2

Prediction with Expert Advice

The model of prediction with expert advice, introduced in this chapter, provides the foundations to the theory of prediction of individual sequences that we develop in the rest of the book.

Prediction with expert advice is based on the following protocol for sequential decisions: the decision maker is a forecaster whose goal is to predict an unknown sequence $y_1, y_2 \dots$ of elements of an *outcome space* $\mathcal{Y}$. The forecaster's predictions $\widehat{p}_1, \widehat{p}_2 \dots$ belong to a *decision space* $\mathcal{D}$, which we assume to be a convex subset of a vector space. In some special cases we take $\mathcal{D} = \mathcal{Y}$, but in general $\mathcal{D}$ may be different from $\mathcal{Y}$.

The forecaster computes his predictions in a sequential fashion, and his predictive performance is compared to that of a set of reference forecasters that we call *experts*. More precisely, at each time t the forecaster has access to the set $\{f_{E,t} : E \in \mathcal{E}\}$ of expert predictions $f_{E,t} \in \mathcal{D}$, where $\mathcal{E}$ is a fixed set of indices for the experts. On the basis of the experts' predictions, the forecaster computes his own guess $\widehat{p}_t$ for the next outcome y_t. After $\widehat{p}_t$ is computed, the true outcome y_t is revealed.

The predictions of forecaster and experts are scored using a nonnegative *loss function* $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$.

This prediction protocol can be naturally viewed as the following repeated game between "forecaster," who makes guesses $\widehat{p}_t$, and "environment," who chooses the expert advice $\{f_{E,t} : E \in \mathcal{E}\}$ and sets the true outcomes y_t.

PREDICTION WITH EXPERT ADVICE

Parameters: decision space $\mathcal{D}$, outcome space $\mathcal{Y}$, loss function ℓ, set $\mathcal{E}$ of expert indices.

For each round $t = 1, 2, \dots$

(1) the environment chooses the next outcome y_t and the expert advice $\{f_{E,t} \in \mathcal{D} : E \in \mathcal{E}\}$; the expert advice is revealed to the forecaster;
(2) the forecaster chooses the prediction $\widehat{p}_t \in \mathcal{D}$;
(3) the environment reveals the next outcome $y_t \in \mathcal{Y}$;
(4) the forecaster incurs loss $\ell(\widehat{p}_t, y_t)$ and each expert E incurs loss $\ell(f_{E,t}, y_t)$.

The forecaster's goal is to keep as small as possible the *cumulative regret* (or simply *regret*) with respect to each expert. This quantity is defined, for expert E, by the sum

$$R_{E,n} = \sum_{t=1}^{n} \bigl(\ell(\widehat{p}_t, y_t) - \ell(f_{E,t}, y_t) \bigr) = \widehat{L}_n - L_{E,n},$$

where we use $\widehat{L}_n = \sum_{t=1}^{n} \ell(\widehat{p}_t, y_t)$ to denote the forecaster's cumulative loss and $L_{E,n} = \sum_{t=1}^{n} \ell(f_{E,t}, y_t)$ to denote the cumulative loss of expert E. Hence, $R_{E,n}$ is the difference between the forecaster's total loss and that of expert E after n prediction rounds. We also define the instantaneous regret with respect to expert E at time t by $r_{E,t} = \ell(\widehat{p}_t, y_t) - \ell(f_{E,t}, y_t)$, so that $R_{E,n} = \sum_{t=1}^{n} r_{E,t}$. One may think about $r_{E,t}$ as the regret the forecaster feels of not having listened to the advice of expert E right after the tth outcome y_t has been revealed.

