ln N/L^*(y_1^n)}, 1\right)$ *and assume the conditions of Theorem 4. If* $L^*(y_1^n) \geq \ln N$ *then*

$$\widehat{L}(y_1^n) - L^*(y_1^n) \leq 2\sqrt{L^*(y_1^n)\ln N} + \ln N$$

and $\widehat{L}(y_1^n) - L^*(y_1^n) \leq 3\ln N$ *otherwise.*

The bound of this corollary "interpolates" between the general bound of Theorem 3 and the $\ln N$-type bound of the extreme case with $L^*(y_1^n) = 0$. In practice, of course, the value of $L^*(y_1^n)$ is not known in advance. This difficulty may be easily overcome by periodically updating the value of η, similarly as in Section 1.4 (see the exercises). The constant factor 2 in the bound may be improved to $\sqrt{2}$ by a somewhat more careful analysis (see the exercises).

The proof of Theorem 4 is similar to that of Theorem 3 except that instead of Lemma 2 we use the following inequality.

Lemma 3 *For any random variable* X *with* $X \in [0,1]$*, for any* $\eta > 0$*,*

$$\ln \mathbb{E}\left[e^{-\eta X}\right] \leq -\eta\mathbb{E}X + \frac{\eta^2}{2}\mathbb{E}X \ .$$

PROOF.

$$\begin{aligned}
\ln \mathbb{E}\left[e^{-\eta X}\right] &\leq \ln\left(1 - (1-e^{-\eta})\mathbb{E}X\right) \\
&\qquad \text{(since, by convexity, } e^{-\eta x} \leq 1 - (1-e^{-\eta})x \text{ if } x \in [0,1]) \\
&\leq -(1-e^{-\eta})\mathbb{E}X \\
&\qquad \text{(since } \ln(1-x) \leq -x \text{ for } x \leq 1) \\
&\leq \left(\frac{\eta^2}{2} - \eta\right)\mathbb{E}X \\
&\qquad \text{(since, by Taylor's theorem, } e^{-\eta} \leq 1 - \eta + \eta^2/2) \ . \quad \square
\end{aligned}$$

Proof of Theorem 4. Recall from the proof of Theorem 3 that

$$-\eta L^*(y_1^n) - \ln N \leq \ln\frac{W_n}{W_0} = \sum_{t=1}^n \ln\frac{W_t}{W_{t-1}} = \sum_{t=1}^n \ln\frac{\sum_{i=1}^N w_{i,t-1}e^{-\eta\ell(f_{i,t},y_t)}}{\sum_{i=1}^N w_{i,t-1}} \ .$$

By Lemma 3 the right-hand side may be upper bounded by

$$\sum_{t=1}^{n}\left(\frac{\eta^2}{2}-\eta\right)\frac{\sum_{i=1}^{N} w_{i,t-1}\ell(f_{i,t},y_t)}{\sum_{i=1}^{N} w_{i,t-1}} = \sum_{t=1}^{n}\left(\frac{\eta^2}{2}-\eta\right)\ell(\widehat{p}_t,y_t) = \left(\frac{\eta^2}{2}-\eta\right)\widehat{L}(y_1^n)\ .$$

Thus, we have obtained that

$$\left(1-\frac{\eta}{2}\right)\widehat{L}(y_1^n) - L^*(y_1^n) \le \frac{\ln N}{\eta}\ .$$

Multiply both sides of the inequality by $1+\eta$ and note that for $\eta \le 1$, $(1-\eta/2)(1+\eta) \ge 1$ to finish the proof. ∎

1.9 Bibliographic remarks

Blackwell [4], Hannan [27], and Cover [12] were among the first to carry out worst-case analysis for different sequential prediction problems. Blackwell's [4] method is equivalent to the weighted average predictor based on $\phi(x) = x_+^2$. Vovk [44] and Littlestone and Warmuth [36] developed and analyzed versions of the exponentially weighted average algorithm. Prediction algorithms satisfying condition (2) were considered by Hart and Mas-Colell [28] and, in the context of binary classification, by Grove, Littlestone, and Schuurmans [25]. In learning theory algorithms based on the exponential potential have been intensively studied and applied to a variety of problems (see, e.g., [7, 23, 44, 45]). A version of the doubling trick appears in [7]. Randomized predictors have been considered in various different setups, see, for example, Feder, Merhav, and Gutman [19], Foster and Vohra [21], Cesa-Bianchi and Lugosi [8]. Theorem 2 appears in Kivinen and Warmuth [32]. For $\mathcal{Y} = \{0,1\}$, Haussler, Kivinen, and Warmuth [29] characterize the loss functions for which $\ln N$-type upper bounds are possible. See also Vovk [46]. Hoeffding's Lemma 2 was originally proved to obtain bounds for the probability that a sum of bounded independent random variables exceeds its mean by a certain amount. For the original application only positive values of s were considered, but it is easy to see the it remains true for $s < 0$, the range that's interesting here. Versions of Theorem 3 appear in Cesa-Bianchi [6] and in Cesa-Bianchi and Lugosi

[8]. The "halving" algorithm described at the beginning of Section 1.8 was proposed by Littlestone [35]. For the absolute loss Theorem 4 appears in [7]. For surveys we refer to Foster and Vohra [21], Vovk [46], Merhav and Feder [37].

1.10 Exercises

Exercise 1 Show that the difference between the random cumulative loss $\sum_{t=1}^{n} \tilde{\ell}(\widehat{p}_t, y_t)$ (defined as in Section 1.5) and its expected value $\sum_{t=1}^{n} \bar{\ell}(\widehat{p}_t, y_t)$ is, with overwhelming probability, not larger than something of the order of $\sqrt{n}$.

Exercise 2 Investigate the same problem as in the previous exercise but now for the modification of the randomized predictor defined as

$$w_{i,t} = \frac{\phi'(\tilde{R}_{i,t-1})}{\sum_{j=1}^{N} \phi'(\tilde{R}_{j,t-1})} .$$

where $\tilde{R}_{i,t} = \sum_{s=1}^{t} (\tilde{\ell}(\widehat{p}_t, y_t) - \ell(f_{i,s}, y_t))$. *Warning:* there is an important difference here between different choices of ϕ. $\phi(x) = e^{\eta x}$ is the easiest case.

Exercise 3 Assume that you have to predict, sequentially, a sequence $y_1, y_2, \ldots$ and you know that the outcomes are i.i.d. random variables taking values in $\{0, 1\}$ (but you do not know the distribution) and your decision space is $[0, 1]$. How would you proceed? Try to estimate the cumulative loss of your predictor and compare it to the cumulative loss of the best of the two experts, one of which always predicts one, and the other always predicts zero. Which are the most "difficult" distributions? How does your (expected) excess loss compare to the one of the weighted average algorithm (which does not "know" that the outcome sequence is i.i.d.)?

Exercise 4 Let $\mathcal{Y} = \mathcal{X} = [0, 1]$ and consider the *Hellinger loss*

$$\ell(z, y) = \frac{1}{2}\left(\left(\sqrt{x} - \sqrt{y}\right)^2 + \left(\sqrt{1-x} - \sqrt{1-y}\right)^2\right) .$$

Determine the values of η for which the function $F(z)$ defined in Theorem 2 in concave.

Exercise 5 Let $\mathcal{Y} = \mathcal{X}$ be the closed ball of radius $r > 0$, centered at the origin in $\mathbb{R}^d$ and consider the squared loss $\ell(\mathbf{z}, \mathbf{y}) = \|\mathbf{z} - \mathbf{y}\|^2$ (where $\|\cdot\|$ denotes the euclidean norm in $\mathbb{R}^d$). Show that $F(\mathbf{z})$ of Theorem 2 is concave whenever $\eta \le 1/(8r^2)$.

Exercise 6 Let $\mathcal{X}$ be a convex subset of $\mathbb{R}^d$, and assume that ℓ is twice differentiable in its first argument. Show that if for a $\mathbf{y} \in \mathcal{Y}$ the function $F(\mathbf{z}) = e^{-\eta\ell(\mathbf{z},\mathbf{y})}$ is concave then $\ell(\mathbf{z}, \mathbf{y})$ is a convex function of $\mathbf{z}$.

Exercise 7 Let $\mathcal{Y} = \{0, 1\}$, $\mathcal{X} = [0, 1]$ and $\ell(\widehat{p}, y) = |x - y|$. Prove that the cumulative loss $\widehat{L}(y_1^n)$ of the exponentially weighted average predictor (3) is always at least as large as the cumulative loss $\min_{i=1,\dots,N} L_i(y_1^n)$ of the best expert. Show that for other loss functions such as the squared loss this is not necessarily so. What can you say about the randomized predictors of Section 1.5? *Hint:* Try to reverse the proof of Theorem 3.

Exercise 8 State and prove the analogues of Theorems 3 and 4 in the setup of randomized predictors of Section 1.5.

Exercise 9 Use Lemma 2 to prove the following result, known as *Hoeffding's inequality*. Let $X_1, \dots, X_n$ be independent real-valued random variables such that for each $i = 1, \dots, n$ there exist some $a_i \le b_i$, such that $\mathbb{P}[X_i \in [a_i, b_i]] = 1$. Then for every $\varepsilon > 0$,

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i - \mathbb{E}\sum_{i=1}^{n} X_i\right| > \varepsilon\right] \le 2e^{-2\varepsilon^2 \Big/ \sum_{i=1}^{n}(b_i - a_i)^2} .$$

Compare the result with what you would get using Chebyshev's inequality and with the central limit theorem.

Exercise 10 Combine Theorem 4 with the doubling trick of Section 1.4 to construct a predictor which, without any previous knowledge of $L^*(y_1^n)$, achieves, for all n,

$$\widehat{L}(y_1^n) - L^*(y_1^n) \le c_1\sqrt{L^*(y_1^n)\ln N} + c_2 \ln N$$

whenever ℓ is a bounded loss function convex in its first argument, where c_1 and c_2 are appropriate constants.

Exercise 11 Show that the bound of Corollary 4 may be improved to

$$\widehat{L}(y_1^n) - L^*(y_1^n) \le \sqrt{2L^*(y_1^n)\ln N} + \ln N$$

by choosing $\eta = \ln(1 + \sqrt{2L^*(y_1^n)/\ln N})$. (Freund and Schapire [23].) *Hint:* In Lemma 3, instead of bounding $-(1-e^{-\eta})$ by $\eta^2/2-\eta$, keep the expression with $-(1-e^{-\eta})$. This leads to the bound $\widehat{L}(y_1^n) \le L^*(y_1^n)\eta/(1-e^{-\eta}) + \ln N/(1-e^{-\eta})$. Show that $\eta \le (1-e^{-2\eta})/(2e^{-\eta})$ and use this inequality to obtain a bound for $\widehat{L}(y_1^n)$ which, with the given choice of η yields the desired bound.

Exercise 12 Define the *adaptive exponentially weighted average* predictor as the exponentially weighted average predictor (3) but with changing value of the parameter η: at time t use $\eta_t = \sqrt{\ln N/\min_{i\le N} L_i(y_1^{t-1})}$. Show that

$$\widehat{L}(y_1^n) - L^*(y_1^n) \le c_1\sqrt{L^*(y_1^n)\ln N} + c_2 \ln N$$

whenever ℓ is a bounded loss function convex in its first argument, where c_1 and c_2 are appropriate constants. (See Auer, Cesa-Bianchi, and Gentile [1].) *Warning:* This exercise is hard.

2 Simulatable experts, absolute loss

2.1 Absolute loss

In this chapter we take a closer look at the sequential prediction problem when the outcome space is $\mathcal{Y} = \{0, 1\}$, the decision space is $\mathcal{X} = [0, 1]$, and the loss function is the "absolute loss" $\ell(\widehat{p}, y) = |\widehat{p} - y|$. This setup serves as a prototype for a variety of situations. In particular, many of the results shown in this section generalize in a straightforward way to the problem of randomized prediction introduced in Section 1.5 but to make the ideas more transparent we stay with this simple example. Note that this setup is equivalent to the case when $\mathcal{Y} = \mathcal{X} = \{0, 1\}$, $\ell(\widehat{p}, y) = \mathbb{I}_{[\widehat{p} \neq y]}$ and the predictor is allowed to randomize.

2.2 Different models for experts

The problem of sequential prediction with expert advice discussed in the previous chapter may be considered as a game between the predictor and the environment. There are various versions of this game, depending on what exactly we mean by an "expert". In the version studied up to now the experts' predictions $f_{i,t}$ may depend on the sequence y_1^{t-1} seen so far, and even on previous predictions $\widehat{p}_1, \ldots, \widehat{p}_{t-1}$ of the predictor.

To make the definition rigorous, we may describe the prediction problem as an *iterated game* between the *predictor* and the *environment*. The game is parametrized by two positive integers n and N, representing the length of the sequence to be predicted (i.e., number of game rounds to play) and the number of experts. In each round $t = 1, \ldots, n$:

(1) The environment picks $f_{1,t}, \ldots, f_{N,t} \in \mathcal{X}$;
(2) The predictor picks $\widehat{p}_t \in \mathcal{X}$;
(3) The environment picks $y_t \in \mathcal{Y}$.

The predictor's goal is to minimize the worst-case excess cumulative loss

$$\sup_{y_1^n \in \mathcal{Y}^n} \left(\widehat{L}(y_1^n) - \min_{i=1,\ldots,N} L_i(y_1^n) \right) .$$

This minimum value is denoted by

$$V_n^{(N)} = \inf \sup \sup_{y_1^n \in \mathcal{Y}^n} \left(\widehat{L}(y_1^n) - \min_{i=1,\dots,N} L_i(y_1^n) \right)$$

(where the inf is taken over all feasible prediction strategies and the first sup is taken over all feasible strategies of the environment) and is called the *value of the game* of the *minimax loss*. We have seen it in Section 1.7 that for all n and N, $V_n^{(N)} \le \sqrt{(n/2)\ln N}$.

In many cases it is more natural to restrict the notion of an expert. In the remaining part of this chapter, an expert f means a sequence of functions $f_t : \mathcal{Y}^{t-1} \to \mathcal{X}$, $t = 1, 2, \dots, n$ which maps sequences of past outcomes y_1^{t-1} into elements of the decision space. The value $f_t(y_1^{t-1})$ is interpreted as the expert's prediction at time t. Note that here we do not allow the expert's prediction to depend on $\widehat{p}_s$ $(s < t)$ anymore. We will see that this restriction is not a serious one: experts which are allowed to depend on past predictions of the predictor are, in a certain sense, not more "powerful" than those which are not. We will make this statement precise later.

The corresponding iterated game is now parametrized by a positive integer n and by a set $\mathcal{F}$ of experts, where $\mathcal{F}$ may not be a finite set. In each round $t = 1, \dots, n$:

(1) The predictor picks $\widehat{p}_t \in \mathcal{X}$;
(2) The environment picks $y_t \in \mathcal{Y}$.

The goal of the predictor is the same as before: to minimize the worst-case excess cumulative loss

$$\sup_{y_1^n \in \mathcal{Y}^n} \left(\widehat{L}(y_1^n) - \min_{f \in \mathcal{F}} L_f(y_1^n) \right)$$

where $L_f(y_1^n) = \sum_{t=1}^n \ell(f_t(y_1^{t-1}), y_t)$ denotes the cumulative loss of expert f. The minimum value is denoted by

$$V_n(\mathcal{F}) = \inf \sup_{y_1^n \in \mathcal{Y}^n} \left(\widehat{L}(y_1^n) - \min_{f \in \mathcal{F}} L_f(y_1^n) \right)$$

where the infimum is taken over all feasible predictors. In this setup the predictor is, just like the experts, a sequence of functions $\widehat{p}_t : \mathcal{Y}^{t-1} \to \mathcal{X}$, $t = 1, 2, \dots, n$. To simplify notation, we will often write $\widehat{p}_t$ instead of $\widehat{p}_t(y_1^{t-1})$. Clearly, if the cardinality of $\mathcal{F}$ is $|\mathcal{F}| = N$ then

$$V_n(\mathcal{F}) \le V_n^{(N)} .$$

We may interpret this second game as if the predictor "knew", for all $f \in \mathcal{F}$, the entire sequence of functions $\{f_t\}_{t=1}^n$ in advance as opposed to the first game where the predictor is only allowed to base its predictions on the predictions $f_s(y_1^{s-1})$ $s = 1, \dots, t$ of the experts made so far. In many situations it is reasonable to assume that the experts are completely known by the predictor and this information may clearly be helpful. Such "known" experts have been called *predictable* or *simulatable*. We adopt the latter expression.

2.3 Optimal algorithm for simulatable experts

The purpose of this section is to present, in the case of simulatable experts (i.e., the second iterated game of the previous section), a predictor which is optimal in the sense that it minimizes, among all feasible predictors, the worst-case excess cumulative loss

$$\sup_{y_1^n \in \{0,1\}^n} \left(\widehat{L}(y_1^n) - \min_{f \in \mathcal{F}} L_f(y_1^n) \right) ,$$

that is, for which

$$\sup_{y_1^n \in \{0,1\}^n} \left(\widehat{L}(y_1^n) - \min_{f \in \mathcal{F}} L_f(y_1^n) \right) = V_n(\mathcal{F}) .$$

Before describing the optimal predictor, we note that, since the experts are simulatable, the predictor may calculate the loss of each expert for any particular outcome sequence. In particular, for all $y_1^n \in \mathcal{Y}^n$, the predictor may compute $\inf_{f \in \mathcal{F}} L_f(y_1^n)$.

We determine the optimal predictor "backwards", starting with $\widehat{p}_n$, the prediction at time n. Assume that the first $n-1$ outcomes y_1^{n-1} have been revealed and we want to determine, optimally, the prediction $\widehat{p}_n = \widehat{p}_n(y_1^{n-1})$. Since our goal is to minimize the worst-case excess loss, we need to determine $\widehat{p}_n$ to minimize

$$\max\left(\widehat{L}(y_1^{n-1}) + \ell(\widehat{p}_n, 0) - \inf_{f\in\mathcal{F}} L_f(y_1^{n-1}0);\ \widehat{L}(y_1^{n-1}) + \ell(\widehat{p}_n, 1) - \inf_{f\in\mathcal{F}} L_f(y_1^{n-1}1)\right)$$

where $y_1^{n-1}0$ denotes the string of n bits whose first $n-1$ bits are y_1^{n-1} and the last bit is 0. Minimizing this quantity is equivalent to minimizing

$$\max\left(\widehat{p}_n - \inf_{f\in\mathcal{F}} L_f(y_1^{n-1}0);\ 1 - \widehat{p}_n - \inf_{f\in\mathcal{F}} L_f(y_1^{n-1}1)\right)$$

Clearly, if we write $A_n(y_1^n) = -\inf_{f\in\mathcal{F}} L_f(y_1^n)$ then this is achieved by

$$\widehat{p}_n = \begin{cases} 0 & \text{if } A_n(y_1^{n-1}0) > A_n(y_1^{n-1}1) + 1 \\ 1 & \text{if } A_n(y_1^{n-1}0) + 1 < A_n(y_1^{n-1}1) \\ \frac{A_n(y_1^{n-1}1) - A_n(y_1^{n-1}0) + 1}{2} & \text{otherwise} \end{cases}$$

A crucial observation is that $\widehat{p}_n$ does not depend on the previous predictions $\widehat{p}_1, \ldots, \widehat{p}_{n-1}$. Define

$$A_{n-1}(y_1^{n-1}) \stackrel{\text{def}}{=} \min_{\widehat{p}_n\in[0,1]} \max\left(\widehat{p}_n - \inf_{f\in\mathcal{F}} L_f(y_1^{n-1}0);\ 1 - \widehat{p}_n - \inf_{f\in\mathcal{F}} L_f(y_1^{n-1}1)\right) .$$

This may be rewritten as

$$A_{n-1}(y_1^{n-1}) = \min_{\widehat{p}_n\in[0,1]} \max\left(\widehat{p}_n + A_n(y_1^{n-1}0);\ 1 - \widehat{p}_n + A_n(y_1^{n-1}1)\right) .$$

So far we have calculated the optimal prediction at the last time instance $\widehat{p}_n$. Next we determine optimally $\widehat{p}_{n-1}$, assuming that at time n the optimal prediction is used. Determining $\widehat{p}_{n-1}$ is clearly equivalent to minimizing

$$\max\left(\widehat{L}(y_1^{n-2}) + \ell(\widehat{p}_{n-1}, 0) + A_{n-1}(y_1^{n-2}0);\ \widehat{L}(y_1^{n-2}) + \ell(\widehat{p}_{n-1}, 1) + A_{n-1}(y_1^{n-2}1)\right) ,$$

or equivalently, to minimizing

$$\max\left(\widehat{p}_{n-1} + A_{n-1}(y_1^{n-2}0);\ 1 - \widehat{p}_{n-1} + A_{n-1}(y_1^{n-2}1)\right) .$$

The solution is, as before,

$$\widehat{p}_{n-1} = \begin{cases} 0 & \text{if } A_{n-1}(y_1^{n-2}0) > A_{n-1}(y_1^{n-2}1) + 1 \\ 1 & \text{if } A_{n-1}(y_1^{n-2}0) + 1 < A_{n-1}(y_1^{n-2}1) \\ \frac{A_{n-1}(y_1^{n-2}1) - A_{n-1}(y_1^{n-2}0) + 1}{2} & \text{otherwise} \end{cases}$$

Clearly, the procedure may be continued in the same way until we determine $\widehat{p}_1$.

Formally, given a class $\mathcal{F}$ of experts and a positive integer n, a predictor whose worst-case excess loss equals $V_n(\mathcal{F})$ is determined by the following recursion:

1. (Initialization.) $A_n(y_1^n) = -\inf_{f\in\mathcal{F}} L_f(y_1^n)$
2. (Recurrence.) For $t = n, n-1, \ldots, 1$,

$$A_{t-1}(y_1^{t-1}) = \min_{\widehat{p}\in[0,1]} \max\left(\widehat{p} + A_t(y_1^{t-1}0);\ 1 - \widehat{p} + A_t(y_1^{t-1}1)\right) .$$

and

$$\widehat{p}_t = \begin{cases} 0 & \text{if } A_t(y_1^{t-1}0) > A_t(y_1^{t-1}1) + 1 \\ 1 & \text{if } A_t(y_1^{t-1}0) + 1 < A_t(y_1^{t-1}1) \\ \frac{A_t(y_1^{t-1}1) - A_t(y_1^{t-1}0) + 1}{2} & \text{otherwise} \end{cases}$$

Note that the recurrence for A_t may also be written as

$$A_{t-1}(y_1^{t-1}) = \begin{cases} A_t(y_1^{t-1}0) & \text{if } A_t(y_1^{t-1}0) > A_t(y_1^{t-1}1) + 1 \\ A_t(y_1^{t-1}1) & \text{if } A_t(y_1^{t-1}0) + 1 < A_t(y_1^{t-1}1) \\ \frac{A_t(y_1^{t-1}1) + A_t(y_1^{t-1}0) + 1}{2} & \text{otherwise} \end{cases} \tag{5}$$

The algorithm for calculating the optimal predictor has an important by-product: the value A_0 of the quantity $A_{t-1}(y_1^{t-1})$ at the last step ($t = 1$) of the recurrence clearly gives

$$A_0 = \max_{y_1^n} \left(\sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) + A_n(y_1^n) \right) = V_n(\mathcal{F}) .$$

Thus, the same algorithm also calculates the minimal worst-case excess loss. In the next section we will see some useful consequences of this fact.

2.4 Static experts

In this section we focus our attention to a special type of experts which we call *static*. An expert f is called static if for all $t \le n$ and $y_1^{t-1} \in \mathcal{Y}^{t-1}$, $f_t(y_1^{t-1}) = f_t \in [0,1]$. In other words, a static expert's predictions do not depend on the past outcomes, they are fixed in advance. For example, the expert which predicts always 0 regardless of the past outcomes is static, but the expert whose prediction is the average of all previously seen outcomes is not static.

The next simple technical result has some surprising consequences. The simple inductive proof is left as an exercise.

Lemma 4 *Let $\mathcal{F}$ be an arbitrary class of static experts. Then for all $t \le n$ and $y_1^{t-1} \in \{0,1\}^{t-1}$,*

$$|A_t(y_1^{t-1}1) - A_t(y_1^{t-1}0)| \le 1 \ .$$

This lemma implies the following interesting result.

Theorem 5 *If $\mathcal{F}$ is a class of static experts, then*

$$V_n(\mathcal{F}) = \frac{n}{2} - \frac{1}{2^n} \sum_{y_1^n \in \{0,1\}^n} \inf_{f \in \mathcal{F}} L_f(y_1^n) \ .$$

PROOF. Lemma 4 and (5) imply that for all $t = 1, \ldots, n$,

$$A_{t-1}(y_1^{t-1}) = \frac{A_t(y_1^{t-1}1) + A_t(y_1^{t-1}0) + 1}{2} \ .$$

Applying this equation recursively, we obtain

$$A_0 = \frac{1}{2^n} \sum_{y_1^n \in \{0,1\}^n} A_n(y_1^n) + \frac{n}{2} \ .$$

Recalling that $V_n(\mathcal{F}) = A_0$ and $A_n(y_1^n) = -\inf_{f \in \mathcal{F}} L_f(y_1^n)$ we conclude the proof. ∎

To understand better the behavior of the value $V_n(\mathcal{F})$ it will be advantageous to reformulate the obtained expression. Recall that, since experts are static, each expert f is represented by a vector $(f_1, \dots, f_n) \in [0,1]^n$ where f_t is the prediction of expert f at time t. Since $\ell(f_t, y) = |f_t - y|$, $L_f(y_1^n) = \sum_{t=1}^n |f_t - y_t|$. Also, the average over all possible outcome sequences appearing in the expression of $V_n(\mathcal{F})$ may be treated as an expected value. To this end, introduce i.i.d. symmetric Bernoulli random variables $Y_1, \dots, Y_n$ (i.e., $\mathbb{P}[Y_t = 0] = \mathbb{P}[Y_t = 1] = 1/2$). Then, by Theorem 5,

$$\begin{aligned} V_n(\mathcal{F}) &= \frac{n}{2} - \mathbb{E}\left[\inf_{f\in\mathcal{F}} \sum_{t=1}^n |f_t - Y_t|\right] \\ &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^n \left(\frac{1}{2} - |f_t - Y_t|\right)\right] \\ &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^n \left(\frac{1}{2} - f_t\right)\sigma_t\right] \end{aligned} \tag{6}$$

where $\sigma_t = 1 - 2Y_t$ are i.i.d. *Rademacher* random variables.

Putting $V_n(\mathcal{F})$ in this form opens the toolbox of probability theory in which expected values of this form have been very well understood.

For illustration purposes first we consider a simple example which already offers us important messages. In Section 2.7 we obtain general upper and lower bounds for $V_n(\mathcal{F})$ based on the characterization in terms of Rademacher averages.

2.5 A simple example

Here we consider the case when $\mathcal{F}$ contains the two "extreme" static experts $f^{(1)} \equiv 0$ and $f^{(2)} \equiv 1$. Recall that by Theorem 3

$$V_n(\mathcal{F}) \le V_n^{(2)} \le \sqrt{\frac{n \ln 2}{2}} \approx 0.5887\sqrt{n} \ .$$

Next we contrast this bound with bounds obtained directly using (6). Since $f_t^{(1)} = 0$ and $f_t^{(2)} = 1$ for all t,

$$\begin{aligned} V_n(\mathcal{F}) &= \frac{1}{2}\mathbb{E}\left[\max\left\{\sum_{t=1}^{n} \sigma_t, \sum_{t=1}^{n} -\sigma_t\right\}\right] \\ &= \frac{1}{2}\mathbb{E}\left|\sum_{t=1}^{n} \sigma_t\right|. \end{aligned}$$

We may easily bound this quantity, from above, using the Cauchy-Schwarz inequality:

$$V_n(\mathcal{F}) = \frac{1}{2}\mathbb{E}\left|\sum_{t=1}^{n} \sigma_t\right| \le \frac{1}{2}\sqrt{\mathbb{E}\left(\sum_{t=1}^{n} \sigma_t\right)^2} = \frac{\sqrt{n}}{2}.$$

Observe that this bound has the same order of magnitude as the bound obtained by Theorem 3 but it has a slightly better constant.

We may obtain a similar lower bound as an easy consequence of Khinchine's inequality which we recall here. (See the Appendix for the proof.)

Theorem 6 KHINCHINE'S INEQUALITY. *Let $a_1, \dots, a_n$ be real numbers, and let $\sigma_1, \dots, \sigma_n$ be i.i.d. sign variables with $\mathbb{P}\{\sigma_1 = 1\} = \mathbb{P}\{\sigma_1 = -1\} = 1/2$. Then*

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} a_i\sigma_i\right|\right] \ge \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n} a_i^2}.$$

Applying Theorem 6 to our problem we obtain the lower bound

$$V_n(\mathcal{F}) \ge \sqrt{\frac{n}{8}}.$$

Summarizing the upper and lower bounds, for every n, we have

$$0.3535 \le \frac{V_n(\mathcal{F})}{\sqrt{n}} \le 0.5.$$

For example, for $n = 100$, there exists a prediction strategy such that for any sequence $y_1, \dots, y_{100}$ the number of mistakes is not more than that of the best expert plus 5, but for any prediction strategy there exists a sequence $y_1, \ldots, y_{100}$ such that the number of excess errors is at least 3. The exact asymptotic value is also easy to calculate: $V_n(\mathcal{F})/\sqrt{n} \to 1/\sqrt{2\pi} \approx 0.3989$ (see the exercises).

Another important corollary of the obtained upper bound is that the "$\sqrt{n}$-type" performance bounds we obtained for various versions of the weighted average predictor cannot be improved in general for bounded loss functions. The absolute loss is an example for which no predictor can achieve, in a worst-case sense, an excess loss which is $o(\sqrt{n})$. (Recall from Section 1.6 that for some loss functions significantly smaller worst-case excess loss may be achieved.)

2.6 A general lower bound

In this section we prove a general lower bound for the quantity $V_n^{(N)}$. Recall that Theorem 3 implies that for all n and N,

$$\frac{V_n^{(N)}}{\sqrt{(n/2)\ln N}} \leq 1 \ .$$

The main result of this section shows that, in a sense, this bound cannot be improved further. It also implies that the exponentially weighted average predictor is, in a sense, optimal.

Theorem 7

$$\sup_{n,N} \frac{V_n^{(N)}}{\sqrt{(n/2)\ln N}} \geq 1 \ .$$

Proof. Clearly, $V_n^{(N)} \geq \sup_{\mathcal{F}:|\mathcal{F}|=N} V_n(\mathcal{F})$ where we take the supremum over classes of N *static* experts. We lower bound this supremum by taking an average over an appropriately chosen set of expert classes containing N experts. This may be done by replacing each expert $f = (f_1, \ldots, f_n)$ by a sequence of symmetric i.i.d. Bernoulli random variables. More precisely, let $\{Z_{i,t}\}$ be an $N \times n$ array of i.i.d. Rademacher random variables with distribution $\mathbb{P}[Z_{i,t} = -1] = \mathbb{P}[Z_{i,t} = 1] = 1/2$. Then

$$\begin{aligned}
\sup_{\mathcal{F}:|\mathcal{F}|=N} V_n(\mathcal{F}) &= \sup_{\mathcal{F}:|\mathcal{F}|=N} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \left(\frac{1}{2} - f_t\right)\sigma_t\right] \\
&\quad \text{(by Theorem 5)} \\
&\geq \frac{1}{2}\mathbb{E}\left[\sup_{i=1,\ldots,N} \sum_{t=1}^{n} Z_{i,t}\sigma_t\right] \\
&= \frac{1}{2}\mathbb{E}\left[\sup_{i=1,\ldots,N} \sum_{t=1}^{n} Z_{i,t}\right]
\end{aligned}$$

By the Central Limit Theorem, for each $i = 1, \ldots, N$, $n^{-1/2}\sum_{t=1}^{n} Z_{i,t}$ converges to a standard normal random variable. In fact, it is not difficult to show (see Lemma 13 in the Appendix for the details) that

$$\lim_{n\to\infty} \mathbb{E}\left[\max_{i=1,\ldots,N} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_{i,t}\right] = \mathbb{E}\left[\max_{i=1,\ldots,N} G_i\right]$$

where $G_1, \ldots, G_N$ are independent standard normal random variables. But it is well known (see Lemma 14 in the Appendix) that

$$\lim_{N\to\infty} \frac{\mathbb{E}\left[\max_{i=1,\ldots,N} G_i\right]}{\sqrt{2\ln N}} = 1$$

which concludes the proof. ∎

2.7 Bounds for classes of static experts

In this section we use Theorem 5 and the rich theory of empirical processes to obtain upper and lower bounds for $V_n(\mathcal{F})$ for general classes $\mathcal{F}$ of static experts. To obtain intuition why the geometric structure of $\mathcal{F}$ should play an important role in the size of $V_n(\mathcal{F})$, consider the simple case when $\mathcal{F}$ contains only two experts. The prediction problem is clearly much "easier" if these two experts are close to each other than if they are very different. For example, if the two experts predictions agree for almost all t and only differ in one position then clearly $V_n(\mathcal{F}) \le 1/2$. However, if the two experts are very different, then this value may be much larger, just recall the example of Section 2.5, where we had $V_n(\mathcal{F}) \ge \sqrt{n/8}$.

A way to capture the geometric structure of the class of experts $\mathcal{F}$ is to consider its *covering numbers*. The covering numbers suitable for our analysis are defined as follows. For any class $\mathcal{F}$ of static experts, let $N_2(\mathcal{F}, r)$ be the minimum cardinality of a set $\mathcal{F}_r$ of static experts (possibly not all belonging to $\mathcal{F}$) such that

$$\text{for all } f \in \mathcal{F} \text{ there exists a } g \in \mathcal{F}_r \text{ such that } \qquad \sqrt{\sum_{t=1}^{n} (f_t - g_t)^2} \le r \ .$$

The following bound shows how $V_n(\mathcal{F})$ may be bounded from above in terms of these covering numbers.

Theorem 8 *For any class $\mathcal{F}$ of static experts,*

$$V_n(\mathcal{F}) \le 12 \int_0^{\sqrt{n}/2} \sqrt{\ln N_2(\mathcal{F}, r)}\, dr \ .$$

The result is a straightforward corollary of Theorem 5, Lemma 2, and the following classical result of empirical process theory.

To state the result in a general form, consider a family

$$\{T_f : f \in \mathcal{F}\}$$

of zero mean random variables indexed by a metric space $(\mathcal{F}, \rho)$. Let $N_\rho(\mathcal{F}, r)$ denote the covering number of the metric space $\mathcal{F}$ with respect to the metric ρ. The family is called *subgaussian* in the metric ρ whenever

$$\mathbb{E}\left[e^{\lambda(T_f - T_g)}\right] \le e^{\lambda^2 \rho(f,g)^2/2}$$

holds for any $f, g \in \mathcal{F}$ and $\lambda > 0$. The family is called *sample continuous* if for any sequence $f^{(1)}, f^{(2)}, \ldots \in \mathcal{F}$ converging to some $f \in \mathcal{F}$, we have $T_{f^{(n)}} - T_f \to 0$ almost surely.

Theorem 9 *If* $\{T_f : f \in \mathcal{F}\}$ *is subgaussian and sample continuous in the metric* ρ, *then*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} T_f\right] \leq 12 \int_0^{D/2} \sqrt{\ln N_\rho(\mathcal{F}, \varepsilon)}\, d\varepsilon$$

where D *is the diameter of* $\mathcal{F}$.

The proof is given in the Appendix.

For completeness, and without proof, we mention a lower bound corresponding to Theorem 8. Once again, the inequality is a straightforward corollary of Theorem 5 and known lower bounds for the expected maximum of Rademacher averages.

Theorem 10 *Let* $\mathcal{F}$ *be an arbitrary class of static experts containing* f *and* g *such that* $f_t = 0$ *and* $g_t = 1$ *for all* $t = 1, \ldots, n$. *Then, there exists a universal constant* $K > 0$ *such that*

$$V_n(\mathcal{F}) \geq K \sup_r r \sqrt{\ln N_2(\mathcal{F}, r)}\ .$$

The bound of Theorem 10 is typically of the same order of magnitude as that of Theorem 8. For examples we refer to the exercises.

2.8 Bounds for general classes

All bounds of the previous sections are based on Theorem 5, a characterization of $V_n(\mathcal{F})$ in terms of expected maxima of Rademacher processes. However, this important tool does not work if the class $\mathcal{F}$ contains nonstatic experts. This section discusses some of the possibilities.

We begin with a simple but very useful bound for expert classes which are subsets of "convex hulls" of just finitely many experts.

Theorem 11 *Assume that the class of experts* $\mathcal{F}$ *is such that there exist* N *experts* $f^{(1)}, \ldots, f^{(N)}$ *(not necessarily in* $\mathcal{F}$*) such that for all* $f \in \mathcal{F}$ *there exist convex coefficients* $q_1, \ldots, q_N \geq 0$ *with* $\sum_{j=1}^N q_j = 1$ *such that* $f_t(y_1^{t-1}) \sum_{j=1}^N q_j f_t^{(j)}(y_1^{t-1})$ *for all* $t = 1, \ldots, n$ *and* $y_1^{t-1} \in \{0,1\}^{t-1}$. *Then*

$$V_n(\mathcal{F}) \leq \sqrt{(n/2) \ln N}.$$

PROOF. Proof The key property is that for any bit sequence $y_1^n \in \{0,1\}^n$ and expert $f = \sum_{j=1}^N q_j f^{(j)} \in \mathcal{F}$ there exists an expert among $f^{(1)}, \ldots, f^{(N)}$ whose loss on y_1^n is not larger than that of f. To see this note that

$$\begin{aligned}
L_f(y_1^n) &= \sum_{t=1}^n |f_t(y_1^{t-1}) - y_t| \\
&= \sum_{t=1}^n \left| \sum_{j=1}^N q_j f_t^{(j)}(y_1^{t-1}) - y_t \right| \\
&= \sum_{t=1}^n \sum_{j=1}^N q_j \left| f_t^{(j)}(y_1^{t-1}) - y_t \right| \\
&= \sum_{j=1}^N q_j \sum_{t=1}^n \left| f_t^{(j)}(y_1^{t-1}) - y_t \right| \\
&= \sum_{j=1}^N q_j L_{f^{(j)}}(y_1^n) \\
&\geq \min_{j=1,\ldots,N} L_{f^{(j)}}(y_1^n) \ .
\end{aligned}$$

Thus, for all $y_1^n \in \{0,1\}^n$, $\inf_{f \in \mathcal{F}} L_f(y_1^y) \geq \min_{j=1,\ldots,N} L_{f^{(j)}}(y_1^n)$. This implies that if $\widehat{p}$ is the exponentially weighted average predictor based on the finite class $f^{(1)}, \ldots, f^{(N)}$, then, by Theorem 3,

$$\widehat{L}(y_1^n) - \inf_{f \in \mathcal{F}} L_f(y_1^y) \leq \widehat{L}(y_1^n) - \min_{j=1,\ldots,N} L_{f^{(j)}}(y_1^n) \leq \sqrt{\frac{n \ln N}{2}}$$

which completes the proof. ▮

Next we review two basic examples.

Example. (LINEAR PREDICTORS.) As a first example, consider the class $\mathcal{L}_k$ of k-th order autoregressive linear predictors, where $k \geq 2$ is a fixed positive integer. The class $\mathcal{L}_k$ contains all experts f such that

$$f_t(y_{1-k}^{t-1}) = \sum_{i=1}^k q_i y_{t-i}$$

for some $q_1, \ldots, q_k \geq 0$ with $\sum_{i=1}^k q_i = 1$. In other words, an expert f predicts according to a convex combination of the k most recent outcomes of the sequence. Convexity of the coefficients q_i assures that $f_t(y_{1-k}^{t-1}) \in [0,1]$.

As each prediction of a k-th order autoregressive linear predictor is determined by the last k bits observed, we add a prefix $y_{-k+1}, \ldots, y_0$ to the sequence y_1^n to be predicted. We use y_{1-k}^n to denote the resulting sequence of $n+k$ bits. Accordingly, for such predictors the value $V_n(\mathcal{F})$ is re-defined by

$$V_n(\mathcal{F}) = \inf \max_{y_{1-k}^n \in \{0,1\}^{n+k}} \left(\widehat{L}(y_{1-k}^n) - \inf_{f \in \mathcal{F}} L_f(y_{1-k}^n) \right) ,$$

where

$$\widehat{L}(y_{1-k}^n) = \sum_{t=1}^{n} \left| \widehat{p}_t(y_{1-k}^{t-1}) - y_t \right|$$

and $L_f(y_{1-k}^n)$ is defined similarly.

Corollary 5 *For any positive integers n and $k \geq 2$,*

$$V_n(\mathcal{L}_k) \leq \sqrt{\frac{n \ln k}{2}} .$$

PROOF. The statement is a direct consequence of Theorem 11 if we observe that $\mathcal{L}_k$ is the convex hull of the k experts $f^{(1)}, \ldots, f^{(k)}$ defined by $f_t^{(i)}(y_1^{t-1}) = y_{t-i}$, $i = 1, \ldots, k$. ∎

Example. (MARKOV PREDICTORS.) For an arbitrary $k \geq 1$, we consider the class $\mathcal{M}_k$ of k-th order Markov experts defined as follows. The class $\mathcal{M}_k$ is indexed by the set $[0,1]^{2^k}$ so that the index of any $f \in \mathcal{M}_k$ is the vector $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_{2^k-1})$ with $\alpha_s \in [0,1]$ for $0 \leq s < 2^k$. If f has index α then $f_t(y_{-k+1}^{t-1}) = \alpha_s$ for all $1 \leq t \leq n$ and for all $y_{-k+1}^{t-1} \in \{0,1\}^{t+k-1}$, where s has binary expansion $y_{t-k}, \ldots, y_{t-1}$.

As each prediction of a k-th order Markov expert is determined by the last k bits observed, we add a prefix $y_{-k+1}, \ldots, y_0$ to the sequence to predict in the same way we did in the previous example for the autoregressive experts. Thus, the function f_t is now defined over the set $\{0,1\}^{t+k-1}$.

Once more, Theorem 11 immediately implies a bound for $V_n(\mathcal{M}_k)$:

Corollary 6 *For any positive integers* n *and* $k \geq 2$*,*

$$V_n(\mathcal{M}_k) \leq \sqrt{\frac{2^k n \ln 2}{2}} \ .$$

Next we study how one can derive lower bounds for $V_n(\mathcal{F})$ for general classes of experts. Even though Theorem 5 cannot be generalized to arbitrary classes of experts, its analogue remains true as an inequality:

Theorem 12 *For any expert class* $\mathcal{F}$*,*

$$V_n(\mathcal{F}) \geq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \left(\frac{1}{2} - f_t(Y_1^{t-1})\right)(1-2Y_t)\right] ,$$

where $Y_1, \ldots, Y_n$ *are independent Bernoulli* $(1/2)$ *random variables,*

Proof. For any prediction strategy, if Y_t is a Bernoulli $(1/2)$ random variable then $\mathbb{E}|\widehat{p}_t(y_1^{t-1}) - Y_t| = 1/2$ for each y_1^{t-1}. Hence

$$\begin{aligned}
V_n(\mathcal{F}) &\geq \max_{y_1^n\in\{0,1\}^n}\left(\widehat{L}(y_1^n) - \inf_{f\in\mathcal{F}} L_f(y_1^n)\right) \\
&\geq \mathbb{E}\left[\widehat{L}(Y_1^n) - \inf_{f\in\mathcal{F}} L_f(Y_1^n)\right] \\
&= \frac{n}{2} - \mathbb{E}\left[\inf_{f\in\mathcal{F}} L_f(Y_1^n)\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \left(\frac{1}{2} - f_t(Y_1^{t-1})\right)(1-2Y_t)\right] . \quad \square
\end{aligned}$$

We demonstrate how to use this inequality for the example of the class of linear predictors described above. In fact, the following result shows that the bound of Corollary 7 is optimal.