Throughout the rest of this chapter we assume that the number of experts is finite, $\mathcal{E} = \{1, 2, \ldots, N\}$, and use the index $i = 1, \ldots, N$ to refer to an expert. The goal of the forecaster is to predict so that the regret is as small as possible for all sequences of outcomes. For example, the forecaster may want to have a vanishing per-round regret, that is, to achieve

$$\max_{i=1,\ldots,N} R_{i,n} = o(n) \quad \text{or, equivalently,} \quad \frac{1}{n} \left(\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \right) \stackrel{n\to\infty}{\longrightarrow} 0,$$

where the convergence is uniform over the choice of the outcome sequence and the choice of the expert advice. In the next section we show that this ambitious goal may be achieved by a simple forecaster under mild conditions.

The rest of the chapter is structured as follows. In Section 2.1 we introduce the important class of weighted average forecasters, describe the subclass of potential-based forecasters, and analyze two important special cases: the polynomially weighted average forecaster and the exponentially weighted average forecaster. This latter forecaster is quite central in our theory, and the following four sections are all concerned with various issues related to it: Section 2.2 shows certain optimality properties, Section 2.3 addresses the problem of tuning dynamically the parameter of the potential, Section 2.4 investigates the problem of obtaining improved regret bounds when the loss of the best expert is small, and Section 2.5 investigates the special case of differentiable loss functions. Starting with Section 2.6, we discover the advantages of rescaling the loss function. This simple trick allows us to derive new and even sharper performance bounds. In Section 2.7 we introduce and analyze a weighted average forecaster for rescaled losses that, unlike the previous ones, is not based on the notion of potential. In Section 2.8 we return to the exponentially weighted average forecaster and derive improved regret bounds based on rescaling the loss function. Sections 2.9 and 2.10 address some general issues in the problem of prediction with expert advice, including the definition of minimax values. Finally, in Section 2.11 we discuss a variant of the notion of regret where discount factors are introduced.

2.1 Weighted Average Prediction

A natural forecasting strategy in this framework is based on computing a *weighted average* of experts' predictions. That is, the forecaster predicts at time t according to

$$\widehat{p}_t = \frac{\sum_{i=1}^N w_{i,t-1} f_{i,t}}{\sum_{j=1}^N w_{j,t-1}},$$

where $w_{1,t-1}, \dots, w_{N,t-1} \geq 0$ are the weights assigned to the experts at time t. Note that $\widehat{p}_t \in \mathcal{D}$, since it is a convex combination of the expert advice $f_{1,t}, \dots, f_{N,t} \in \mathcal{D}$ and $\mathcal{D}$ is convex by our assumptions. As our goal is to minimize the regret, it is reasonable to choose the weights according to the regret up to time $t-1$. If $R_{i,t-1}$ is large, then we assign a large weight $w_{i,t-1}$ to expert i, and vice versa. As $R_{i,t-1} = \widehat{L}_{t-1} - L_{i,t-1}$, this results in weighting more those experts i whose cumulative loss $L_{i,t-1}$ is small. Hence, we view the weight as an arbitrary increasing function of the expert's regret. For reasons that will become apparent shortly, we find it convenient to write this function as the derivative of a nonnegative, convex, and increasing function $\phi : \mathbb{R} \to \mathbb{R}$. We write ϕ' to denote this derivative. The forecaster uses ϕ' to determine the weight $w_{i,t-1} = \phi'(R_{i,t-1})$ assigned to the ith expert. Therefore, the prediction $\widehat{p}_t$ at time t of the weighted average forecaster is defined by

$$\widehat{p}_t = \frac{\sum_{i=1}^N \phi'(R_{i,t-1}) f_{i,t}}{\sum_{j=1}^N \phi'(R_{j,t-1})} \qquad \text{(weighted average forecaster)}.$$

Note that this is a legitimate forecaster as $\widehat{p}_t$ is computed on the basis of the experts' advice at time t and the cumulative regrets up to time $t-1$.