Corollary 7

$$\liminf_{k\to\infty} \liminf_{n\to\infty} \frac{V_n(\mathcal{L}_k)}{\sqrt{n \ln k}} = \frac{1}{\sqrt{2}} \ .$$

PROOF. We only sketch the proof, the details are left to the reader as an easy exercise. By Theorem 12, if we write $Z_t = 1-2Y_t$ for $t = -k+1, \ldots, n$, then

$$V_n(\mathcal{L}_k) \geq \frac{1}{2}\mathbb{E}\left[\sup_{f\in\mathcal{L}_k}\sum_{t=1}^{n}(1-2f_t(Y_1^{t-1}))Z_t\right] \geq \frac{1}{2}\mathbb{E}\left[\max_{1\leq i\leq k}\sum_{t=1}^{n} Z_t Z_{t-i}\right]$$

The proof is now similar to the proof of Theorem 7 (see Lemma 13 in the Appendix) with the exception that, instead of the ordinary Central Limit Theorem, we use a generalization of it to martingales.

Consider the k-vector $X_n = (X_{n,1}, \ldots, X_{n,k})$ of components

$$X_{n,i} \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}}\sum_{t=1}^{n} Z_t Z_{t-i}, \quad i = 1, \ldots, k.$$

By the "Cramér-Wold device" the sequence of vectors $\{X_n\}$ converges in distribution to a vector random variable $N = (N_1, \ldots, N_k)$ if and only if $\sum_{i=1}^{k} a_i X_{n,i}$ converges in distribution to $\sum_{i=1}^{k} a_i N_i$ for all possible choices of the coefficients $a_1, \ldots, a_k$. Thus, consider

$$\sum_{i=1}^{k} a_i X_{n,i} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} Z_t \sum_{i=1}^{k} a_i Z_{t-i} \; .$$

It is easy to see that the sequence of random variables $\sqrt{n}X_{n,i}$, $n = 1, 2, \ldots$, forms a martingale with respect to the sequence of σ-algebras $\mathcal{G}_t$ generated by $Z_{-k+1}, \ldots, Z_t$. Furthermore, by the martingale central limit theorem (see, e.g., Hall and Heyde [26, Theorem 3.2]) $\sum_{i=1}^{k} a_i X_{n,i}$ converges in distribution, as $n \to \infty$, to a zero-mean normal random variable with variance $\sum_{i=1}^{k} a_i^2$. Then, by the Cramér-Wold device, as $n \to \infty$ the vector X_n converges in distribution to $N = (N_1, \ldots, N_k)$, where $N_1, \ldots, N_k$ are independent standard normal random variables. The rest of the proof is identical to that of Lemma 13 in the Appendix except that Hoeffding's inequality needs to be replaced by its analogue for bounded martingale differences due to Azuma [2]. ∎

2.9 Bibliographic remarks

The predictor presented in Section 2.3 appears in Chung [11] who also gives an optimal algorithm for nonsimulatable experts (see the exercises). The form of Khinchine's inequality cited here is due to Szarek [43]. The proof given here is due to Devroye and Györfi [17]. The example described in Section 2.3 was first studied in detail by Cover [12] and was generalized substantially by Feder, Merhav, and Gutman [19]. Theorem 7 is due to Cesa-Bianchi, Freund, Haussler, Hembold, Schapire, and Warmuth [7]. Similar results for other loss functions appear in Haussler, Kivinen, and Warmuth [29]. More general lower bounds for the absolute loss were proved by Cesa-Bianchi and Lugosi [8]. Theorem 9 is a simple version of Dudley's metric entropy bound [18]. Theorem 10 follows from a result of Sudakov, see Ledoux and Talagrand [34]. The Markov experts appearing in Section 2.8 were first considered by Feder, Merhav, and Gutman [19]. See also [8] where the lower bounds of Section 2.8 appear. We refer to [8] for more information on upper and lower bounds for general classes of experts.

2.10 Exercises

Exercise 13 Calculate $V_n^{(1)}$, $V_1^{(2)}$, $V_2^{(2)}$, and $V_3^{(2)}$. Calculate $V_1(\mathcal{F})$, $V_2(\mathcal{F})$, and $V_3(\mathcal{F})$ when $\mathcal{F}$ contains two experts: $f_{1,t} = 0$ and $f_{2,t} = 1$ for all t.

Exercise 14 Prove or disprove:

$$\sup_{\mathcal{F}:|\mathcal{F}|=N} V_n(\mathcal{F}) = V_n^{(N)} .$$

Exercise 15 Prove Lemma 4. *Hint:* Proceed with a backward induction, starting with $t = n$. Visualizing the possible sequences of outcomes in a rooted binary tree may help.

Exercise 16 Use Lemma 4 to show that if $\mathcal{F}$ is a class of static experts then the optimal predictor of Section 2.4 may be written as

$$\widehat{p}_t(y_1^{t-1}) = \frac{1}{2} + \mathbb{E}\left[\frac{\inf_{f\in\mathcal{F}} L_f(y_1^{t-1}0Y_1^{n-t}) - \inf_{f\in\mathcal{F}} L_f(y_1^{t-1}1Y_1^{n-t})}{2}\right]$$

where $Y_1, \ldots, Y_n$ are i.i.d. symmetric Bernoulli random variables. (Chung [11].)

Exercise 17 Give an appropriate modification of the optimal prediction algorithm of Section 2.3 for the case of "nonsimulatable" experts, that is, give a prediction algorithm which achieves the worst-case excess loss $V_n^{(N)}$ in the setup of the first game of Section 2.2. (See [7] and [11].) *Warning:* This exercise requires work.

Exercise 18 Show that Lemma 4 and Theorem 5 are not necessarily true if the experts in $\mathcal{F}$ are not static.

Exercise 19 Show that for the class of experts discussed in Section 2.5

$$\lim_{n\to\infty} \frac{V_n(\mathcal{F})}{\sqrt{n}} = \frac{1}{\sqrt{2\pi}} .$$

Exercise 20 Consider the simple class of experts described in Section 2.5. Feder, Merhav, and Gutman [19] proposed the following predictor. Let $\widehat{p}_t = (1/t)\sum_{s=1}^{t} \mathbb{I}_{[y_s=1]}$ denote the proportion of 1's appeared so far in the outcome sequence. Then the predictor is defined by $\widehat{p}_t = \psi_t(\widehat{p}_t)$ where for $x \in [0,1]$,

$$\psi(x) = \begin{cases} 0 & \text{if } x < 1/2 - \varepsilon_t \\ 1 & \text{if } x > 1/2 + \varepsilon_t \\ (1/2) + (x - 1/2)/2\varepsilon_t & \text{otherwise} \end{cases}$$

where $\varepsilon_t > 0$ is some positive number. Show that if $\varepsilon_t = 1/(2\sqrt{t+2}$ then for all $n > 1$ and $y_1^n \in \{0,1\}^n$,

$$\widehat{L}(y_1^n) - \min_{i=1,2} L_i(y_1^n) \le \sqrt{n+1} + \frac{1}{2} .$$

Warning: This is not an easy exercise. *Hint:* Show first that among all sequences containing $n_1 < n/2$ ones the predictor performs worst for the sequence which starts alternating zeroes and ones n_1 times and ends with $n - 2n_1$ zeroes.

Exercise 21 After reading the previous exercise you may wonder whether the simpler predictor defined with

$$\psi'(x) = \begin{cases} 0 & \text{if } x < 1/2 \\ 1 & \text{if } x > 1/2 \\ 1/2 & \text{if } x = 1/2 \end{cases}$$

also does the job. Show that this is not true: show that there exists a sequence y_1^n such that $\widehat{L}(y_1^n) - \min_{i=1,2} L_i(y_1^n) \approx n/4$ for large n. (See [19].)

Exercise 22 Let $\mathcal{F}$ be the class of all static experts $f \in \mathcal{F}$ of the form $F_t = p$ regardless of t. The class contains all such experts with $p \in [0,1]$. Estimate the covering numbers $N_2(\mathcal{F}, r)$ and compare the upper and lower bounds of Theorems 8 and 10 for this case.

Exercise 23 Consider the class $\mathcal{F}$ of all static experts which predict in a monotonic way, that is, for each $f \in \mathcal{F}$, either $f_t \le f_{t+1}$ for all $t \ge 1$ or $f_t \ge f_{t+1}$ for all $t \ge 1$. Apply Theorem 8 to conclude that

$$V_n(\mathcal{F}) = O\left(\sqrt{n \log n}\right) .$$

What do you obtain using Theorem 10? Can you apply Theorem 11 in this case?

Exercise 24 Construct a predictor such that for *all* $k = 1, 2, \ldots$ and all sequences $y_1, y_2, \ldots,$

$$\limsup_{n\to\infty} \frac{1}{n}\left(\widehat{L}(y_1^n) - \inf_{f \in \mathcal{M}_k} L_f(y_1^n)\right) = 0 .$$

In other words, the predictor predicts asymptotically as well as any Markov predictor of any order. (The existence of such a predictor was shown in [19].) *Hint:* For an easy construction use the doubling trick.

Exercise 25 Show that

$$\liminf_{k\to\infty} \liminf_{n\to\infty} \frac{V_n(\mathcal{M}_k)}{\sqrt{2^k n \ln 2}} = \frac{1}{\sqrt{2}} .$$

Hint: Mimic the proof of Corollary 7.

3 Logarithmic loss

3.1 Sequential probability assignment

This chapter is entirely devoted to the investigation of a special loss function, the *logarithmic* loss, sometimes also called *self information* loss. The reason for this special attention is that this loss function has a meaningful interpretation in various sequential decision problems, including repeated gambling and data compression which we briefly describe below. Sequential prediction which aims at minimizing this loss function is also intimately related to maximizing benefits by repeated investment on the stock market. This application is studied is a separate subsequent chapter.

Now we describe the setup for the entire chapter. Let $m > 1$ be a fixed positive integer, and let the outcome space be $\mathcal{Y} = \{1, 2, \ldots, m\}$. The decision space is the probability simplex

$$\mathcal{X} = \left\{ p = (p(1), \ldots, p(m)) : \sum_{j=1}^{m} p(j) = 1,\ p(j) \geq 0, j = 1, \ldots, m \right\} \subset \mathbb{R}^m \ .$$

A vector $p \in \mathcal{X}$ is often interpreted as a probability distribution over the set $\mathcal{Y}$. Indeed, in some cases, the predictor is required to assign a probability to each possible outcomes, representing the predictor's belief. For example, wheather forecasters predictions usually have the form "the possibility of rain is 40%".

In the entire chapter we consider the model of *simulatable* experts introduced in Section 2.2. Thus, an expert f is a sequence of functions $f_t : \mathcal{Y}^{t-1} \to \mathcal{X}$, $t = 1, 2, \ldots$ so that after having seen the past outcomes y_1^{t-1}, expert f outputs the probability vector $f_t(\cdot|y_1^{t-1}) \in \mathcal{X}$. For the components of this vector we will write

$$f_t(1|y_1^{t-1}), \ldots, f_t(m|y_1^{t-1}) \ .$$

This notation emphasizes further the analogy between an expert and a probability distribution. Indeed, the j-th component of the vector $f_t(j|y_1^{t-1})$ may be interpreted as the *conditional* probability f assigns to the j-th element of $\mathcal{Y}$ given the past y_1^{t-1}.

Similarly, at each time instant the predictor chooses a probability vector

$$\widehat{p}_t(\cdot|y_1^{t-1}) = \left(\widehat{p}_t(1|y_1^{t-1}), \ldots, \widehat{p}_t(m|y_1^{t-1})\right) .$$

Once again, appealing to the analogy with probability distributions, we may introduce, for all $n \geq 1$ and $y_1^n \in \mathcal{Y}^n$, the notation

$$f_n(y_1^n) = \prod_{t=1}^n f_t(y_t|y_1^{t-1}) , \qquad \widehat{p}_n(y_1^n) = \prod_{t=1}^n \widehat{p}_t(y_t|y_1^{t-1}) .$$

Observe that

$$\sum_{y_1^n \in \mathcal{Y}^n} f_n(y_1^n) = \sum_{y_1^n \in \mathcal{Y}^n} \widehat{p}_n(y_1^n) = 1$$

and therefore expert f as well as the predictor define probability distributions over the set of all sequences of length n. Conversely, any probability distribution $\widehat{p}_n(y_1^n)$ over the set $\mathcal{Y}^n$ defines a predictor by the induced conditional distributions

$$\widehat{p}_t(y_t|y_1^{t-1}) = \frac{\widehat{p}_t(y_1^t)}{\widehat{p}_{t-1}(y_1^{t-1})}$$

where $\widehat{p}_t(y_1^t) = \sum_{y_{t+1}^n \in \mathcal{Y}^{n-t}} \widehat{p}_n(y_1^n)$.

The loss function we consider throughout this chapter is defined by

$$\ell(p, y) = \sum_{j=1}^m \mathbb{I}_{[y=j]} \ln \frac{1}{p(j)} = \ln \frac{1}{p(y)} , \quad y \in \mathcal{Y}, p \in \mathcal{X} .$$

For an outcome sequence $y_1, \ldots, y_n$, the cumulative loss of expert f and the predictor are, respectively,

$$L_f(y_1^n) = \sum_{t=1}^n \ln \frac{1}{f_t(y_t|y_1^{t-1})} \qquad \widehat{L}(y_1^n) = \sum_{t=1}^n \ln \frac{1}{\widehat{p}_t(y_t|y_1^{t-1})} .$$

Given a class $\mathcal{F}$ of experts, the difference between the cumulative loss of the predictor and that of the best expert may now be written as

$$\begin{aligned}
\widehat{L}(y_1^n) - \inf_{f \in \mathcal{F}} L_f(y_1^n) &= \sum_{t=1}^n \ln \frac{1}{\widehat{p}_t(y_t|y_1^{t-1})} - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \ln \frac{1}{f_t(y_t|y_1^{t-1})} \\
&= \sup_{f \in \mathcal{F}} \ln \frac{f_n(y_1^n)}{\widehat{p}_n(y_1^n)} .
\end{aligned}$$

Thus, the difference of the cumulative losses may be interpreted as the logarithm of the ratio of the total probabilities that are sequentially assigned to the outcome sequence by the predictor and the experts.

Remark. In these notes we restrict our discussion to the case when the outcome space is a finite set. Note, however, that most results may be extended easily to the more general case when $\mathcal{Y}$ is a measurable space. In such cases the decision space becomes the set of all densities over $\mathcal{Y}$ with respect to some fixed common dominating measure, and the loss is the negative logarithm of the density evaluated at the outcome.

3.2 Mixture predictors

Recall from Section 1.6 that the logarithmic loss function is exp-concave and Theorem 2 applies. Thus, if the class $\mathcal{F}$ of experts is finite and $|\mathcal{F}| = N$, then the exponentially weighted average predictor with parameter $\eta = 1$ satisfies

$$\widehat{L}(y_1^n) - \inf_{f \in \mathcal{F}} L_f(y_1^n) \leq \ln N$$

or equivalently,

$$\widehat{p}_n(y_1^n) \geq \frac{1}{N} \sup_{f \in \mathcal{F}} f_n(y_1^n) \ .$$

It is worth noting that the exponentially weighted average predictor has an interesting interpretation in this special case. Observe that

$$\widehat{p}_t(y_t | y_1^{t-1}) = \frac{\sum_{f \in \mathcal{F}} f_t(y_t | y_1^{t-1}) e^{-\eta L_f(y_1^{t-1})}}{\sum_{f \in \mathcal{F}} e^{-\eta L_f(y_1^{t-1})}}$$

with $\eta = 1$ reduces to

$$\widehat{p}_t(y_t | y_1^{t-1}) = \frac{\sum_{f \in \mathcal{F}} f_t(y_t | y_1^{t-1}) f_{t-1}(y_1^{t-1})}{\sum_{f \in \mathcal{F}} f_{t-1}(y_1^{t-1})} = \frac{\sum_{f \in \mathcal{F}} f_t(y_1^t)}{\sum_{f \in \mathcal{F}} f_{t-1}(y_1^{t-1})} \ .$$

Thus, the total probability the predictor assigns to a sequence y_1^n is just

$$\widehat{p}_n(y_1^n) = \prod_{t=1}^{n} \frac{\sum_{f \in \mathcal{F}} f_t(y_1^t)}{\sum_{f \in \mathcal{F}} f_{t-1}(y_1^{t-1})} = \frac{\sum_{f \in \mathcal{F}} f_n(y_1^n)}{N}$$

(recall that $f(y_1^0)$ is defined to be equal to one). In other words, the probability distribution the exponentially weighted average predictor $\widehat{p}$ defines over the set $\mathcal{Y}^n$ of all strings of length n is just the uniform *mixture* of the distributions defined by the experts. Thus, we will sometimes call the exponentially weighted average predictor *mixture predictor*. This observation suggests a straightforward extension to the case when $\mathcal{F}$ may contain countably many experts. Denote these experts by $f^{(1)}, f^{(2)}, \ldots$. To define the mixture predictor, we assign a nonnegative number $q_i \geq 0$ to each expert $f^{(i)} \in \mathcal{F}$ such that $\sum_{i=1}^{\infty} q_i = 1$. The numbers q_i may be called *prior probabilities*. Now we introduce the predictor $\widehat{p}$ "backwards", by defining the joint probability it assigns to each sequence:

$$\widehat{p}_n(y_1^n) = \sum_{i=1}^{\infty} q_i f^{(i)}(y_1^n) \ .$$

Note that $\widehat{p}$ indeed defines a valid probability distribution over $\mathcal{Y}^n$. Mimicking the argument above, we see that this probability distribution defines the weighted average predictor

$$\widehat{p}_t(y_t|y_1^{t-1}) = \frac{\sum_{i=1}^{\infty} q_i f_t^{(i)}(y_t|y_1^{t-1}) f_{t-1}^{(i)}(y_1^{t-1})}{\sum_{i=1}^{\infty} q_i f_{t-1}^{(i)}(y_1^{t-1})} = \frac{\sum_{i=1}^{\infty} q_i f_t^{(i)}(y_t|y_1^{t-1}) e^{-\eta L_{f^{(i)}}(y_1^{t-1})}}{\sum_{i=1}^{\infty} q_i e^{-\eta L_{f^{(i)}}(y_1^{t-1})}} \ .$$

Using the trivial bound $\widehat{p}_n(y_1^n) = \sum_{i=1}^{\infty} q_i f^{(i)}(y_1^n) \geq q_i f^{(i)}(y_1^n)$ for all i, we obtain the following.

Theorem 13 *Let $\mathcal{F} = \{f^{(1)}, f^{(2)}, \ldots\}$ be an arbitrary countable class of experts and let $\{q_i : i = 1, 2 \ldots\}$ be a probability distribution over the set of positive integers. Then the mixture predictor defined above satisfies, for all $y_1^n \in \mathcal{Y}$,*

$$\widehat{L}(y_1^n) \leq \inf_{f^{(i)} \in \mathcal{F}} \left(L_{f^{(i)}}(y_1^n) + \ln \frac{1}{q_i} \right)$$

Thus, the cumulative loss of the predictor exceeds the loss of each expert by at most a constant, but the constant depends on the expert. If we write the predictor in the form

$$\widehat{p}_t(y_t|y_1^{t-1}) = \frac{\sum_{i=1}^{\infty} f_t^{(i)}(y_t|y_1^{t-1}) e^{-\eta\left(L_{f^{(i)}}(y_1^{t-1}) + \frac{1}{\eta}\ln\frac{1}{q_i}\right)}}{\sum_{i=1}^{\infty} e^{-\eta\left(L_{f^{(i)}}(y_1^{t-1}) + \frac{1}{\eta}\ln\frac{1}{q_i}\right)}}$$

then we see that the quantity $(1/\eta)\ln(1/q_i)$ may be regarded as a "penalty" we add to the cumulative loss of expert i at each time t. Theorem 13 is a so-called "oracle inequality" which states that the mixture predictor achieves a cumulative loss matching the best penalized cumulative loss of the experts.

Later we will extend the idea of a mixture predictor to certain uncountably infinite classes of experts.

3.3 Gambling in the horse race, data compression

Imagine that we want to bet in a horse race in which m horses run repeatedly many times. In the t-th race we bet our entire fortune on the m horses according to proportions $\widehat{p}_t(1), \ldots, \widehat{p}_t(m)$. If horse j wins the t-th race, we multiply our money bet on this horse by a factor of $o_t(j)$ and we lose it otherwise. In other words, if y_t denotes the index of the winner of the t-th race, we multiply our capital by a factor of

$$\sum_{j=1}^{m} \mathbb{I}_{[y_t=j]} \widehat{p}_t(j) o_t(j) = \widehat{p}_t(y_t) o_t(y_t) \ .$$

To make it explicit that both the odds $o_t(j)$ and the proportions $\widehat{p}_t(j)$ according to which we bet may depend on the results of previous races, we write $o_t(j) = o_t(j|y_1^{t-1})$ and $\widehat{p}_t(j|y_1^{t-1})$. If we dispose of an initial capital of C units, after n races our capital is

$$C \prod_{t=1}^{n} \widehat{p}_t(y_t|y_1^{t-1}) o_t(y_t|y_1^{t-1}) \ .$$

Now assume that before each race we ask the advice of a class of experts and our goal is to win as much as the best of these experts. If expert f divides his capital in the t-th race according to proportions $f_t(j|y_1^{t-1})$, $j = 1, \ldots, m$ and starts with the same initial capital then his capital after n races is

$$C \prod_{t=1}^{n} f_t(y_t|y_1^{t-1}) o_t(y_t|y_1^{t-1}) \ .$$

The proportion of the money of the best expert and ours is, thus,

$$\frac{\sup_{f\in\mathcal{F}} C\prod_{t=1}^{n} f_t(y_t|y_1^{t-1})o_t(y_t|y_1^{t-1})}{C\prod_{t=1}^{n} \widehat{p}_t(y_t|y_1^{t-1})o_t(y_t|y_1^{t-1})} = \sup_{f\in\mathcal{F}} \frac{f_n(y_1^n)}{\widehat{p}_n(y_1^n)} ,$$

independently of the odds. The logarithm of this quantity is just the difference of the cumulative logarithmic loss of the predictor $\widehat{p}$ and that of the best expert described in the previous section.

Another important motivation for the study of the logarithmic loss function has its roots in information theory, more concretely in lossless data compression. Instead of describing the data-compression problem in detail, we briefly mention that it is well known that any probability distribution f_n over the set $\mathcal{Y}^n$ defines a *code* which assigns, to each string $y_1^n \in \mathcal{Y}^n$ a *codeword*, that is, a string of bits of length about $-\log_2 f_n(y_1^n)$, and, conversely, any code satisfying a natural condition (i.e., unique decodability) defines a probability distribution. Given a class of codes, or equivalently a class of experts $\mathcal{F}$, the best compression of a string y_1^n is achieved by the code which minimizes the length $-\log_2 f_n(y_1^n)$, that is, which minimizes the logarithmic cumulative loss $L_f(y_1^n)$. Now assume that the symbols of the string y_1^n are revealed one by one and the goal is to compress it almost as well as the best code in the class. It turns out that for any predictor $\widehat{p}_n$ (i.e., sequential probability assignment) it is possible to construct, sequentially, a codeword of length about $-\log_2 \widehat{p}_n(y_1^n)$ using a method called *arithmetic coding*. The difference

$$-\log_2 \widehat{p}_n(y_1^n) - \inf_{f\in\mathcal{F}}(-\log_2 f_n(y_1^n)) = \frac{1}{\ln 2}\left(\widehat{L}(y_1^n) - \inf_{f\in\mathcal{F}} L_f(y_1^n)\right)$$

is called the *redundancy* of the code with respect to the class of codes $\mathcal{F}$. In this respect, the problem of sequential lossless data compression is equivalent to sequential prediction under the logarithmic loss.

3.4 The minimax optimal predictor

In this section we investigate the *minimax loss* defined in Section 2.2 for the logarithmic loss. Recall that for a given class $\mathcal{F}$ of experts, and integer

$n > 0$, the minimax loss is defined by

$$V_n(\mathcal{F}) = \inf_{\widehat{p}} \sup_{y_1^n \in \mathcal{Y}^n} \left(\widehat{L}(y_1^n) - \inf_{f \in \mathcal{F}} L_f(y_1^n) \right) = \inf_{\widehat{p}} \sup_{y_1^n \in \mathcal{Y}^n} \ln \frac{\sup_{f \in \mathcal{F}} f_n(y_1^n)}{\widehat{p}_n(y_1^n)} .$$

If, for a given predictor $\widehat{p}$ we define the worst-case cumulative regret by

$$V_n(\widehat{p}, \mathcal{F}) = \sup_{y_1^n \in \mathcal{Y}^n} \left(\widehat{L}(y_1^n) - \inf_{f \in \mathcal{F}} L_f(y_1^n) \right) ,$$

then $V_n(\mathcal{F}) = \inf_{\widehat{p}} V_n(\widehat{p}, \mathcal{F})$. Interestingly, in the case of the logarithmic loss it is possible to identify explicitly the unique predictor achieving the minimax loss. The next theorem shows that the predictor p^* defined by the *normalized maximum likelihood* probability

$$p_n^*(y_1^n) = \frac{\sup_{f \in \mathcal{F}} f_n(y_1^n)}{\sum_{x_1^n \in \mathcal{Y}^n} \sup_{f \in \mathcal{F}} f_n(x_1^n)}$$

has this property. Note that p_n^* is indeed a probability distribution over the set $\mathcal{Y}^n$ and recall that all probability distributions define a predictor by the corresponding conditional probabilities $p_t^*(y_t | y_1^{t-1})$.

Theorem 14 *For any class $\mathcal{F}$ of experts and integer $n > 0$, the normalized maximum likelihood predictor p^* is the unique predictor such that*

$$\sup_{y_1^n \in \mathcal{Y}^n} \left(\widehat{L}(y_1^n) - \inf_{f \in \mathcal{F}} L_f(y_1^n) \right) = V_n(\mathcal{F}) .$$

Moreover, p^* is an equalizer, that is, for all $y_1^n \in \mathcal{Y}^n$,

$$\ln \frac{\sup_{f \in \mathcal{F}} f_n(y_1^n)}{p_n^*(y_1^n)} = \ln \sum_{x_1^n \in \mathcal{Y}^n} \sup_{f \in \mathcal{F}} f_n(x_1^n) = V_n(\mathcal{F}) .$$

Proof. First we show the second part of the statement. Note that by the definition of p^*, its cumulative loss satisfies

$$\begin{aligned} \widehat{L}(y^n) - \inf_{f \in \mathcal{F}} L_f(y_1^n) &= \ln \frac{\sup_{f \in \mathcal{F}} f_n(y_1^n)}{p_n^*(y_1^n)} \\ &= \ln \sum_{x^n \in \mathcal{Y}^n} \sup_{f \in \mathcal{F}} f_n(x^n), \end{aligned}$$

which is independent of y^n and so p^* is indeed an equalizer.

To show that p^* is minimax optimal, let $p \neq p^*$ be an arbitrary predictor. Then since $\sum_{y_1^n \in \mathcal{Y}^n} p_n(y^n) = \sum_{y_1^n \in \mathcal{Y}^n} p_n^*(y_1^n) = 1$, for some $y_1^n \in \mathcal{Y}^n$ we must have $p_n(y_1^n) < p_n^*(y_1^n)$. But then, for this y_1^n,

$$\ln \frac{\sup_{f \in \mathcal{F}} f_n(y_1^n)}{p_n(y_1^n)} > \ln \frac{\sup_{f \in \mathcal{F}} f_n(y_1^n)}{p_n^*(y_1^n)} = \text{const.} = V_n(p^*, \mathcal{F})$$

by the equalizer property. Hence,

$$V_n(p, \mathcal{F}) = \sup_{y_1^n \in \mathcal{Y}^n} \ln \frac{\sup_{f \in \mathcal{F}} f_n(y_1^n)}{p_n(y_1^n)} > V_n(p^*, \mathcal{F})$$

which proves the theorem. ∎

Even though we have been able to determine the minimax optimal predictor explicitely, note that the practical implementation of the predictor may be problematic. First of all, we have determined the predictor via the joint probabilities it assigns to all strings of length n, and calculation of the actual predictions $p_t^*(y_t|y_1^{t-1})$ involves sums of exponentially many terms. Also, as it is pointed out in the exercises, previous knowledge of the total length n of the sequence to be predicted is necessary to determine p^*.

The theorem not only describes the minimax optimal predictor but also gives a useful formula for the minimax cumulative loss $V_n(\mathcal{F})$ which we study in detail in subsequent sections.

3.5 Examples

Next we work out a few simple examples to understand better the behavior of the minmax loss $V_n(\mathcal{F})$.

Finite classes. To start with the simplest possible case, consider a finite class of experts with $|\mathcal{F}| = N$. Then clearly

$$\begin{aligned} V_n(\mathcal{F}) &= \ln \sum_{y_1^n \in \mathcal{Y}^n} \sup_{f \in \mathcal{F}} f_n(y_1^n) \\ &\leq \ln \sum_{y_1^n \in \mathcal{Y}^n} \sum_{f \in \mathcal{F}} f(y_1^n) \\ &= \ln \sum_{f \in \mathcal{F}} \sum_{y_1^n \in \mathcal{Y}^n} f(y_1^n) \\ &= \ln N \ . \end{aligned}$$

Of course, we already knew this. In fact, the mixture predictor described in Section 3.1 achieves the same bound. The mixture predictor has obvious computational advantages over the normalized maximum likelihood predictor, and does not suffer from the "horizon-dependence" of the latter mentioned above. These bounds suggest that one does not lose much by using the simple uniform mixture predictor instead of the optimal but horizon-dependent predictor. We will see later that this fact remains true in more general settings, even for certain infinite classes of experts. However, as it is pointed out in the exercises, even if $\mathcal{F}$ is finite, in some cases $V_n(\mathcal{F})$ may be significantly smaller than the worst-case loss achievable by any mixture predictor.

Constant experts. Next we consider the class $\mathcal{F}$ of all experts such that for each $f \in \mathcal{F}$, $f_t(j|y_1^{t-1}) = f(j)$ (with $f(j) \geq 0$, $\sum_{j=1}^m f(j) = 1$) independently of t and y_1^{t-1}. In other words, $\mathcal{F}$ contains all predictors f_n such that the associated probability distribution over $\mathcal{Y}^n$ is a product distribution with identical components. Here we only consider the case when $m = 2$. This will simplify the notation and the calculations while the main ideas are all present. Generalization to $m > 2$ is straightforward though sometimes cumbersome.

If $m = 2$, (i.e., $\mathcal{Y} = \{1, 2\}$ and $\mathcal{X} = \{(q, 1-q) \in \mathbb{R}^2 : q \in [0, 1]\}$), each expert in $\mathcal{F}$ may be identified with a number $q \in [0, 1]$ representing $q = f(1)$. Thus, this expert predicts, at each time t, according to the vector $(q, 1-q) \in \mathcal{X}$, regardless of t and the past outcomes $y_1, \ldots, y_{t-1}$. We call this class the class of *constant experts*. Next we determine the asymptotic value of the minimax loss $V_n(\mathcal{F})$ for this class.

Theorem 15 *The minimax loss $V_n(\mathcal{F})$ of the class $\mathcal{F}$ of all constant experts over the alphabet $\mathcal{Y} = \{1, 2\}$ defined above satisfies*

$$V_n(\mathcal{F}) = \frac{1}{2}\ln n + \frac{1}{2}\ln\frac{\pi}{2} + \varepsilon_n$$

where $\varepsilon_n \to 0$ as $n \to \infty$.

PROOF. Recall that by Theorem 14,

$$V_n(\mathcal{F}) = \ln \sum_{y_1^n \in \mathcal{Y}^n} \sup_{f \in \mathcal{F}} f_n(y_1^n).$$

Now assume that the number of 1's in the sequence $y_1^n \in \{1, 2\}^n$ is n_1 and the number of 2's is $n_2 = n - n_1$. Then for the expert f which predicts according to $(q, 1 - q)$,

$$f_n(y_1^n) = q^{n_1}(1 - q)^{n_2} \ .$$

Then it is easy to see, for example, by differentiating the logarithm of the above expression, that this is maximized for $q = n_1/n$, and therefore

$$V_n(\mathcal{F}) = \ln \sum_{y_1^n \in \mathcal{Y}^n} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2} \ .$$

Since there are $\binom{n}{n_1}$ sequences containing exactly n_1 ones, we have

$$V_n(\mathcal{F}) = \ln \sum_{n_1=1}^{n-1} \binom{n}{n_1} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2} \ .$$

We show the proof of the upper bound $V_n(\mathcal{F}) \le (1/2)\ln n + (1/2)\ln(\pi/2) + o(1)$, while the similar proof of the lower bound is left as an exercise. Recall Stirling's formula

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/(12n)} \le n! \le \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/(12n+1)}$$

(see, e.g., Feller [20]). Using this to approximate the binomial coefficients, each term of the sum may be bounded as

$$\binom{n}{n_1} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2} \le \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{n_1 n_2}} e^{1/(12n+1)} \ .$$

Thus,

$$V_n(\mathcal{F}) \le \ln\left(\sqrt{\frac{n}{2\pi}}e^{1/(12n+1)}\sum_{n_1=1}^{n-1}\frac{1}{\sqrt{n_1 n_2}}\right) .$$

Writing the last sum in the expression as

$$\sum_{n_1=1}^{n-1}\frac{1}{\sqrt{n_1 n_2}} = \sum_{n_1=1}^{n-1}\frac{1}{n}\frac{1}{\sqrt{\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)}} ,$$

we notice that it is just a Riemann approximation of the integral

$$\int_0^1 \frac{1}{\sqrt{x(1-x)}}dx = \pi$$

(see the exercises). This implies that $\lim_{n\to\infty}\sum_{n_1=1}^{n-1} 1/\sqrt{n_1 n_2} = \pi$ and so

$$V_n(\mathcal{F}) \le \ln\left(\sqrt{\frac{n\pi}{2}}(1+o(1))\right)$$

as desired. ∎

Remark. In the general case, when $m \ge 2$ is a positive integer, Theorem 15 becomes

$$V_n(\mathcal{F}) = \frac{m-1}{2}\ln n + \ln\frac{\Gamma(1/2)^m}{\Gamma(m/2)} + o(1)$$

where Γ denotes the Gamma function. The fact that the minimax loss grows as a constant times $\ln n$ the constant being the half of the number of "free parameters" is a general phenomenon. We study $V_n(\mathcal{F})$ in a much more general setting in Section 3.9.

3.6 The Laplace mixture

The purpose of this section is to introduce the idea of mixture predictors discussed in Section 3.2 to certain uncountably infinite classes of experts. For simplicity we show the idea for the class of all constant experts introduced in Section 3.5. Once again, we gain simplicity by considering only the case of $m = 2$. That is, the outcome space is $\mathcal{Y} = \{1, 2\}$ and $\mathcal{X} = \{(q, 1-q) \in \mathbb{R}^2 : q \in [0, 1]$ so that each expert in $\mathcal{F}$ predicts, at each time t, according to the vector $(q, 1-q) \in \mathcal{X}$, regardless of t and the past outcomes $y_1, \ldots, y_{t-1}$. Theorem 15 shows that $V_n(\mathcal{F}) \approx (1/2)\ln n + (1/2)\ln(\pi/2)$.

In Section 3.5 we pointed out that, for finite classes of experts, the exponentially weighted average predictor, to each sequence y_1^n, simply assigns the avergage of the probabilities assigned by each experts, that is, $\widehat{p}_n(y_1^n) = (1/N)\sum_{i=1}^N f_n^{(i)}(y_1^n)$.

This idea may be generalized in a natural way to the class of constant experts. As before, let n_1 and n_2 denote the number of 1's and 2's in a sequence y_1^n. Then the probability assigned to such a sequence by any expert in the class has the form $q^{n_1}(1-q)^{n_2}$. Thus, we may define the *Laplace mixture* of these experts by the one which assigns the average of all these probabilities, averaged according to the uniform distribution over the class $\mathcal{F}$, that is,

$$\widehat{p}(y_1^n) = \int_0^1 q^{n_1}(1-q)^{n_2}dq \ .$$

After calculating this integral, it will be easy to understand the behavior of the predictor.

Lemma 5

$$\int_0^1 q^{n_1}(1-q)^{n_2}dq = \frac{1}{(n+1)\binom{n}{n_1}} \ .$$

Proof. We may prove the equality by a backwards induction with respect to n_1. If $n_1 = n$, we clearly have $\int_0^1 q^n dq = 1/(n+1)$ as desired. On the other hand, assuming

$$\int_0^1 q^{n_1+1}(1-q)^{n_2-1}dq = \frac{1}{(n+1)\binom{n}{n_1+1}}$$

and integrating by parts, we obtain

$$\begin{aligned}\int_0^1 q^{n_1}(1-q)^{n_2}dq &= \frac{n-n_1}{n_1+1}\int_0^1 q^{n_1+1}(1-q)^{n_2-1}dq \\ &= \frac{n-n_1}{n_1+1}\cdot\frac{1}{(n+1)\binom{n}{n_1+1}} \\ &= \frac{1}{(n+1)\binom{n}{n_1}} . \quad \Box\end{aligned}$$

The first thing we observe is that the actual predictions of the Laplace mixture predictor can be calculated very easily, and have a natural interpretation. Assume that the number of 1's and 2's in past outcome sequence y_1^{t-1} are t_1 and t_2. Then the probability the Laplace predictor assigns to the next outcome being 1 is, by Lemma 5,

$$\widehat{p}_t(1|y_1^{t-1}) = \frac{\widehat{p}_t(y_1^{t-1}1)}{\widehat{p}_{t-1}(y_1^{t-1})} = \frac{\frac{1}{(t+1)\binom{t}{t_1+1}}}{\frac{1}{t\binom{t-1}{t_1}}} = \frac{t_1+1}{t+1} .$$

Similarly, of course, $\widehat{p}_t(2|y_1^{t-1}) = (t_2+1)/(t+1)$. We may interpret $\widehat{p}_t(1|y_1^{t-1})$ as a slight modification of the relative frequency $t_1/(t-1)$.

To analyze the cumulative loss of the Laplace mixture, all we need is a simple property of binomial coefficients.

Lemma 6 *For all* $1 \le k \le n$,

$$\binom{n}{k} \le \frac{1}{\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k}} .$$

PROOF. If the random variables $Y_1, \ldots, Y_n$ are drawn i.i.d. according to the distribution $\mathbb{P}\{Y_i = 1\} = 1 - \mathbb{P}\{Y_i = 2\} = k/n$ then the probability that exactly k of them equals 1 is $\binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k}$ and therefore this last expression cannot be larger than 1. ∎

Theorem 16 *The excess cumulative logarithmic loss of the Laplace mixture predictor satisfies*

$$\sup_{y_1^n \in \{1,2\}^n} \left(\widehat{L}(y_1^n) - \inf_{f \in \mathcal{F}} L_f(y_1^n) \right) = \ln(n+1) \ .$$

PROOF. Let n_1 and n_2 denote the number of 1's and 2's in the sequence y_1^n. We have already observed in the proof of Theorem 15 that

$$\sup_{q \in [0,1]} q^{n_1}(1-q)^{n_2} = \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2} \ .$$

Thus, the excess loss of the predictor for such a sequence is

$$\begin{aligned}
\widehat{L}(y_1^n) - \inf_{f \in \mathcal{F}} L_f(y_1^n) &= \ln \frac{\sup_{q \in [0,1]} q^{n_1}(1-q)^{n_2}}{\int_0^1 q^{n_1}(1-q)^{n_2} dq} \\
&= \ln \frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}}{\frac{1}{(n+1)\binom{n}{n_1}}} \quad \text{(by Lemma 5)} \\
&\le \ln(n+1) \quad \text{(by Lemma 6)} \ .
\end{aligned}$$

Equality is achieved for the sequence $y_1^n = (1, 1, \ldots, 1)$. ∎

Thus, the extremely simple Laplace mixture predictor achieves an excess cumulative loss which is of the same order of magnitude as that of the minimax optimal predictor, though the leading constant is 1 instead of 1/2. In the next section we show that a slight modification of the predictor achieves this optimal leading constant as well.

3.7 A refined mixture predictor

In this section we briefly point out that with a small modification of the Laplace predictor we may obtain a mixture predictor which achieves a worst-case cumulative loss comparable to that of the minimax optimal normalized maximum likelihood predictor. We see in the proof of Theorem 16 that the Laplce mixture achieves the largest loss for sequences containing either very few 1's or very few 0's. As the optimal predictor is an equalizer (recall Theorem 14), a good predictor should attempt to achieve a nearly equal regret for all sequences. This may be done by modifying the mixture so that it gives a slightly larger weight to those experts which predict well for these critical sequences. The idea, suggested first by Krichevsky and Trofimov, is to use, instead of the uniform weighting distribution, the so-called Dirichlet-$(1/2, 1/2)$ density $1/\left(\pi\sqrt{q(1-q)}\right)$. (See the exercises for the fact that this indeed a density on $[0,1]$.) Thus, we define the Krichevsky-Trofimov mixture predictor by

$$\widehat{p}(y_1^n) = \int_0^1 \frac{1}{\pi\sqrt{q(1-q)}} q^{n_1}(1-q)^{n_2} dq$$

where we use the notation of the previous section. It is easy to see, by a recursive argument similar to the one seen in the proof of Lemma 5 that the predictions of the Krichevsky-Trofimov mixture may be calculated by

$$\widehat{p}_t(1|y_1^{t-1}) = \frac{t_1 + 1/2}{t} ,$$

a formula very similar to that obtained in the case of the Laplace predictor.

The performance of the predictor is easily bounded once the following lemma is proved:

Lemma 7 *For all* $q \in [0,1]$,

$$\int_0^1 \frac{1}{\pi\sqrt{q(1-q)}} q^{n_1}(1-q)^{n_2} dq \geq \frac{1}{2\sqrt{n}} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2} .$$

The proof is left as an exercise. Based on this lemma we immediately derive the following performace bound.