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

***Lemma* 2.1.** *If the loss function ℓ is convex in its first argument, then*

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

Proof. Using Jensen's inequality, for all $y \in \mathcal{Y}$,

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

Rearranging, we obtain the statement. ■

The simple observation of the lemma above allows us to interpret the weighted average forecaster in an interesting way. To do this, introduce the *instantaneous regret vector*

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

and the corresponding *regret vector* $\mathbf{R}_n = \sum_{t=1}^n \mathbf{r}_t$. It is convenient to introduce also a *potential function* $\Phi : \mathbb{R}^N \to \mathbb{R}$ of the form

$$\Phi(\mathbf{u}) = \psi\left(\sum_{i=1}^N \phi(u_i)\right) \qquad \text{(potential function)},$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is any nonnegative, increasing, and twice differentiable function, and $\psi : \mathbb{R} \to \mathbb{R}$ is any nonnegative, strictly increasing, concave, and twice differentiable auxiliary function.

Using the notion of potential function, we can give the following equivalent definition of the weighted average forecaster

$$\widehat{p}_t = \frac{\sum_{i=1}^N \nabla\Phi(\mathbf{R}_{t-1})_i \, f_{i,t}}{\sum_{j=1}^N \nabla\Phi(\mathbf{R}_{t-1})_j}$$

where $\nabla\Phi(\mathbf{R}_{t-1})_i = \partial\Phi(\mathbf{R}_{t-1})/\partial R_{i,t-1}$. We say that a forecaster defined as above is *based on the potential* Φ. Even though the definition of the weighted average forecaster is independent of the choice of ψ (the derivatives ψ' cancel in the definition of $\widehat{p}_t$ above), the proof of the main result of this chapter, Theorem 2.1, reveals that ψ plays an important role in the analysis. We remark that convexity of ϕ is not needed to prove Theorem 2.1, and this is the reason why convexity is not mentioned in the above definition of potential function. On the other hand, all forecasters in this book that are based on potential functions and have a vanishing per-round regret are constructed using a convex ϕ (see also Exercise 2.2).

The statement of Lemma 2.1 is equivalent to

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

The notation $\mathbf{u} \cdot \mathbf{v}$ stands for the the inner product of two vectors defined by $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \cdots + u_N v_N$. We call the above inequality *Blackwell condition* because of its similarity to a key property used in the proof of the celebrated Blackwell's approachability theorem. The theorem, and its connection to the above inequality, are explored in Sections 7.7 and 7.8. Figure 2.1 shows an example of a prediction satisfying the Blackwell condition.

The Blackwell condition shows that the function Φ plays a role vaguely similar to the potential in a dynamical system: the weighted average forecaster, by forcing the regret vector to point away from the gradient of Φ irrespective to the outcome y_t, tends to keep the point $\mathbf{R}_t$ close to the minimum of Φ. This property, in fact, suggests a simple analysis because the increments of the potential function Φ may now be easily bounded by Taylor's theorem. The role of the function ψ is simply to obtain better bounds with this argument.

The next theorem applies to any forecaster satisfying the Blackwell condition (and thus not only to weighted average forecasters). However, it will imply several interesting bounds for different versions of the weighted average forecaster.

***Theorem* 2.1.** *Assume that a forecaster satisfies the Blackwell condition for a potential* $\Phi(\mathbf{u}) = \psi\left(\sum_{i=1}^N \phi(u_i)\right)$. *Then, for all* $n = 1, 2, \ldots,$

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

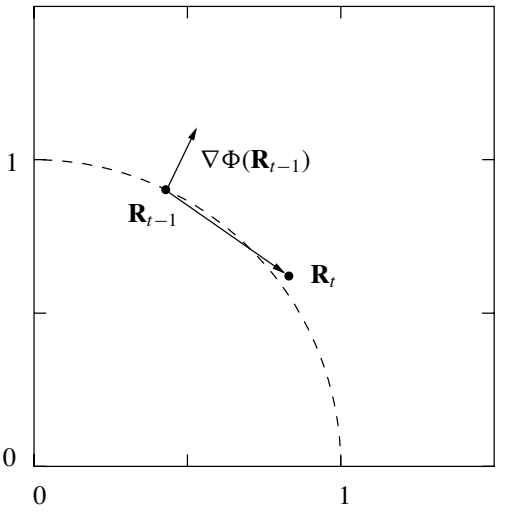

Figure 2.1. An illustration of the Blackwell condition with $N = 2$. The dashed line shows the points in regret space with potential equal to 1. The prediction at time t changed the potential from $\Phi(\mathbf{R}_{t-1}) = 1$ to $\Phi(\mathbf{R}_t) = \Phi(\mathbf{R}_{t-1} + \mathbf{r}_t)$. Though $\Phi(\mathbf{R}_t) > \Phi(\mathbf{R}_{t-1})$, the inner product between $\mathbf{r}_t$ and the gradient $\nabla\Phi(\mathbf{R}_{t-1})$ is negative, and thus the Blackwell condition holds.

where

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

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

$$\begin{aligned}
\Phi(\mathbf{R}_t) &= \Phi(\mathbf{R}_{t-1} + \mathbf{r}_t) \\
&= \Phi(\mathbf{R}_{t-1}) + \nabla\Phi(\mathbf{R}_{t-1}) \cdot \mathbf{r}_t + \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \left.\frac{\partial^2\Phi}{\partial u_i \partial u_j}\right|_\xi r_{i,t} r_{j,t} \\
&\qquad \text{(where } \xi \text{ is some vector in } \mathbb{R}^N\text{)} \\
&\le \Phi(\mathbf{R}_{t-1}) + \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \left.\frac{\partial^2\Phi}{\partial u_i \partial u_j}\right|_\xi r_{i,t} r_{j,t}
\end{aligned}$$

where the inequality follows by the Blackwell condition. Now straightforward calculation shows that

$$\begin{aligned}
&\sum_{i=1}^N \sum_{j=1}^N \left.\frac{\partial^2\Phi}{\partial u_i \partial u_j}\right|_\xi r_{i,t} r_{j,t} \\
&\quad = \psi''\left(\sum_{i=1}^N \phi(\xi_i)\right) \sum_{i=1}^N \sum_{j=1}^N \phi'(\xi_i)\phi'(\xi_j) r_{i,t} r_{j,t} \\
&\qquad + \psi'\left(\sum_{i=1}^N \phi(\xi_i)\right) \sum_{i=1}^N \phi''(\xi_i) r_{i,t}^2
\end{aligned}$$

$$
\begin{aligned}
&= \psi''\left(\sum_{i=1}^N \phi(\xi_i)\right)\left(\sum_{i=1}^N \phi'(\xi_i) r_{i,t}\right)^2 + \psi'\left(\sum_{i=1}^N \phi(\xi_i)\right)\sum_{i=1}^N \phi''(\xi_i) r_{i,t}^2 \\
&\le \psi'\left(\sum_{i=1}^N \phi(\xi_i)\right)\sum_{i=1}^N \phi''(\xi_i) r_{i,t}^2 \quad \text{(since } \psi \text{ is concave)} \\
&\le C(\mathbf{r}_t)
\end{aligned}
$$

where at the last step we used the definition of $C(\mathbf{r}_t)$. Thus, we have obtained $\Phi(\mathbf{R}_t) - \Phi(\mathbf{R}_{t-1}) \le C(\mathbf{r}_t)/2$. The proof is finished by summing this inequality for $t = 1, \ldots, n$. ■

Theorem 2.1 can be used as follows. By monotonicity of ψ and ϕ,

$$
\psi\left(\phi\left(\max_{i=1,\ldots,N} R_{i,n}\right)\right) = \psi\left(\max_{i=1,\ldots,N} \phi(R_{i,n})\right) \le \psi\left(\sum_{i=1}^N \phi(R_{i,n})\right) = \Phi(\mathbf{R}_n).
$$

Note that ψ is invertible by the definition of the potential function. If ϕ is invertible as well, then we get

$$
\max_{i=1,\ldots,N} R_{i,n} \le \phi^{-1}\left(\psi^{-1}\left(\Phi(\mathbf{R}_n)\right)\right),
$$

where $\Phi(\mathbf{R}_n)$ is replaced with the bound provided by Theorem 2.1. In the first of the two examples that follow, however, ϕ is not invertible, and thus $\max_{i=1,\ldots,N} R_{i,n}$ is directly majorized using a function of the bound provided by Theorem 2.1.