Theorem 17 *The excess cumulative logarithmic loss of the Krichevsky-Trofimov mixture predictor satisfies*

$$\sup_{y_1^n \in \{1,2\}^n} \left(\widehat{L}(y_1^n) - \inf_{f \in \mathcal{F}} L_f(y_1^n)\right) \leq \frac{1}{2}\ln n + \ln 2 .$$

PROOF.

$$\begin{aligned}
\widehat{L}(y_1^n) - \inf_{f\in\mathcal{F}} L_f(y_1^n) &= \ln \frac{\sup_{q\in[0,1]} q^{n_1}(1-q)^{n_2}}{\int_0^1 \frac{1}{\pi\sqrt{q(1-q)}} q^{n_1}(1-q)^{n_2}\,dq} \\
&= \ln \frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}}{\int_0^1 \frac{1}{\pi\sqrt{q(1-q)}} q^{n_1}(1-q)^{n_2}\,dq} \\
&\leq \ln\left(2\sqrt{n}\right) \quad \text{(by Lemma 7)}
\end{aligned}$$

as desired. ∎

Remark. The Krichevsky-Trofimov mixture estimate may be generalized to the class of all constant experts when the outcome space is $\mathcal{Y} = \{1, \ldots, m\}$ where $m \geq 2$ is an arbitrary integer. In this case the mixture is calculated with respect to the so-called Dirichlet$(1/2, \cdots, 1/2)$ density

$$\phi(p) = \frac{\Gamma(m/2)}{\Gamma(1/2)^m} \prod_{j=1}^{m} \frac{1}{\sqrt{p(j)}}$$

over the probability simplex $\mathcal{X}$ containing all vectors $p = (p(1), \ldots, p(m)) \in \mathbb{R}^m$ with nonnegative components adding up to one. Analogously to the bound of Theorem 17 above, one may show that the worst-case excess loss of the obtained predictor

$$\widehat{p}(y_1^n) = \int_{\mathcal{X}} \prod_{j=1}^{m} p(j)^{n_j} \phi(p)\,dp$$

(where $n_1, \ldots, n_m$ denote the number of occurences of each symbol in the string y_1^n) is upper bounded by

$$\frac{m-1}{2}\ln n + \ln\frac{\Gamma(1/2)^m}{\Gamma(m/2)} + \frac{m-1}{2}\ln 2 + o(1)$$

This upper bound exceeds the minimax loss by just a constant $\frac{m-1}{2}\ln 2$.

3.8 Prediction with side information

The purpose of this section is to extend the framework of prediction when the predictor has access to certain "side information". At each time instant, before making his prediction, a side-information symbol drawn from a finite set is revealed to the predictor. In our setup this side information is completely arbitrary, it may contain any external information and it may even depend on the sequence to be predicted. The formal setup is the following.

We consider prediction of sequences taking values in a finite alphabet $\mathcal{Y} = \{1, \ldots, m\}$. Let K be a positive integer, and let $\mathcal{G}_1, \ldots, \mathcal{G}_K$ be "base" classes of static predictors. (We assume the static property for convenience, the definitions and the results below may be generalized easily.) The class $\mathcal{G}_j$ contains predictors of the form

$$g^{(j)}(y^n) = \prod_{t=1}^{n} g_t^{(j)}(y_t)$$

where for each $t = 1, \ldots, n$, the vector $(g_t^{(j)}(1), \ldots, g_t^{(j)}(m))$ is an element of the probability simplex $\mathcal{X}$ in $\mathbb{R}^m$. At each time t, a side-information symbol $z_t \in \mathcal{Z} = \{1, \ldots, K\}$ becomes available to the predictor. The class $\mathcal{F}$ of predictors against which the predictor competes contains all predictors f of the form

$$f_t(y|y_1^{t-1}, z_t) = f_t(y|z_t) = g_t^{(z_t)}(y).$$

In other words, each f ignores the past, and uses the side information symbol z_t to pick a class index and use a predictor from $\mathcal{G}_{z_t}$. Note that in some sense f is static, but z_t may depend in an arbitrary manner on the sequence of past (or even future) outcomes.

The loss of $f \in \mathcal{F}$ for a given sequence of outcomes $y^n \in \mathcal{Y}^n$ and side information $z^n \in \mathcal{Z}^n$ is

$$-\sum_{t=1}^{n} \ln f_t(y_t|z_t) = -\sum_{t=1}^{n} \ln g_t^{(z_t)}(y_t).$$

Our goal is to define in predictors $\widehat{p}_t$ whose cumulative loss $-\sum_{t=1}^{n} \ln \widehat{p}_t(y_t|y_1^{t-1}, z_t)$ is close to that of the best expert $\inf_{f\in\mathcal{F}} - \sum_{t=1}^{n} \ln f_t(y_t|y_1^{t-1}, z_t)$ for all outcome sequences $y_1^n \in \mathcal{Y}^n$ and side information sequences $z_1^n \in \mathcal{Z}^n$. Assume that for each static expert class $\mathcal{G}_j$ we have a predictor $q^{(j)}$ with worst-case excess cululative loss

$$V_n(q^{(j)}, \mathcal{G}_j) = \sup_{y^n\in\mathcal{Y}^n} \sup_{g^{(j)}\in\mathcal{G}_j} \sum_{t=1}^{n} \ln \frac{g_t^{(j)}(y_t)}{q_t^{(j)}(y_t|y_1^{t-1})}.$$

Based on these "elementary" predictors, we may define, in a natural way, the following predictor p with side information:

$$p_t(y|y_1^{t-1}, z_t) = q_{\overline{t}_{z_t}}^{(z_t)}(y|\overline{y}_{z_t}^{t}),$$

where for each j, $\overline{y}_j^t$ denotes the sequence of those y_s's $(s < t)$ such that $z_s = j$ and $\overline{t}_j$ is the length of $\overline{y}_j^t$. In other words, the predictor p looks back at the past sequence $y_1, \ldots, y_{t-1}$ and considers only those time instants at which the side information symbol agreed with the current symbol z_t. The prediction of p is just that of $q^{(z_t)}$ based on these past symbols. The performance of p may be bounded easily as follows.

Theorem 18 *For any outcome sequence $y_1^n \in \mathcal{Y}^n$ and side-information sequence $z_1^n \in \mathcal{Z}^n$, the excess loss of predictor p with respect to all predictors in class $\mathcal{F}$ satisfies*

$$\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \ln \frac{f_t(y_t|z_t)}{p_t(y_t|y_1^{t-1}, \overline{y}_{z_t}^t)} \le \sum_{j=1}^{K} V_{\overline{n}_j}(q^{(j)}, \mathcal{G}_j),$$

where $\overline{n}_j = \sum_{t=1}^{n} I_{z_t=j}$ is the number of occurences of symbol j in the side information sequence.

PROOF.

$$
\begin{aligned}
& \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \ln \frac{f_t(y_t|z_t)}{p_t(y_t|y_1^{t-1}, z_t)} \\
&= \sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \ln \frac{g_t^{(z_t)}(y_t)}{q_t^{(z_t)}(y_t|\overline{y}_{z_t}^{t})} \quad \text{(by definition of } p\text{)} \\
&= \sup_{f\in\mathcal{F}} \sum_{j=1}^{K} \left(\sum_{t=1}^{n} \mathbb{I}_{z_t=j} \ln \frac{g_t^{(j)}(y_t)}{q_t^{(j)}(y_t|\overline{y}_j^{t})} \right) \\
&= \sum_{j=1}^{K} \sup_{g^{(j)}\in\mathcal{G}_j} \sum_{t=1}^{n} \mathbb{I}_{z_t=j} \ln \frac{g_t^{(j)}(y_t)}{q_t^{(j)}(y_t|\overline{y}_j^{t})} \\
&\qquad \text{(since the expression within the parentheses only depends on } g^{(j)}\text{)} \\
&\le \sum_{j=1}^{K} V_{\overline{n}_j}(q^{(j)}, \mathcal{G}_j). \quad \square
\end{aligned}
$$

The next simple example sheds some light on the power of this simple result.

Example. (MARKOV PREDICTORS.) Let $k \ge 1$ and consider the class $\mathcal{M}_k$ of all k-th order stationary Markov experts defined over the binary alphabet $\mathcal{Y} = \{1, 2\}$. More precisely, $\mathcal{M}_k$ contains all predictors f for which $f_t(y|y_1^{t-1})$ is a function of the last k outcomes $(y_{t-k}, \ldots, y_{t-1})$ (independently of t and outcomes y_s for $s < t - k$). Such predictors are also considered in Section 2.8 for different loss functions. As each prediction of a k-th order Markov expert is determined by the last k bits observed, we add a prefix $y_{-k+1}, \ldots, y_0$ to the sequence to predict in the same way as in Section 2.8.

To obtain an upper bound for the minimax loss $V_n(\mathcal{M}_k)$, we may use Theorem 18 in a simple way. The side information z_t is now defined as $(y_{t-k}, \ldots, y_{t-1})$, that is, z_t takes one of $K = 2^k$ values. If we define $\mathcal{G}_1 = \cdots = \mathcal{G}_K$ as the class $\mathcal{G}$ of all constant experts over the alphabet $\mathcal{Y} = \{1, 2\}$ as in the previous sections, then it is easy to see that the class $\mathcal{F}$ of all predictors using the side information $(y_{t-k}, \ldots, y_{t-1})$ is just the class $\mathcal{M}_k$. Let $q^{(1)} = \cdots = q^{(k)}$ be the Krichevsky-Trofimov predictor for class $\mathcal{G}$, and define the predictor f as in Theorem 18. Then, according to Theorem 18, for any sequence of outcomes y_{-k+1}^{n} the loss of f may be bounded as

$$\widehat{L}(y_1^n) - \inf_{f\in\mathcal{M}_k} L_f(y_1^n) \le \sum_{j=1}^{2^k} V_{\overline{n}_j}(\mathcal{G}).$$

According to Theorem 17,

$$V_n(\mathcal{G}) \le \frac{1}{2}\ln n + \ln 2$$

Using the bound above, we obtain,

$$\begin{aligned}
\widehat{L}(y_1^n) - \inf_{f\in\mathcal{M}_k} L_f(y_1^n) &\le \sum_{j=1}^{2^k} \frac{1}{2}\ln \overline{n}_j + 2^k \ln 2 \\
&\le \frac{1}{2}\ln\left(\frac{1}{2^k}\sum_{j=1}^{2^k}\overline{n}_j\right)^{2^k} + 2^k\ln 2 \\
&= \frac{2^k}{2}\ln\frac{n}{2^k} + 2^k \ln 2,
\end{aligned}$$

where we used the arithmetic-geometric mean inequality.

3.9 A general upper bound

Next we investigate the minimax loss $V_n(\mathcal{F})$ for general classes of experts. We derive a general bound which shows how the "size" of the class $\mathcal{F}$ effects the excess cumulative loss.

To any class $\mathcal{F}$ of experts, we associate the metric d defined as

$$d(f,g) = \sqrt{\sum_{t=1}^{n} \sup_{y_1^t}\left(\ln f(y_t|y_1^{t-1}) - \ln g(y_t|y_1^{t-1})\right)^2}\ .$$

Denote by $N(\mathcal{F},\varepsilon)$ the ε-covering number of $\mathcal{F}$ under the metric d. Recall that for any $\varepsilon > 0$, the ε-covering number is the cardinality of the smallest subset $\mathcal{F}' \subset F$ such that for all $f \in \mathcal{F}$ there exists a $g \in \mathcal{F}'$ such that $d(f,g) \le \varepsilon$.

Theorem 19 *For any class $\mathcal{F}$ of experts,*

$$V_n(\mathcal{F}) \leq \inf_{\varepsilon>0} \left(\ln N(\mathcal{F}, \varepsilon) + 24 \int_0^{\varepsilon} \sqrt{\ln N(\mathcal{F}, \delta)}\, d\delta \right) .$$

Note that if $\mathcal{F}$ is a finite class then as $\varepsilon \to 0$, the right-hand side converges to $\ln |\mathcal{F}|$ and therefore we recover our earlier general bound. However, even for finite classes, the right-hand side may be significantly smaller than $\ln |\mathcal{F}|$. Also, this result allows us to derive upper bounds for very general classes of experts.

As a first step in the proof of Theorem 19 we obtain a weak bound for $V_n(\mathcal{F})$. This will be later refined to prove the stronger bound of Theorem 19.

Lemma 8 *For any class $\mathcal{F}$ of experts,*

$$V_n(\mathcal{F}) \leq 24 \int_0^{D/2} \sqrt{\ln N(\mathcal{F}, \varepsilon)}\, d\varepsilon$$

where D is the diameter of $\mathcal{F}$.

PROOF. Recall that by the equalizer property of the normalized maximum likelihood predictor p^* established in Theorem 14, for all $y_1^n \in \mathcal{Y}^n$, $V_n(\mathcal{F}) = \sup_{f\in\mathcal{F}} \ln \frac{f_n(y_1^n)}{p_n^*(y_1^n)}$. Since the right-hand side is a constant function of y_1^n, we may take any weighted average of it without changing its value. The trick is to weight the average according to the probability distribution defined by p^*. Thus, we may write

$$V_n(\mathcal{F}) = \sum_{y_1^n\in\mathcal{Y}^n} \left(\sup_{f\in\mathcal{F}} \ln \frac{f_n(y_1^n)}{p_n^*(y_1^n)} \right) p_n^*(y_1^n) .$$

If we introduce a vector $Y_1^n = (Y_1, \ldots, Y_n)$ of random variables distributed according to p_n^*, we obtain

$$\begin{aligned} V_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \ln \frac{f_n(Y_1^n)}{p_n^*(Y_1^n)} \right] \\ &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \ln \frac{f_t(Y_t|Y_1^{t-1})}{p_t^*(Y_t|Y_1^{t-1})} \right] \\ &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \left(\ln \frac{f(Y_t|Y_1^{t-1})}{p^*(Y_t|Y_1^{t-1})} - \mathbb{E}\left[\ln \frac{f(Y_t|Y_1^{t-1})}{p^*(Y_t|Y_1^{t-1})} \middle| Y_1^{t-1} \right] \right) \right] . \end{aligned}$$

where the last step follows from the nonnegativity of the Kullback-Leibler divergence of the conditional densities, that is, the fact that

$$\mathbb{E}\left[\ln \frac{p^*(Y_t|Y_1^{t-1}=y_1^{t-1})}{f(Y_t|Y_1^{t-1}=y_1^{t-1})}\right] \geq 0$$

(see, e.g., [15]). Now, for each $f \in \mathcal{F}$ let

$$T_f(y_1^n) = \frac{1}{2}\sum_{t=1}^{n}\left(\ln \frac{f(y_t|y_1^{t-1})}{p^*(y_t|y_1^{t-1})} - \mathbb{E}\left[\ln \frac{f(Y_t|Y_1^{t-1})}{p^*(Y_t|Y_1^{t-1})}\middle| Y_1^{t-1}\right]\right)$$

so that we have $V_n(\mathcal{F}) \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} T_f\right]$ where we write $T_f = T_f(Y_1^n)$.

To obtain a suitable upper bound for this quantity, we apply Theorem 9. To do this, we need to show that the process $\{T_f : f \in \mathcal{F}\}$ is indeed a subgaussian family under the metric d. (Sample continuity of the process is obvious.) To this end, note that for any $f, g \in \mathcal{F}$,

$$T_f(y_1^n) - T_g(y_1^n) = \sum_{t=1}^{n} Z_t(y_1^t) \ ,$$

where

$$Z_t(y_1^t) = \frac{1}{2}\left(\ln \frac{f(y_t \mid y_1^{t-1})}{g(y_t \mid y_1^{t-1})} - \mathbb{E}\left[\ln \frac{f(Y_t \mid Y_1^{t-1}=y_1^{t-1})}{g(Y_t \mid Y_1^{t-1}=y_1^{t-1})}\right]\right) \ .$$

Now it is easy to see that $T_f - T_g = T_f(Y_1^n) - T_g(Y_1^n)$ is a sum of bounded martingale differences with respect to the sequence $Y_1, Y_2, \ldots, Y_n$, that is, each term Z_t has zero conditional mean and range bounded by $2d_t(f, g)$. Then Lemma 16 in the Appendix implies that, for all $\lambda > 0$,

$$\mathbb{E}\left[e^{\lambda(T_f - T_g)}\right] \leq \exp\left(\frac{\lambda^2}{2} d(f, g)^2\right) \ .$$

Thus, the family $\{T_f : f \in \mathcal{F}\}$ is indeed subgaussian. Hence, using that $V_n(\mathcal{F}) \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} T_f\right]$ and applying Theorem 9 we obtain the statement of the lemma. ∎

Proof of Theorem 19. To prove the main inequality, we partition $\mathcal{F}$ into small subclasses and calculate the minimax predictor for each subclass.

Lemma 8 is applied in each subclass. Finally, the optimal predictors for these subclasses are combined using the weighted average method.

Fix an arbitrary $\varepsilon > 0$ and let $\mathcal{G} = \{g_1, \ldots, g_N\}$ be an ε-covering of $\mathcal{F}$ of minimum size $N = N(\mathcal{F}, \varepsilon)$. Determine the subsets $\mathcal{F}_1, \ldots, \mathcal{F}_N$ of $\mathcal{F}$ by

$$\mathcal{F}_i = \{f \in \mathcal{F} : d(f, g_i) \leq d(f, g_j) \text{ for all } j = 1, \ldots, N\} ,$$

that is, $\mathcal{F}_i$ contains all experts which are closest to g_i in the covering. Clearly, the union of $\mathcal{F}_1, \ldots, \mathcal{F}_N$ is $\mathcal{F}$. For each $i = 1, \ldots, N$, let $g^{(i)}$ denote the normalized maximum likelihood predictor for $\mathcal{F}_i$:

$$g_n^{(i)}(y_1^n) = \frac{\sup_{f \in \mathcal{F}_i} f_n(y_1^n)}{\sum_{y_1^n \in \mathcal{Y}^n} \sup_{f \in \mathcal{F}_i} f_n(y_1^n)} .$$

Now let the predictor p_ε be the uniform mixture of the "experts" $g^{(1)}, \ldots, g^{(N)}$. Clearly, $V_n(\mathcal{F}) \leq \inf_{\varepsilon > 0} V_n(p_\varepsilon, \mathcal{F})$. So all we have to do is to bound the regret of p_ε.

To this end, fix any $y_1^n \in \mathcal{Y}^n$ and let $k = k(y_1^n)$ be such that

$$\ln \sup_{f \in \mathcal{F}} f(y_1^n) = \ln \sup_{f \in \mathcal{F}_k} f(y_1^n) .$$

Then

$$\ln \frac{\sup_{f \in \mathcal{F}} f(y_1^n)}{p_\varepsilon(y_1^n)} = \ln \frac{g^{(k)}(y_1^n)}{p_\varepsilon(y_1^n)} + \ln \frac{\sup_{\mathcal{F}_k} f(y_1^n)}{g^{(k)}(y_1^n)} .$$

On the one hand, by the upper bound for the loss of the mixture predictor,

$$\sup_{y_1^n} \ln \frac{g^{(k)}(y_1^n)}{p_\varepsilon(y_1^n)} \leq \ln N .$$

On the other hand,

$$\sup_{y_1^n} \ln \frac{\sup_{\mathcal{F}_k} f(y_1^n)}{g^{(k)}(y_1^n)} \leq \max_{1 \leq i \leq N} \sup_{y_1^n} \ln \frac{\sup_{\mathcal{F}_i} f(y_1^n)}{g^{(i)}(y_1^n)} = \max_{1 \leq i \leq N} V_n(\mathcal{F}_i) .$$

Hence, we get

$$V_n(p_\varepsilon, \mathcal{F}) \leq \ln N + \max_{1 \leq i \leq N} V_n(\mathcal{F}_i) .$$

Now note that the diameter of each element of the partition $\mathcal{F}_1, \ldots, \mathcal{F}_N$ is at most 2ε. Hence, applying Lemma 8 to each $\mathcal{F}_i$ we find that

$$
\begin{aligned}
V_n(p_\varepsilon, \mathcal{F}) &\leq \ln N + \max_{1 \leq i \leq N} 24 \int_0^\varepsilon \sqrt{\ln N(\mathcal{F}_i, \delta)}\, d\delta \\
&\leq \ln N(\mathcal{F}, \varepsilon) + 24 \int_0^\varepsilon \sqrt{\ln N(\mathcal{F}, \delta)}\, d\delta
\end{aligned}
$$

concluding the proof. ▮

Theorem 19 requires the existence of finite coverings of $\mathcal{F}$ in the metric d. But since the definition of d involves the logarithm of the experts' predictions, this is only possible if all $f_t(j|y_1^{t-1})$ are bounded away from zero. If the class of experts $\mathcal{F}$ does not satisfy this property, one may appeal to the following simple property. For simplicity, we state the lemma for the binary-alphabet case $m = 2$. Its extension to $m > 2$ is obvious.

Lemma 9 *Let $m = 2$, and let $\mathcal{F}$ be a class of experts. Define the class $\mathcal{F}^{(\delta)}$ as the set of all experts $f^{(\delta)}$ of the form*

$$f_t^{(\delta)}(1|y_1^{t-1}) = \tau_\delta(f_t(1|y_1^{t-1}))$$

where

$$
\tau_\delta(x) = \begin{cases} x & \text{if } x \in [\delta, 1-\delta] \\ \delta & \text{if } x < \delta \\ 1-\delta & \text{if } x > 1-\delta \end{cases}
$$

for some fixed $0 < \delta < 1/2$. Then $V_n(\mathcal{F}) \leq V_n(\mathcal{F}^{(\delta)}) + 2n\delta$.

Thus, to obtain bounds for $V_n(\mathcal{F})$, we may first calculate a bound for the truncated class $\mathcal{F}^{(\delta)}$ using Theorem 19 and then choose δ to optimize the right-hand side of the inequality of Lemma 9. The bound of the lemma is convenient but quite crude, and the resulted bounds are not always optimal.

PROOF. Simply observe that for any sequence y_1^n, and any $f \in \mathcal{F}$,

$$
\begin{aligned}
&\ln f_n(y_1^n) \\
&= \sum_{t=1}^{n} \ln f_t(y_t|y_1^{t-1}) \\
&\le \sum_{t=1}^{n} \left(\mathbb{I}_{\left[f_t^{(\delta)}(y_t|y_1^{t-1})<1-\delta\right]} \ln f_t^{(\delta)}(y_t|y_1^{t-1}) + \mathbb{I}_{\left[f_t^{(\delta)}(y_t|y_1^{t-1})=1-\delta\right]} \ln 1 \right) \\
&\le \sum_{t=1}^{n} \left(\mathbb{I}_{\left[f_t^{(\delta)}(y_t|y_1^{t-1})<1-\delta\right]} \ln f_t^{(\delta)}(y_t|y_1^{t-1}) + \mathbb{I}_{\left[f_t^{(\delta)}(y_t|y_1^{t-1})=1-\delta\right]} (2\delta + \ln(1-\delta)) \right) \\
&\quad \text{(since } \ln((1-\delta)+\delta) \le \ln(1-\delta) + \delta/(1-\delta) \text{ by concavity)} \\
&\le \sum_{t=1}^{n} \left(2\delta + \ln f_t^{(\delta)}(y_t|y_1^{t-1}) \right) \\
&= 2n\delta + \ln f_n^{(\delta)}(y_1^n) \ .
\end{aligned}
$$

Thus, for any predictor $\widehat{p}$,

$$
\begin{aligned}
\widehat{L}(y_1^n) - \inf_{f\in\mathcal{F}} L_f(y_1^n) &= -\ln \widehat{p}_n(y_1^n) + \sup_{f\in\mathcal{F}} \ln f_n(y_1^n) \\
&\le -\ln \widehat{p}_n(y_1^n) + \sup_{f^{(\delta)}\in\mathcal{F}^{(\delta)}} \ln f_n^{(\delta)}(y_1^n) + 2n\delta \\
&= \widehat{L}(y_1^n) - \inf_{f^{(\delta)}\in\mathcal{F}^{(\delta)}} L_{f^{(\delta)}}(y_1^n) + 2n\delta \ . \quad \square
\end{aligned}
$$

3.10 Further examples

Parametric classes. Consider first classes $\mathcal{F}$ such that there exist positive constants k and c such that for all $\varepsilon > 0$,

$$
\ln N(\mathcal{F}, \varepsilon) \le k \ln \frac{c\sqrt{n}}{\varepsilon} \ .
$$

This is the case for most "parametric" classes, that is, classes which can be parametrized by a bounded subset of $\mathbb{R}^k$ in some "smooth" way provided that all experts' predictions are bounded away from zero.

Corollary 8 *Assume that the covering numbers of the class $\mathcal{F}$ satisfy the inequality above. Then*

$$
V_n(\mathcal{F}) \le \frac{k}{2} \ln n + o(\ln n) \ .
$$

The main term $\frac{k}{2}\ln n$ is the same as the one we have seen in the case of the class of all constant experts. In that case we could derive much sharper expressions for $V_n(\mathcal{F})$ even without requiring that the experts' predictions be bounded away from zero. On the other hand, this corollary allows us to handle, in a simple way, much more complicated classes of experts. An example is provided in the exercises.

PROOF. Substituting the assumed bound for the covering numbers in the upper bound of Theorem 19, the first term of the expression is bounded by $\frac{k}{2}\ln n + k\ln c - k\ln\varepsilon$. Then the second term may be bounded as follows:

$$\begin{aligned}
24\int_0^{\varepsilon}\sqrt{\ln N(\mathcal{F},\delta)}\,d\delta \;&\le\; 48c\sqrt{kn}\int_{a_n}^{\infty} x^2 e^{-x^2}\,dx \\
&\qquad \text{(by substituting } x=\sqrt{\ln(c\sqrt{n}/\delta)} \\
&\qquad\quad \text{and writing } a_n=\sqrt{\ln(c\sqrt{n}/\varepsilon)}) \\
&=\; 48c\sqrt{kn}\left[\frac{a_n}{2c\sqrt{n}/\varepsilon}+\frac{1}{2}\int_{a_n}^{\infty} e^{-x^2}\,dx\right] \\
&\qquad \text{(by integrating by parts)} \\
&\le\; 48c\sqrt{kn}\left[\frac{a_n}{2c\sqrt{n}/\varepsilon}+\frac{1}{2a_n c\sqrt{n}/\varepsilon}\right] \\
&\qquad \text{(by using the gaussian tail estimate} \\
&\qquad\quad \textstyle\int_t^{\infty} e^{-x^2}\,dx \le e^{-t^2}/(2t)) \\
&\le\; 48\sqrt{k}a_n\varepsilon \quad (\text{whenever } e\varepsilon \le c\sqrt{n}) \\
&\le\; 48\sqrt{2}\varepsilon\sqrt{k\ln(c\sqrt{n})} \quad (\text{whenever } \varepsilon \ge 1/(c\sqrt{n}))
\end{aligned}$$

The obtained upper bound is minimized for

$$\varepsilon = \frac{1}{48\sqrt{2}}\sqrt{\frac{k}{\ln(c\sqrt{n})}}\ ,$$

so for each n so large that

$$c\sqrt{n} \ge 48\sqrt{2}\sqrt{\ln(c\sqrt{n})/k}\ ,$$

we have

$$V_n(\mathcal{F}) \le \frac{k}{2}\ln n + \frac{k}{2}\ln\frac{\ln(c\sqrt{n})}{k} + k\ln c + 6k$$

proving the statement. ∎

Nonarametric classes. Theorem 19 may also be used to handle much larger, "nonparametric" classes. In such cases the minimax loss $V_n(\mathcal{F})$ may be of a much larger order of magnitude than the logarithmic bounds characteristic of parametric classes. We work out one simple example here.

Let $\mathcal{Y} = \{0, 1\}$ be a binary alphabet, and consider the class $\mathcal{F}^{(\delta)}$ of all experts $f = \prod_t f_t$ such that $f(1|y_1^{t-1}) = f_t(1) \in [\delta, 1-\delta]$, where $\delta \in (0, 1/2)$ is some fixed constant, and for each $t = 2, 3, \ldots, n$, $f_t(1) \geq f_{t-1}(1)$. (The case when $\delta = 0$ will be treated latex by Lemma 9.) In other words, $\mathcal{F}^{(\delta)}$ contains all static experts which assign a probability to the outcome "1" in a monotonically increasing manner. To estimate the covering numbers of $\mathcal{F}^{(\delta)}$, consider the finite subclass $\mathcal{G}$ of $\mathcal{F}^{(\delta)}$ containing only those monotone experts $g = \prod_t g_t$ which take values of the form $g_t(1) = \delta + (i/k)(1-2\delta)$, $i = 0, \ldots, k$, where k is a positive integer to be specified later. It is easy to see that $|\mathcal{G}| = \binom{n+k}{k} \leq (2n)^k$ if $k \leq n$, and $|\mathcal{G}| \leq 2^k$ otherwise. On the other hand, for any $f \in \mathcal{F}^{(\delta)}$, if g is the expert in $\mathcal{G}$ which is closest to f, then for each $t \leq n$,

$$\begin{aligned} \max_{y\in\{0,1\}} |\ln f_t(y) - \ln g_t(y)| &\leq \frac{1}{\delta} \max_{y\in\{0,1\}} |f_t(y) - g_t(y)| \\ &= \frac{1}{\delta} |f_t(1) - g_t(1)| \\ &\leq \frac{1}{\delta k}. \end{aligned}$$

Thus, $d(f, g) \leq \sqrt{n}/(\delta k)$. By taking $k = \sqrt{n}/(\delta\varepsilon)$, it follows that the covering number of $\mathcal{F}^{(\delta)}$ is bounded as

$$N(\mathcal{F}^{(\delta)}, \varepsilon) \leq \begin{cases} (2n)^{\sqrt{n}/(\delta\varepsilon)} & \text{if } \varepsilon \geq \frac{1}{\delta\sqrt{n}} \\ 2^{\sqrt{n}/(\delta\varepsilon)} & \text{otherwise.} \end{cases}$$

Substituting this bound into Theorem 19, it is a matter of straightforward calculation to obtain

$$V_n(\mathcal{F}^{(\delta)}) = O\left(n^{1/3}\delta^{-2/3} \ln^{2/3} n\right).$$

Note that the radius optimizing the bound of Theorem 19 is about $\varepsilon \approx n^{1/6}\delta^{-1/3}\ln^{1/3} n$. Finally, by Lemma 9, if $\mathcal{F}$ is the class of all monotonincally predicting experts, without restricting predictions in $[\delta, 1-\delta]$, then by Lemma 9, $V_n(\mathcal{F}) \le V_n(\mathcal{F}^{(\delta)}) + 2n\delta$. By optimizing the value of δ in the obtained upper bound, we get $V_n(\mathcal{F}) = O\left(n^{3/5}\ln^{2/5} n\right)$.

3.11 Bibliographic remarks

A recent survey summarizing most of the literature of prediction under the logarithmic loss is by Merhav and Feder [37]. An excellent account of the basics of lossless data compression, including arithmetic coding, may be found in Cover and Thomas [15]. Theorem 14 is due to Shtarkov [42] just like Theorem 15, see also Freund [22], Xie and Barron [48]. The Laplace mixture predictor was introduced, in the context of universal coding, by Davisson [16], also investigated by Rissanen [40]. The refined mixture predictor presented in Section 3.7 was suggested by Krichevsky and Trofimov [33]. Lemma 7 appears in Willems, Shtarkov, and Tjalkens [47]. It is shown in [48] and [22] that the Krichevsky-Trofimov mixture, in fact, achieves a regret $(1/2)\ln n + (1/2)\ln(\pi/2) + o(1)$ optimal even in the additive constant for all sequences except for those containing very few 1's or 2's. Xie and Barron [48] refine further the mixture so that it achieves a worst-case excess cumulative loss $(1/2)\ln n + (1/2)\ln(\pi/2) + o(1)$ matching the performance of the minimax optimal predictor. Xie and Barron also derive the analogue of all these results in the general case $m \ge 2$. Theorem 19 is due to Cesa-Bianchi and Lugosi [10]. A weaker version, for classes of static experts had been proven by Opper and Haussler [38]. The subgaussian inequality Lemma 16 used in the proof of Theorem 19 is due to Hoeffding [31] and Azuma [2]. Other general bounds for certain regular parametric classes have been shown by Rissanen [41] and Yamanishi [49].

3.12 Exercises

Exercise 26 Let $\mathcal{F}$ be the class of all experts such that for each $f \in \mathcal{F}$, $f_t(j|y_1^{t-1}) = f(j)$ (with $f(j) \ge 0$, $\sum_{j=1}^{m} f(j) = 1$) independently of t and

y_1^{t-1}. For a particular sequence $y_1^n \in \mathcal{Y}^n$, determine the best expert and its cumulative loss.

Exercise 27 Formulate and prove the analog of Theorem 13 for an arbitrary exp-convave loss function.

Exercise 28 Assume that you want to bet in the horse race only once, and that you know that the j-th horse wins with probability p_j and the odds are $o_1, \ldots, o_m$. How would you distribute your money to maximize your expected winnings? Contrast your result with the setup described in Section 3.3 where the optimal betting strategy is independent of the odds.

Exercise 29 Show that there exists a class $\mathcal{F}$ of experts with cardinality $\mathcal{F}| = N$ such that for all $n \geq \log_m N$, $V_n(\mathcal{F}) \geq \ln N$. This exercise shows that the bound $\ln N$ achieved by the uniform mixture predictor is not improvable for some classes.

Exercise 30 Show that the normalized maximum likelihood predictor p_n^* is horizon-dependent in the sense that if p_t^* denotes the normalized maximum likelihood predictor for some $t < n$ (i.e., p_t' achieves the minimax loss $V_t(\mathcal{F})$) then it is not true in general that

$$\sum_{y_{t+1}^n \in \mathcal{Y}^{n-t}} p_n^*(y_1^n) = p_t'(y_1^t) \ .$$

Exercise 31 Show that

$$\int_0^1 \frac{1}{\sqrt{x(1-x)}} dx = \pi \ .$$

Hint: substitute x by $\sin^2 \alpha$.

Exercise 32 Show that for every $n > 1$ there exists a class $\mathcal{F}_n$ of two static experts such that, if $\widehat{p}$ denotes the exponentially weighted average (or mixture) predictor, then

$$\frac{V_n(\widehat{p}, \mathcal{F}_n)}{V_n(\mathcal{F}_n)} \geq c\sqrt{n} \ ,$$

for some universal constant c (see Cesa-Bianchi and Lugosi [10]). *Hint:* Let $\mathcal{Y} = \{0, 1\}$ and let $\mathcal{F}_n$ contain the two experts f, g defined by $f(1|y_1^{t-1}) = \frac{1}{2}$ and $g(1|y_1^{t-1}) = \frac{1}{2} + \frac{1}{2n}$. Show, on the one hand, that $V_n(\mathcal{F}_n) \le c_1 n^{-1/2}$ and on the other hand that $V_n(\widehat{p}, \mathcal{F}_n) \ge c_2$ for appropriate constants c_1, c_2.

Exercise 33 Complete the proof of Theorem 15 by showing that $V_n(\mathcal{F}) \ge (1/2)\ln n + (1/2)\ln(\pi/2) + o(1)$. *Hint:* The proof goes the same way as that of the upper bound, but to get the right constant you need to be a bit careful when n_1/n or n_2/n are small.

Exercise 34 Define the Laplace predictor for the class of all constant experts over the alphabet $\mathcal{Y} = \{1, 2, \ldots, m\}$ for $m > 2$. Extend the arguments of Section 3.6 to this general case. In particular, show that the worst-case excess cumulative loss of the uniformly weighted mixture predictor satisfies

$$\sup_{y_1^n \in \mathcal{Y}^n} \left(\widehat{L}(y_1^n) - \inf_{f \in \mathcal{F}} L_f(y_1^n) \right) = \ln \binom{n+m-1}{m-1} \le (m-1)\ln(n+1) \; .$$

Exercise 35 Prove Lemma 7. *Hint:* Show that the ratio of the two sides decreases if we replace y_1^n by y_1^{n+1} and thus achieves its minimum when $n = 1$. *Warning:* This requires some work.

Exercise 36 Experts with fading memory. Let $\mathcal{Y} = \{0, 1\}$, and consider the one-parameter class $\mathcal{F}$ of distributions on $\{0, 1\}^n$ containing all experts $f^{(a)}$ with $a \in [0, 1]$, where each $f^{(a)}$ is defined by its conditionals as: $f_1^{(a)}(1) = 1/2$, $f_2^{(a)}(1|y_1) = y_1$, and

$$f_t^{(a)}(1|y_1^{t-1}) = \frac{1}{t-1} \sum_{i=1}^{t-1} y_i \left(1 + \frac{a(2i-t)}{t-2} \right) ,$$

for all $y_1^{t-1} \in \{0, 1\}^{t-1}$ and for all $t > 2$. Show that

$$V_n(\mathcal{F}) = O(\ln n) + o(\ln n) \; .$$

Hint: First show using Theorem 19 that $V_n(\mathcal{F}^{(\delta)}) \le \frac{1}{2}\ln n + \frac{1}{2}\ln\ln\frac{\sqrt{n}}{\delta} + \ln\frac{1}{\delta} + O(1)$ and then use Lemma 9.

4 Sequential investment

4.1 Portfolio selection

This chapter is devoted to the investigation of sequential investment in the stock market. Imagine a market of m assets (stocks) in which the price of each stock may vary, in an arbitrary way, in each trading period (day). The investor, at the beginning of each trading period, is allowed to distribute his wealth among the m assets. His wealth changes according to the behavior of the market in that trading period. The goal of the investor is to maximize his wealth after n days of trading. Just like in other parts of these notes, we avoid any statistical assumptions about the nature of the stock market. (In fact, statistical modelling of real stock markets is a notoriously difficult problem.) The investor's wealth is compared to the maximum wealth achieved by a class of reference investment strategies (the "experts").

In the idealized stock market we assume there are no transaction costs and the amount of each stock that can be bought at any trading period is only limited by the investor's wealth at that time. Similarly, the investor can sell any quantity of the stocks he possesses at any time at the actual market price.

The model may be formalized as follows. A *market vector* $\mathbf{x} = (x_1, \ldots, x_m)$ for m assets is a vector of nonnegative real numbers representing price relatives for a given trading period. In other words, the quantity $x_i \geq 0$ denotes the ratio of closing to opening price of the i-th asset for that period. Hence, an initial wealth invested in the m assets according to fractions $Q_1, \ldots, Q_m$ multiplies by a factor of $\sum_{i=1}^{m} x_i Q_i$ at the end of period. The market behavior during n trading periods is represented by a sequence $\mathbf{x}_1^n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of market vectors. $x_{t,j}$, the j-th component of $\mathbf{x}_t$, denotes the the factor by which the wealth invested in asset j increases in the t-th period. As in Chapter 3, we denote the probability simplex in $\mathbb{R}^m$ by $\mathcal{X}$.

An *investment strategy* Q for n trading periods is a sequence $\mathbf{Q}_1, \ldots, \mathbf{Q}_n$ of vector-valued functions $\mathbf{Q}_t : \mathbb{R}_+^{t-1} \to \mathcal{X}$, where the i-th component $Q_{t,i}(\mathbf{x}_1^{t-1})$ of the vector $\mathbf{Q}_t(\mathbf{x}_1^{t-1})$ denotes the fraction of the current wealth invested in the i-th asset at the beginning of the t-th period based on the past market behavior $\mathbf{x}_1^{t-1}$. We use

$$S_n(Q, \mathbf{x}_1^n) = \prod_{t=1}^{n} \left(\sum_{i=1}^{m} x_{t,i} \, Q_{t,i}(\mathbf{x}_1^{t-1}) \right)$$

to denote the wealth factor of strategy Q after n trading periods.

Example. (BUY-AND-HOLD STRATEGIES.) The simplest investment strategies are the so called *buy-and-hold* strategies. An investor following such a strategy simply distributes his initial wealth among the m assets according to some distribution $\mathbf{Q}_1 \in \mathcal{X}$ before the first trading period, and does not trade anymore. The wealth factor of such a strategy, after n periods, is simply

$$S_n(Q, \mathbf{x}_1^n) = \sum_{j=1}^{m} Q_{1,j} \prod_{t=1}^{n} x_{t,j}.$$

Clearly, this wealth factor is at most as large as the gain $\max_{j=1,\ldots,m} \prod_{t=1}^{n} x_{t,j}$ of the best stock over the investment period, and achieves this maximal wealth if $\mathbf{Q}_1$ concentrates on the best stock.

Example. (CONSTANTLY REBALANCED PORTFOLIOS.) Another simple and important class of investment strategies is the class of *constantly rebalanced portfolios*. Such a strategy B is parametrized by a probability vector $\mathbf{B} = (B_1, \ldots, B_m) \in \mathcal{X}$, and simply $\mathbf{Q}_t(\mathbf{x}_1^{t-1}) = \mathbf{B}$ regardless of t and the past market behavior $\mathbf{x}_1^{t-1}$. Thus, an investor following such a strategy rebalances, at every trading period, his current wealth according to the distribution $\mathbf{B}$ by investing a proportion B_1 of his wealth in the first stock, a proportion B_2 in the second stock, etc. The wealth factor achieved after n trading periods is

$$S_n(\mathbf{B}, \mathbf{x}_1^n) = \prod_{t=1}^{n} \left(\sum_{i=1}^{m} x_{t,i} \, B_i \right).$$

To understand the power of constantly rebalanced strategies, consider the simple market of $m = 2$ stocks such that the sequence of market vectors is $(1, 1/2), (1, 2), (1, 1/2), (1, 2), \ldots$. Thus, the first stock maintains its value stable while the second stock is more volatile: on even days it doubles its price, while on odd days it loses half of its value. Clearly, on a long run, none of the two stocks (and therefore no buy-and-hold strategy) yields any gain. On the other hand, the invesment strategy which rebalances every day uniformly (i.e., with $\mathcal{B} = (1/2, 1/2)$) achieves an exponentially increasing wealth at a rate $(9/8)^{n/2}$.

4.2 Comparison classes, the minimax wealth ratio

The investor's objective is to achieve a wealth comparable to the best of a certain class of investment strategies regardless of the market behavior. Thus, given a class $\mathcal{Q}$ of investment strategies, we define the *worst-case logarithmic wealth ratio* of strategy P by

$$W_n(P, \mathcal{Q}) = \sup_{x_1^n} \sup_{Q \in \mathcal{Q}} \ln \frac{S_n(Q, x_1^n)}{S_n(P, x_1^n)} .$$

Clearly, the investor's goal is to choose a strategy P for which $W_n(P, \mathcal{Q})$ is as small as possible. $W_n(P, \mathcal{Q}) = o(n)$ means that the investment strategy P achieves the same exponent of growth as the best reference strategy in the class $\mathcal{Q}$ for all possible market behaviors. The *minimax logarithmic wealth ratio* is just the best possible worst-case logarithmic wealth ratio achievable by any investment strategy P:

$$W_n(\mathcal{Q}) = \inf_P W_n(P, \mathcal{Q}) .$$

Example. (FINITE CLASSES.) Assume that the investor competes against a finite class $\mathcal{Q} = \{Q^{(1)}, \dots, Q^{(N)}\}$ of investment strategies. A very simple strategy P divides the initial wealth in N equal parts and invests each part according to the "experts" $Q^{(i)}$. Then the total wealth of the strategy is

$$S_n(P, x_1^n) = \frac{1}{N} \sum_{i=1}^{N} S_n(Q^{(i)}, x_1^n)$$

and the worst-case logarithmic wealth ratio is bounded as

$$
\begin{aligned}
W_n(P, \mathcal{Q}) &= \sup_{x_1^n} \ln \frac{\max_{i=1,\dots,N} S_n(Q^{(i)}, \mathbf{x}_1^n)}{\frac{1}{N}\sum_{i=1}^{N} S_n(Q^{(i)}, \mathbf{x}_1^n)} \\
&\le \sup_{x_1^n} \ln \frac{\max_{i=1,\dots,N} S_n(Q^{(i)}, \mathbf{x}_1^n)}{\frac{1}{N}\max_{i=1,\dots,N} S_n(Q^{(i)}, \mathbf{x}_1^n)} \\
&= \ln N \ .
\end{aligned}
$$

4.3 Prediction and investment

In this section we point out an intimate connection between the sequential investment problem and the problem of prediction under the logarithmic loss studied in the previous chapter.

The first thing we observe is that any investment strategy Q over m assets may be used to define a predictor of sequences $y_1^n \in \mathcal{Y}^n$ where $\mathcal{Y} = \{1, \dots, m\}$. To do this, we simply restrict our attention to those market vectors $\mathbf{x}$ which have one component which equals to 1 and all other components equal to zero. Such market vectors are called *Kelly market vectors*. If $\mathbf{x}_1, \dots, \mathbf{x}_n$ are Kelly market vectors and we denote the index of the only nonzero component of each vector $\mathbf{x}_t$ by y_t, then we may define predictor f by

$$f_t(y|y_1^{t-1}) = Q_{t,y}(\mathbf{x}_1^{t-1}).$$

We will say that the predictor f is *induced* by the investment strategy Q. With some abuse of notation we write $S_n(Q, y_1^n)$ for $S_n(Q, \mathbf{x}_1^n)$ when $\mathbf{x}_1^n$ is a sequence of Kelly vectors determined by the sequence of indices y_1^n. Clearly, $S_n(Q, y_1^n) = f_n(y_1^n)$ if f is the predictor induced by Q. Now it is straightforward to see that the investment problem is at least as hard as the corresponding prediction problem:

Lemma 10 *Let $\mathcal{Q}$ be a class of investment strategies, and let $\mathcal{F}$ denote the class of predictors incuced by the strategies in $\mathcal{Q}$. Then the minimax loss* $V_n(\mathcal{F}) = \inf_{\widehat{p}} \sup_{y_1^n} \ln \sup_{f\in\mathcal{F}} f_n(y_1^n)/\widehat{p}(y_1^n)$ *satisfies*

$$W_n(\mathcal{Q}) \ge V_n(\mathcal{F}) \ .$$

PROOF. Let P be any investment strategy and let p be its induced predictor. Then

$$\begin{aligned}
\sup_{\mathbf{x}_1^n} \sup_{Q\in\mathcal{Q}} \ln \frac{S_n(Q,\mathbf{x}_1^n)}{S_n(P,\mathbf{x}_1^n)} &\geq \max_{y_1^n\in\mathcal{Y}^n} \sup_{Q\in\mathcal{Q}} \ln \frac{S_n(Q,y_1^n)}{S_n(P,y_1^n)} \\
&= \max_{y_1^n\in\mathcal{Y}^n} \sup_{f\in\mathcal{F}} \ln \frac{\prod_{t=1}^n f_t(y_t|y_1^{t-1})}{p(y_1^n)} \\
&= V_n(p,\mathcal{F}) \geq V_n(\mathcal{F}) \ . \quad \square
\end{aligned}$$

Surprisingly, as it turns out, the investment problem is not genuinely more difficult than that of prediction. Below we show that in many interesting cases the two problems are, in fact, equivalent in a minimax sense.

Given a prediction strategy p (i.e., a probability distribution on the set $\mathcal{Y}^n$), we define the investment strategy P as follows

$$P_{t,j}(\mathbf{x}_1^{t-1}) = \frac{\sum_{y_1^{t-1}\in\mathcal{Y}^{t-1}} p(j \mid y_1^{t-1}) p(y_1^{t-1}) \left(\prod_{s=1}^{t-1} x_{s,y_s}\right)}{\sum_{y_1^{t-1}} p(y^{t-1}) \left(\prod_{s=1}^{t-1} x_{s,y_s}\right)} \ .$$

Note that $\prod_{t=1}^n x_{t,y_t}$ may be viewed as the return of the "extremal" investment strategy that, on each trading period t, invests everything on the y_t-th asset. Clearly, the obtained investment strategy induces p, so we will say that p and P induce each other.

The next result is the key in relating the minimax wealth ratio to the minimax loss of predictors.

Theorem 20 *Let P be an investment strategy induced by a predictor p, and let $\mathcal{Q}$ be an arbitrary class of investment strategies. Then for any market sequence $\mathbf{x}_1^n$,*

$$\sup_{Q\in\mathcal{Q}} \ln \frac{S_n(Q,\mathbf{x}_1^n)}{S_n(P,\mathbf{x}_1^n)} \leq \max_{y_1^n\in\mathcal{Y}^n} \sup_{Q\in\mathcal{Q}} \ln \frac{\prod_{t=1}^n Q_{t,y_t}(\mathbf{x}_1^{t-1})}{p_n(y^n)} \ .$$

The proof of the theorem uses the following two simple lemmas. The first is an elementary inequality whose proof is left as an exercise.

Lemma 11 *Let $a_1,\ldots,a_n,b_1,\ldots,b_n$ be nonnegative numbers. Then*

$$\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \leq \max_{i=1,\ldots,n} \frac{a_i}{b_i} \ .$$

Lemma 12 *The wealth factor achieved by an investment strategy* Q *may be written as*

$$S_n(Q, \mathbf{x}_1^n) = \sum_{y_1^n \in \mathcal{Y}^n} \left(\prod_{t=1}^n x_{t,y_t} \right) \left(\prod_{t=1}^n Q_{t,y_t}(\mathbf{x}_1^{t-1}) \right) .$$

If the investment strategy P *is induced by a predictor* p*, then*

$$S_n(P, \mathbf{x}_1^n) = \sum_{y_1^n \in \mathcal{Y}^n} \left(\prod_{t=1}^n x_{t,y_t} \right) p_n(y^n) .$$

Proof. By straightforfard calculation,

$$\begin{aligned}
S_n(Q, \mathbf{x}_1^n) &= \prod_{t=1}^n \left(\sum_{j=1}^m x_{t,j}\, Q_{t,j}(\mathbf{x}_1^{t-1}) \right) \\
&= \sum_{y_1^n \in \mathcal{Y}^n} \left(\prod_{t=1}^n x_{t,y_t}\, Q_{t,y_t}(\mathbf{x}_1^{t-1}) \right) \\
&= \sum_{y_1^n \in \mathcal{Y}^n} \left(\prod_{t=1}^n x_{t,y_t} \right) \left(\prod_{t=1}^n Q_{t,y_t}(\mathbf{x}_1^{t-1}) \right) .
\end{aligned}$$

On the other hand, if an investment strategy is induced by a predictor p, then

$$\begin{aligned}
S_n(P, \mathbf{x}_1^n) &= \prod_{t=1}^n \left(\sum_{j=1}^m x_{t,j} P_{t,j}(\mathbf{x}_1^{t-1}) \right) \\
&= \prod_{t=1}^n \frac{\sum_{j=1}^m \sum_{y_1^{t-1}} p_t(y^{t-1}j) \left(\prod_{s=1}^t x_{s,y_s} \right)}{\sum_{y_1^{t-1}} p_{t-1}(y^{t-1}) \left(\prod_{s=1}^{t-1} x_{s,y_s} \right)} \\
&= \prod_{t=1}^n \frac{\sum_{y_1^t \in \mathcal{Y}^t} \left(\prod_{s=1}^t x_{s,y_s} \right) p_t(y_1^t)}{\sum_{y_1^{t-1}} \left(\prod_{s=1}^{t-1} x_{s,y_s} \right) p_{t-1}(y_1^{t-1})} \\
&= \sum_{y_1^n} \left(\prod_{t=1}^n x_{t,y_t} \right) p_n(y^n) . \quad \square
\end{aligned}$$

Proof of Theorem 20. Fix any market sequence $\mathbf{x}_1^n$ and choose a reference strategy $Q' \in \mathcal{Q}$ such that $S_n(Q', \mathbf{x}_1^n) = \sup_{Q \in \mathcal{Q}} S_n(Q, \mathbf{x}_1^n)$. To simplify notation, we write $S_n(y_1^n, \mathbf{x}_1^n) = \prod_{t=1}^n x_{t,y_t}$. Then using the expressions derived above for $S_n(Q', \mathbf{x}_1^n)$ and $S_n(P, \mathbf{x}_1^n)$, we have

$$\begin{aligned}
\frac{S_n(Q', \mathbf{x}_1^n)}{S_n(P, \mathbf{x}_1^n)} &= \frac{\sum_{y_1^n \in \mathcal{Y}} S_n(y_1^n, \mathbf{x}_1^n) \left(\prod_{t=1}^n Q'_{t,y_t}(\mathbf{x}_1^{t-1})\right)}{\sum_{y_1^n} S_n(y_1^n, \mathbf{x}_1^n) p_n(y_1^n)} \\
&\leq \max_{y_1^n \in \mathcal{Y}^n : S_n(y_1^n, \mathbf{x}_1^n) > 0} \frac{S_n(y_1^n, \mathbf{x}_1^n) \prod_{t=1}^n Q'_{t,y_t}(\mathbf{x}_1^{t-1})}{S_n(y_1^n, \mathbf{x}_1^n) p(y_1^n)} \\
&\qquad \text{(by Lemma 11)} \\
&= \max_{y_1^n \in \mathcal{Y}} \frac{\prod_{t=1}^n Q'_{t,y_t}(\mathbf{x}_1^{t-1})}{p_n(y^n)} \\
&\leq \max_{y_1^n \in \mathcal{Y}^n} \sup_{Q \in \mathcal{Q}} \frac{\prod_{t=1}^n Q_{t,y_t}(\mathbf{x}_1^{t-1})}{p_n(y_1^n)} \ . \quad \square
\end{aligned}$$

An important and immediate corollary of Theorem 20 is that the minimax logarithmic wealth ratio $W_n(\mathcal{Q})$ of any class $\mathcal{Q}$ of static strategies equals the minimax cumulative logarithmic loss associated to the class of the induced predictors. To make this statement precise, similarly to the notion of static experts in prediction problems, we introduce the notion of static investment strategies. A *static* investment strategy Q satisfies $\mathbf{Q}_t(\mathbf{x}_1^{t-1}) = \mathbf{Q}_t \in \mathcal{X}$ for each $t = 1, \ldots, n$. Thus, the allocation of wealth $\mathbf{Q}_t$ for each trading period does not depend on the past market behavior.

Theorem 21 *Let $\mathcal{Q}$ be a class of static investment strategies, and let $\mathcal{F}$ denote the class of predictors induced by strategies in $\mathcal{Q}$. Then*

$$W_n(Q) = V_n(\mathcal{F}) \ .$$

Furthermore, the minimax optimal investment strategy is defined by

$$P^*_{t,j}(\mathbf{x}_1^{t-1}) = \frac{\sum_{y_1^{t-1} \in \mathcal{Y}^{t-1}} p_t^*(j \mid y_1^{t-1}) p^*_{t-1}(y_1^{t-1}) \left(\prod_{s=1}^{t-1} x_{s,y_s}\right)}{\sum_{y_1^{t-1} \in \mathcal{Y}^{t-1}} p^*_{t-1}(y^{t-1}) \left(\prod_{s=1}^{t-1} x_{s,y_s}\right)}$$

where p^* is the normalized maximum likelihood predictor

$$p_n^*(y_1^n) = \frac{\sup_{Q \in \mathcal{Q}} \prod_{t=1}^n Q_{t,y_t}}{\sum_{y_1^n \in \mathcal{Y}^n} \sup_{Q \in \mathcal{Q}} \prod_{t=1}^n Q_{t,y_t}}$$

PROOF. By Lemma 10 we have $W_n(\mathcal{Q}) \geq V_n(\mathcal{F})$, so it suffices to prove that $W_n(\mathcal{Q}) \leq V_n(\mathcal{F})$. Recall from Theorem 14 that the normalized maximum likelihood predictor p^* is minimax optimal for the class $\mathcal{F}$, that is,

$$\max_{y_1^n \in \mathcal{Y}^n} \ln \sup_{Q \in \mathcal{Q}} \frac{\prod_{t=1}^n Q_{t,y_t}}{p_n^*(y^n)} = V_n(\mathcal{F}) \ .$$

Now let P^* be the prediction strategy induced by the minimax predictor p^* for $\mathcal{Q}$. By Lemma 20 we get

$$W_n(\mathcal{Q}) \leq \sup_{x_1^n} \sup_{Q \in \mathcal{Q}} \frac{S_n(Q, x_1^n)}{S_n(P^*, x_1^n)} \leq \max_{y_1^n \in \mathcal{Y}^n} \sup_{Q \in \mathcal{Q}} \frac{\prod_{t=1}^n Q_{t,y_t}}{p_n^*(y^n)} = V_n(\mathcal{F}) \ .$$

The fact that the worst-case wealth ratio $\sup_{x_1^n} \sup_{Q \in \mathcal{Q}} S_n(Q, x_1^n)/S_n(P^*, x_1^n)$ achieved by the strategy P^* equals W_n follows by the inequality above and the fact that $V_n(\mathcal{F}) \leq W_n(\mathcal{Q})$. ∎

Example. (CONSTANTLY REBALANCED PORTFOLIOS) Consider now the class $\mathcal{Q}$ of all constantly rebalanced portfolios. It is immediate to see that the stategies of this class induce the "constant" predictors studied in Sections 3.5, 3.6, and 3.7. Thus, combining Theorem 21 with the remark following Theorem 15, we obtain the following expression for the behavior of the minimax logarithmic wealth ratio:

$$W_n(\mathcal{Q}) = \frac{m-1}{2} \ln n + \ln \frac{\Gamma(1/2)^m}{\Gamma(m/2)} + o(1)$$

This result shows that the wealth $S_n(P^*, \mathcal{Q})$ of the minimax optimal investment strategy given by Theorem 21 comes within a factor of $n^{(m-1)/2}$ of the best possible constantly rebalanced portfolio, regardless of the market behavior. Since, typically, $\sup_{Q \in \mathcal{Q}} S_n(Q, x_1^n)$ increases exponentially with n, this factor becomes negligible on the long run.

4.4 Universal portfolios

Just like in the case of the prediction problem of Section 3, the minimax optimal solution for the investment problem is not feasible in practice. In this section we introduce computationally more attractive methods, close in spirit to the mixture predictors of Sections 3.6 and 3.7.

For simplicity, and for its importance, in this section we restrict our attention to the class $\mathcal{Q}$ of all constantly rebalanced portfolios. Recall that each strategy Q in this class is determined by a vector $\mathbf{B} = (B_1, \ldots, B_m)$ in the probability simplex $\mathcal{X}$ in $\mathbb{R}^m$. In order to compete with the best strategy in $\mathcal{Q}$, we introduce the *universal portfolio* strategy P by

$$P_{t,j}(\mathbf{x}_1^{t-1}) = \frac{\int_{\mathcal{X}} B_j S_{t-1}(\mathbf{B}, \mathbf{x}_1^{t-1})\phi(\mathbf{B})d\mathbf{B}}{\int_{\mathcal{X}} S_{t-1}(\mathbf{B}, \mathbf{x}_1^{t-1})\phi(\mathbf{B})d\mathbf{B}} \ , j = 1, \ldots, m; t = 1, \ldots, n$$

where ϕ is a density function on $\mathcal{X}$. In the simplest case ϕ is just the uniform density, though we will see that it may be advantageous to consider nonuniform densities such as the Dirichlet$(1/2, \cdots, 1/2)$ density. In any case, the universal portfolio is a weighted average of the strategies in $\mathcal{Q}$, weighted by their past performance.

The wealth achieved by the universal portfolio is just the average of the wealths achieved by the individual strategies in the class. This may be easily seen by observing that

$$\begin{aligned}
S_n(P, \mathbf{x}_1^n) &= \prod_{t=1}^{n} \sum_{j=1}^{m} P_{t,j}(\mathbf{x}_1^{t-1}) x_{t,j} \\
&= \prod_{t=1}^{n} \frac{\int_{\mathcal{X}} \sum_{j=1}^{m} x_{t,j} B_j S_{t-1}(\mathbf{B}, \mathbf{x}_1^{t-1})\phi(\mathbf{B})d\mathbf{B}}{\int_{\mathcal{X}} S_{t-1}(\mathbf{B}, \mathbf{x}_1^{t-1})\phi(\mathbf{B})d\mathbf{B}} \\
&= \prod_{t=1}^{n} \frac{\int_{\mathcal{X}} S_t(\mathbf{B}, \mathbf{x}_1^{t})\phi(\mathbf{B})d\mathbf{B}}{\int_{\mathcal{X}} S_{t-1}(\mathbf{B}, \mathbf{x}_1^{t-1})\phi(\mathbf{B})d\mathbf{B}} \\
&= \int_{\mathcal{X}} S_n(\mathbf{B}, \mathbf{x}_1^n)\phi(\mathbf{B})d\mathbf{B}
\end{aligned}$$

since the product is telescoping.

This simple observation is the key in establishing performance bounds for the universal portfolio:

Theorem 22 *If ϕ is the uniform density on the probability simplex $\mathcal{X}$ in $\mathbb{R}^m$, then the wealth achieved by the universal portfolio satisfies*

$$\sup_{\mathbf{x}_1^n} \sup_{\mathbf{B} \in \mathcal{X}} \frac{S_n(\mathbf{B}, \mathbf{x}_1^n)}{S_n(P, \mathbf{x}_1^n)} \le (m-1)\ln(n+1) \ .$$

If the universal portfolio is defined using the Dirichlet$(1/2, \cdots, 1/2)$ *density ϕ, then*

$$\sup_{\mathbf{x}_1^n} \sup_{\mathbf{B} \in \mathcal{X}} \frac{S_n(\mathbf{B}, \mathbf{x}_1^n)}{S_n(P, \mathbf{x}_1^n)} \le \frac{m-1}{2} \ln n + \ln \frac{\Gamma(1/2)^m}{\Gamma(m/2)} + \frac{m-1}{2} \ln 2 + o(1) \ .$$

The second statement shows that the universal portfolio based on the Dirichlet$(1/2, \cdots, 1/2)$ density comes within a constant of the minimax optimal investment strategy.

Proof. First recall that each constantly rebalanced portfolio strategy indexed by $\mathbf{B}$ is induced by the "constant" predictor $p^{\mathbf{B}}$ which assigns probability

$$p_n^{\mathbf{B}}(y_1^n) = B_1^{n_1} \cdots B_m^{n_m}$$

to each string y_1^n in which the number of occurences of symbol j is n_j $(j = 1, \ldots, m)$. By Lemma 12, the wealth achieved by such a strategy is

$$S_n(\mathbf{B}, \mathbf{x}_1^n) = \sum_{y_1^n} \left(\prod_{t=1}^n x_{t,y_t} \right) p_n^{\mathbf{B}}(y_1^n) \ .$$

Using the fact that the wealth achieved by the universal portfolio is just the average of the wealths achieved by the strategies in $\mathcal{Q}$, we have that

$$\begin{aligned} S_n(P, \mathbf{x}_1^n) &= \int_{\mathcal{X}} S_n(\mathbf{B}, \mathbf{x}_1^n) \phi(\mathbf{B}) d\mathbf{B} \\ &= \sum_{y_1^n} \left(\prod_{t=1}^n x_{t,y_t} \right) \int_{\mathcal{X}} p_n^{\mathbf{B}}(y_1^n) \phi(\mathbf{B}) d\mathbf{B} \ . \end{aligned}$$

The last expression shows that the universal portfolio P is induced by the mixture predictor

$$p_n(y_1^n) = \int_{\mathcal{X}} p_n^{\mathbf{B}}(y_1^n) \phi(\mathbf{B}) d\mathbf{B} \ .$$

(Simply note that by Lemma 12 the wealth achieved by the strategy induced by the mixture predictor is the same as the wealth achieved by the universal portfolio, hence the two strategies must coincide.)

By Theorem 20,

$$\sup_{x_1^n} \sup_{\mathbf{B} \in \mathcal{X}} \ln \frac{S_n(\mathbf{B}, \mathbf{x}_1^n)}{S_n(P, \mathbf{x}_1^n)} \leq \max_{y_1^n \in \mathcal{Y}^n} \sup_{\mathbf{B} \in \mathcal{X}} \ln \frac{p_n^{\mathbf{B}}(y_1^n)}{p_n(y^n)} .$$

In other words, the worst-case logarithmic wealth factor achieved by the universal portfolio is bounded by the worst-case logarithmic loss of the mixture predictor. But we have already studied this latter quantity. In particular, if ϕ is the uniform density then p_n is just the Laplace predictor whose performance is bounded by Theorem 16 (for $m = 2$) and by Exercise 34 (for $m \geq 2$) yielding the first half of the theorem.

The second statement is obtained by noting that if the universal portfolio is defined based on the Dirichlet$(1/2, \cdots, 1/2)$ density, then p_n is the Krichevsky-Trofimov predictor whose loss is bounded in Section 3.7. ∎

4.5 Bibliographic remarks

The problem of sequential investment of arbitrary markets was introduced in a pioneering paper of Cover [13] who defined the "universal portfolio" strategy discussed in Section 4.4. The minimax value $W_n(\mathcal{Q})$ for the class of all constantly rebalanced portfolios was found by Ordentlich and Cover [39]. The bounds for the universal portfolios over the class of constantly rebalanced portfolios was obtained by Cover and Ordentlich [14]. The general results of Theorems 20 and 15 were given in [9]. Alternative approaches to the problem may be found in Hembold, Schapire, Singer, and Warmuth [30] and in Blum and Kalai [5].

4.6 Exercises

Exercise 37 Show that for any constantly rebalanced portfolio strategy $\mathbf{B}$, the achieved wealth $S_n(\mathbf{B}, \mathbf{x}_1^n)$ is invariant under permutations of the sequence $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Exercise 38 Show that the wealth $S_n(P, \mathbf{x}_1^n)$ achieved by the universal portfolio P is invariant under permutations of the sequence $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Exercise 39 UNIVERSAL PORTFOLIO EXCEEDS VALUE LINE INDEX. Let P be the universal portfolio strategy based on the uniform density ϕ. Show that the wealth achieved by P is at least as large as the geometric mean of the wealth achieved by the individual stocks, that is,

$$S_n(P, \mathbf{x}_1^n) \geq \left(\prod_{j=1}^{m} \prod_{t=1}^{n} x_{t,j} \right)^{1/m}$$

(Cover [13].) *Hint:* Use Jensen's inequality twice.

Exercise 40 Let $\mathcal{F}$ be a finite class of predictors. Show that the investment strategy induced by the mixture predictor over this class is just the strategy described in the first example of Section 4.2 based on the class $\mathcal{Q}$ of strategies induced by members of $\mathcal{F}$.

Exercise 41 Prove Lemma 11.

Exercise 42 Consider the following randomized approximation of the universal portfolio. Observe that an interpretation of the identity $S_n(P, \mathbf{x}_1^n) = \int_{\mathcal{X}} S_n(\mathbf{B}, \mathbf{x}_1^n)\phi(\mathbf{B})d\mathbf{B}$ is that the wealth achieved by the universal portfolio is the expected value, with respect to the density ϕ, of the wealth of all constantly rebalanced strategies. This expectation may be approximated by choosing randomly N vectors $\mathbf{B}_1, \ldots, \mathbf{B}_N$ according to the density ϕ and distributing the initial wealth uniformly among them just like in the example of Section 4.2. Investigate the relationship of the wealth achieved by this randomized strategy and that of the universal portfolio. What value of N do you suggest? (Blum and Kalai [5].)

Exercise 43 Consider a market of $m > 2$ assets and the class $\mathcal{Q}$ of all investment strategies which rebalance between two assets. More precisely, $\mathcal{Q}$ is the class of all constantly rebalanced portfolios such that the probability vector $\mathbf{B} \in \mathcal{X}$ characterizing the strategy has at most two nonzero components. Determine tight bounds for the minimax logarithmic wealth ratio $W_n(\mathcal{Q})$. Define and analyze a universal portfolio for this class.

Appendix: tools from probability theory

Proof of Hoeffding's inequality (Lemma 2). Since $\ln \mathbb{E}\left[e^{sX}\right] \le s\mathbb{E}X + \ln \mathbb{E}\left[e^{s(X-\mathbb{E}X)}\right]$, it suffices to show that for any random variable X with $\mathbb{E}X = 0$, $a \le X \le b$,

$$\mathbb{E}\left[e^{sX}\right] \le e^{s^2(b-a)^2/8}.$$

Note that by convexity of the exponential function

$$e^{sx} \le \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa} \quad \text{for } a \le x \le b.$$

Exploiting $\mathbb{E}X = 0$, and introducing the notation $p = -a/(b-a)$ we get

$$\begin{aligned}
\mathbb{E}e^{sX} &\le \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\
&= \left(1 - p + pe^{s(b-a)}\right) e^{-ps(b-a)} \\
&\stackrel{\text{def}}{=} e^{\phi(u)},
\end{aligned}$$

where $u = s(b-a)$, and $\phi(u) = -pu + \log(1 - p + pe^u)$. But by straightforward calculation it is easy to see that the derivative of ϕ is

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}},$$

therefore $\phi(0) = \phi'(0) = 0$. Moreover,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2} \le \frac{1}{4}.$$

Thus, by Taylor's theorem, for some $\theta \in [0, u]$,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) \le \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}. \quad \square$$

Proof of Khinchine's inequality (Lemma 6). Here we give a short proof with a suboptimal constant ($1/\sqrt{3}$ instead of $1/\sqrt{2}$). First note that for any random variable X with finite fourth moment,

$$\mathbb{E}|X| \ge \frac{\left(\mathbb{E}[X^2]\right)^{3/2}}{\left(\mathbb{E}[X^4]\right)^{1/2}}.$$

Indeed, for any $a > 0$, the function $1/x + ax^2$ is minimal on $(0, \infty)$ when $x^3 = 1/(2a)$. Thus,

$$\frac{x + ax^4}{x^2} \geq (2a)^{1/3} + \frac{a}{(2a)^{2/3}} = \frac{3}{2}(2a)^{1/3}.$$

Replace x by $|X|$ and take expectations:

$$\mathbb{E}|X| \geq \frac{3}{2}(2a)^{1/3}\mathbb{E}[X^2] - a\mathbb{E}[X^4].$$

The lower bound, considered as a function of a, is maximized if we take $a = \frac{1}{2}\left(\mathbb{E}[X^2]/\mathbb{E}[X^4]\right)^{3/2}$. Resubstitution yields the claim. Applying the above inequality for $X = \sum_{i=1}^n a_i\sigma_i$ gives

$$\mathbb{E}\left[\left|\sum_{i=1}^n a_i\sigma_i\right|\right] \geq \frac{\left(\sum_{i=1}^n a_i^2\right)^{3/2}}{\sqrt{\sum_{i=1}^n a_i^4 + 3\sum_{i\neq j} a_i^2 a_j^2}} \geq \frac{1}{\sqrt{3}}\sqrt{\sum_{i=1}^n a_i^2},$$

where we used $\sum_{i=1}^n a_i^4 + 3\sum_{i\neq j} a_i^2 a_j^2 \leq 3\left(\sum_{i=1}^n a_i^2\right)^2$. ∎

Lemma 13 *Let $\{Z_{i,t}\}$ be an $N \times n$ array of i.i.d. Rademacher random variables with distribution $\mathbb{P}[Z_{i,t} = -1] = \mathbb{P}[Z_{i,t} = 1] = 1/2$, and let $G_1, \ldots, G_N$ be independent standard normal random variables. Then*

$$\lim_{n\to\infty} \mathbb{E}\left[\max_{i=1,\ldots,N} \frac{1}{\sqrt{n}}\sum_{t=1}^n Z_{i,t}\right] = \mathbb{E}\left[\max_{i=1,\ldots,N} G_i\right]$$

Proof. Define the N-vector $X_n = (X_{n,1}, \ldots, X_{n,N})$ of components

$$X_{n,i} \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}}\sum_{t=1}^n Z_{i,t}, \quad i = 1, \ldots, N\ .$$

By the "Cramér-Wold device" (see, e.g., Billingsley [3, p. 48]), the sequence of vectors $\{X_n\}$ converges in distribution to a vector random variable $G = (G_1, \ldots, G_N)$ if and only if $\sum_{i=1}^N a_i X_{n,i}$ converges in distribution to $\sum_{i=1}^N a_i G_i$ for all possible choices of the coefficients $a_1, \ldots, a_N$. Now clearly,

$\sum_{i=1}^N a_i X_{n,i}$ converges in distribution, as $n \to \infty$, to a zero-mean normal random variable with variance $\sum_{i=1}^N a_i^2$. Then, by the Cramér-Wold device, as $n \to \infty$ the vector X_n converges in distribution to $G = (G_1, \ldots, G_k)$, where $G_1, \ldots, G_k$ are independent standard normal random variables.

Convergence in distribution is equivalent to the fact that for any bounded continuous function $\psi : \mathbb{R}^N \to \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}\left[\psi(X_{n,1}, \ldots, X_{n,N})\right] = \mathbb{E}\left[\psi(G_1, \ldots, G_N)\right] \ . \tag{7}$$

Consider, in particular, the function $\psi(x_1, \ldots, x_N) = \phi_L(\max_i x_i)$, where $L > 0$, and ϕ_L is the "thresholding" function

$$\phi_L(x) = \begin{cases} -L & \text{if } x < -L \\ x & \text{if } |x| \le L \\ L & \text{if } x > L. \end{cases}$$

Clearly, ϕ_L is bounded and continuous. Hence, by (7), we conclude

$$\lim_{n\to\infty} \mathbb{E}\left[\phi_L\left(\max_{1\le i\le N} X_{n,i}\right)\right] = \mathbb{E}\left[\phi_L\left(\max_{1\le i\le N} G_i\right)\right] \ .$$

Now note that for any $L > 0$,

$$\begin{aligned} &\mathbb{E}\left[\max_{1\le i\le N} X_{n,i}\right] \\ &\quad\ge\ \mathbb{E}\left[\phi_L\left(\max_{1\le i\le N} X_{n,i}\right)\right] + \mathbb{E}\left[\left(L + \max_{1\le i\le N} X_{n,i}\right) \mathbb{I}_{\{\max_{1\le i\le N} X_{n,i} < -L\}}\right], \end{aligned}$$

where

$$\begin{aligned}
&\left|\mathbb{E}\left[\left(L+\max_{1\le i\le N} X_{n,i}\right) I_{\{\max_{1\le i\le N} X_{n,i}<-L\}}\right]\right| \\
&\le \mathbb{E}\left[\left(\left|\max_{1\le i\le N} X_{n,i}\right|-L\right) I_{\{|\max_{1\le i\le N} X_{n,i}|-L>0\}}\right] \\
&= \int_0^\infty \mathbb{P}\left\{\left|\max_{1\le i\le N} X_{n,i}\right|>L+u\right\} du \\
&= \int_L^\infty \mathbb{P}\left\{\left|\max_{1\le i\le N} X_{n,i}\right|>u\right\} du \\
&\le \int_L^\infty N \max_{1\le i\le N} \mathbb{P}\{|X_{n,i}|>u\}\, du \\
&\le 2N\int_L^\infty e^{-u^2/2} du \\
&\quad \text{(by Hoeffding's inequality–see Exercise 9)} \\
&\le 2N\int_L^\infty \left(1+\frac{1}{u^2}\right) e^{-u^2/2} du \\
&= \frac{2N}{L} e^{-L^2/2}.
\end{aligned}$$

Therefore, we have that for any $L>0$,

$$\liminf_{n\to\infty} \mathbb{E}\left[\max_{1\le i\le N} X_{n,i}\right] \ge \mathbb{E}\left[\phi_L\left(\max_{1\le i\le N} G_i\right)\right] - \frac{2N}{L}e^{-L^2/2} .$$

Letting $L\to\infty$ on the right-hand side, we see that

$$\liminf_{n\to\infty} \mathbb{E}\left[\max_{1\le i\le N} X_{n,i}\right] \ge \mathbb{E}\left[\max_{1\le i\le N} G_i\right] .$$

The proof that

$$\limsup_{n\to\infty} \mathbb{E}\left[\max_{1\le i\le N} X_{n,i}\right] \le \mathbb{E}\left[\max_{1\le i\le N} G_i\right]$$

is similar. ∎

Lemma 14 *Let* $G_1,\dots,G_N$ *be independent standard normal random variables. Then*

$$\lim_{N\to\infty} \frac{\mathbb{E}\left[\max_{i=1,\dots,N} G_i\right]}{\sqrt{2\ln N}} = 1 .$$

PROOF. See, for example, Galambos [24]. ∎

Proof of Theorem 9. We start the proof with the following simple lemma.

Lemma 15 *Let $\sigma > 0$, and let $X_1, \ldots, X_N$ be real-valued random variables such that for all $\lambda > 0$ and $1 \le i \le N$, $\mathbb{E}\left[e^{\lambda X_i}\right] \le e^{\lambda^2\sigma^2/2}$. Then*

$$\mathbb{E}\left[\max_{i \le N} X_i\right] \le \sigma\sqrt{2\ln N} \ .$$

PROOF. By Jensen's inequality, for all $\lambda > 0$,

$$\begin{aligned} e^{\lambda \mathbb{E}[\max_{i \le N} X_i]} &\le \mathbb{E}\left[e^{\lambda \max_{i \le N} X_i}\right] = \mathbb{E}\left[\max_{i \le N} e^{\lambda X_i}\right] \\ &\le \sum_{i=1}^{N} \mathbb{E}\left[e^{\lambda X_i}\right] \le N e^{\lambda^2\sigma^2/2} \ . \end{aligned}$$

Thus,

$$\mathbb{E}\left[\max_{i \le N} X_i\right] \le \frac{\ln N}{\lambda} + \frac{\lambda\sigma^2}{2} \ ,$$

and taking $\lambda = \sqrt{2\ln N/\sigma^2}$ yields the result. ∎

Now we are prepared to prove Theorem 9. For each $k = 0, 1, 2, \ldots$, let $\mathcal{F}^{(k)}$ be a minimal cover of $\mathcal{F}$ of radius $D2^{-k}$. Note that $|\mathcal{F}^{(k)}| = N(\mathcal{F}, D2^{-k})$. Denote the unique element of $\mathcal{F}^{(0)}$ by f_0.

Let Ω be the common domain where the random variables T_f, $f \in \mathcal{F}$ are defined. Pick $\omega \in \Omega$ and let $f^* \in \mathcal{F}$ be such that $\sup_{f \in \mathcal{F}} T_f(\omega) = T_{f^*}(\omega)$. (Here we implicitly assume that such an element exists. The modification of the proof for the general case is straightforward.)

For each $k \ge 0$, let f_k^* denote an element of $\mathcal{F}^{(k)}$ whose distance to f^* is minimal. Clearly, $\rho(f^*, f_k^*) \le D2^{-k}$, and therefore, by the triangle inequality, for each $k \ge 1$,

$$\rho(f_{k-1}^*, f_k^*) \le \rho(f^*, f_k^*) + \rho(f^*, f_{k-1}^*) \le 3D2^{-k} \ . \tag{8}$$

Clearly, $\lim_{k\to\infty} f_k^* = f^*$, and so by the sample continuity of the process,

$$\sup_f T_f(\omega) = T_{f^*}(\omega) = T_{f_0}(\omega) + \sum_{k=1}^{\infty} \left(T_{f_k^*}(\omega) - T_{f_{k-1}^*}(\omega) \right) ,$$

and therefore

$$\mathbb{E}\left[\sup_f T_f\right] \leq \sum_{k=1}^{\infty} \mathbb{E}\left[\max_{f,g} (T_f - T_g)\right]$$

where the max is taken over all pairs $(f, g) \in \mathcal{F}^{(k)} \times \mathcal{F}^{(k-1)}$ such that $\rho(f, g) \leq 3D2^{-k}$.

Noting that there are at most $N(\mathcal{F}, D2^{-k})^2$ of these pairs, and recalling that $\{T_f : f \in \mathcal{F}\}$ is subgaussian in the metric ρ, we can apply Lemma 15 using (8). Thus, for each $k \geq 1$,

$$\mathbb{E}\left[\max_{f,g} (T_f - T_g)\right] \leq 3D2^{-k}\sqrt{2 \ln N(\mathcal{F}, D2^{-k})^2} .$$

Summing over k, we obtain

$$\begin{aligned} \mathbb{E}\left[\sup_f T_f\right] &\leq \sum_{k=1}^{\infty} 3D2^{-k}\sqrt{2 \ln N(\mathcal{F}, D2^{-k})^2} \\ &= 12 \sum_{k=1}^{\infty} D2^{-(k+1)}\sqrt{\ln N(\mathcal{F}, D2^{-k})} \\ &\leq 12 \int_0^{D/2} \sqrt{\ln N_\rho(\mathcal{F}, \varepsilon)}\, d\varepsilon , \end{aligned}$$

as desired. ∎

Hoeffding-Azuma inequality.

A sequence of random variables $V_1, V_2, \ldots$ is a *martingale difference sequence* with respect to the the sequence of random variables $X_1, X_2, \ldots$ if for every $i > 0$ V_i is a function of $X_1, \ldots, X_i$, and

$$\mathbb{E}[V_{i+1}|X_1, \ldots, X_i] = 0 \text{ with probability one.}$$

Lemma 16 *Let* $V_1, V_2, \dots$ *be a martingale difference sequence with respect to some sequence* $X_1, X_2, \dots,$ *such that* $V_i \in [a_i, b_i]$ *for some constants* $a_i \leq b_i$*. If* $S_k = \sum_{i=1}^{k} V_i$ *then for any* $s > 0$*,*

$$\mathbb{E}\left[e^{sS_n}\right] \leq e^{s^2 \sum_{i=1}^{n} (b_i - a_i)^2/8} \ .$$

PROOF.

$$\begin{aligned}
\mathbb{E}\left[e^{sS_n}\right] &= \mathbb{E}\left[e^{sS_{n-1}} \mathbb{E}\left[e^{sV_n} | X_1, \dots, X_{n-1}\right]\right] \\
&\leq \mathbb{E}\left[e^{sS_{n-1}} e^{s^2 (b_n - a_n)^2/8}\right] \\
&= e^{s^2 (b_n - a_n)^2/8} \mathbb{E}\left[e^{sS_{n-1}}\right]
\end{aligned}$$

where we applied Lemma 2. The desired inequality is obtained by iterating the argument. ∎

References

[1] P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. Technical Report NC-TR-00-083, NeuroCOLT, 2000.

[2] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.

[3] P. Billingsley. *Convergence of Probability Measures*. John Wiley, New York, 1968.

[4] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

[5] A. Blum and A. Kalai. Universal portfolios with and without transaction costs. *Machine earning*, 35:193–205, 1999.

[6] N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. In *Proceedings of the 10th Annual Conference on Computational Learning Theory*, pages 163–170. ACM Press, 1997.

[7] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, D. Haussler, R. Schapire, and M.K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[8] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Annals of Statistics*, 27(6), 1999.

[9] N. Cesa-Bianchi and G. Lugosi. Minimax values and entropy bounds for portfolio selection problems. In *Proceedings of the First World Congress of the Game Theory Society*, 2000.

[10] N. Cesa-Bianchi and G. Lugosi. Worst-case bounds for the logarithmic loss of predictors. *Machine Learning*, 43:247–264, 2001.

[11] T.H. Chung. *Minimax Learning in Iterated Games via Distributional Majorization*. PhD thesis, Stanford University, 1994.

[12] T.M. Cover. Behavior of sequential predictors of binary sequences. In *Proceedings of the 4th Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 263–272. Publishing House of the Czechoslovak Academy of Sciences, Prague, 1965.

[13] T.M. Cover. Universal portfolios. *Mathematical Finance*, 1:1–29, 1991.

[14] T.M. Cover and E. Ordentlich. Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42:348–363, 1996.

[15] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley, New York, 1991.

[16] L. D. Davisson. Universal lossless coding. *IEEE Transactions on Information Theory*, 19:783–795, 1973.

[17] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The* L_1 *View*. John Wiley, New York, 1985.

[18] R.M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1978.

[19] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38:1258–1270, 1992.

[20] W. Feller. *An Introduction to Probability Theory and its Applications, Vol.1*. John Wiley, New York, 1968.

[21] D. Foster and R. Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29:7–36, 1999.

[22] Y. Freund. Predicting a binary sequence as well as the best biased coin. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 89–98. Association of Computing Machinery, New York, 1996.

[23] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[24] J. Galambos. *The Asymptotic Theory of Extreme Order Statistics*. R.E. Kreiger, 1987.

[25] A.J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. In *Proceedings of the 10th Annual Conference on Computational Learning Theory*, pages 171–183. ACM Press, 1997.

[26] P. Hall and C.C. Heyde. *Martingale Limit Theory and its Application*. Academic Press, New York, 1980.

[27] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the theory of games*, 3:97–139, 1957.

[28] S. Hart and A. Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, to appear, 2000.

[29] D. Haussler, J. Kivinen, and M. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44:1906–1925, 1998.

[30] D. P. Hembold, R. E. Schapire, Y. Singer, and M. K. Warmuth. On-line portfolio selection using multiplicative updates. *Mathematical Finance*, 8:325–344, 1998.

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

[32] J. Kivinen and M. K. Warmuth. Averaging expert predictions. In H. U. Simon P. Fischer, editor, *Computational Learning Theory: Proceedings of the Fourth European Conference, EuroCOLT'99*, pages 153–167. Springer, Berlin, 1999. Lecture Notes in Artificial Intelligence 1572.

[33] R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27:199–207, 1981.

[34] M. Ledoux and M. Talagrand. *Probability in Banach Space*. Springer-Verlag, New York, 1991.

[35] N. Littlestone. Learning quickly when irrelevant attributes abound: a new linear threshold algorithm. *Machine Learning*, 2:85–318, 1988.

[36] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[37] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147, 1998.

[38] M. Opper and D. Haussler. Worst case prediction over sequences under log loss. In G. Cybenko, D. O'Leary, and J. Rissanen, editors, *The Mathematics of Information Coding, Extraction, and Distribution*. Springer Verlag, New York, 1997.

[39] E. Ordentlich and T.M. Cover. The cost of achieving the best portfolio in hindsight. *Mathematics of Operations Research*, 23:960–982, 1998.

[40] J. Rissanen. Complexity of strings in the class of Markov sources. *IEEE Transactions on Information Theory*, 32:526–532, 1986.

[41] J. Rissanen. Fischer information and stochastic complexity. *IEEE Transactions on Information Theory*, 42:40–47, 1996.

[42] Y. Shtarkov. Universal sequential coding of single messages. *Problems in Information Transmission*, 23:3–17, 1987.

[43] S.J. Szarek. On the best constants in the Khintchine inequality. *Studia Mathematica*, 63:197–208, 1976.

[44] V.G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 372–383. Association of Computing Machinery, New York, 1990.

[45] V.G. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56:153–173, 1998.

[46] V.G. Vovk. Competitive on-line statistics. *International Statistical Review*, 2001, to appear.

[47] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory*, 41:653–664, 1995.

[48] Q. Xie and A.R. Barron. Asymptotic minimax loss for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46:431–445, 2000.

[49] K. Yamanishi. A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transactions on Information Theory*, 44:1424–1440, 1998.