Polynomially Weighted Average Forecaster

Consider the *polynomially weighted average forecaster* based on the potential

$$
\Phi_p(\mathbf{u}) = \left(\sum_{i=1}^N (u_i)_+^p\right)^{2/p} = \|\mathbf{u}_+\|_p^2 \qquad \text{(polynomial potential)},
$$

where $p \ge 2$. Here $\mathbf{u}_+$ denotes the vector of positive parts of the components of $\mathbf{u}$. The weigths assigned to the experts are then given by

$$
w_{i,t-1} = \nabla\Phi_p(\mathbf{R}_{t-1})_i = \frac{2(R_{i,t-1})_+^{p-1}}{\|(\mathbf{R}_{t-1})_+\|_p^{p-2}}
$$

and the forecaster's predictions are just the weighted average of the experts predictions

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

***Corollary* 2.1.** *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 $n \ge 1$, the*

regret of the polynomially weighted average forecaster satisfies

$$\widehat{L}_n - \min_{i=1,\dots,N} L_{i,n} \le \sqrt{n(p-1)N^{2/p}}.$$

This shows that, for all $p \ge 2$, the per-round regret converges to zero at a rate $O(1/\sqrt{n})$ uniformly over the outcome sequence and the expert advice. The choice $p = 2$ yields a particularly simple algorithm. On the other hand, the choice $p = 2 \ln N$ (for $N > 2$), which approximately minimizes the upper bound, leads to

$$\widehat{L}_n - \min_{i=1,\dots,N} L_{i,n} \le \sqrt{ne(2 \ln N - 1)}$$

yielding a significantly better dependence on the number of experts N.

***Proof of Corollary* 2.1.** Apply Theorem 2.1 using the polynomial potential. Then $\phi(x) = x_+^p$ and $\psi(x) = x^{2/p}$, $x \ge 0$. Moreover

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

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,

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

and the conditions of Theorem 2.1 are satisfied with $C(\mathbf{r}_t) \le 2(p-1) \|\mathbf{r}_t\|_p^2$. Since $\Phi_p(\mathbf{0}) = 0$, Theorem 2.1, together with the boundedness of the loss function, implies that

$$\left(\sum_{i=1}^N \left(R_{i,n} \right)_+^p \right)^{2/p} = \Phi_p(\mathbf{R}_n) \le (p-1) \sum_{t=1}^n \|\mathbf{r}_t\|_p^2 \le n(p-1)N^{2/p}.$$

Finally, since

$$\widehat{L}_n - \min_{i=1,\dots,N} L_{i,n} = \max_{i=1,\dots,N} R_{i,n} \le \left(\sum_{i=1}^N \left(R_{i,n} \right)_+^p \right)^{1/p}$$

the result follows. ■

***Remark* 2.1.** We have defined the polynomial potential as $\Phi_p(\mathbf{u}) = \|\mathbf{u}_+\|_p^2$, which corresponds to taking $\psi(x) = x^{2/p}$. Recall that ψ does not have any influence on the prediction, it only has a role in the analysis. The particular form analyzed here is chosen by convenience, but there are other possibilities leading to similar results. For example, one may argue that it is more natural to take $\psi(x) = x^{1/p}$, which leads to the potential function $\Phi(\mathbf{u}) = \|\mathbf{u}_+\|_p$.

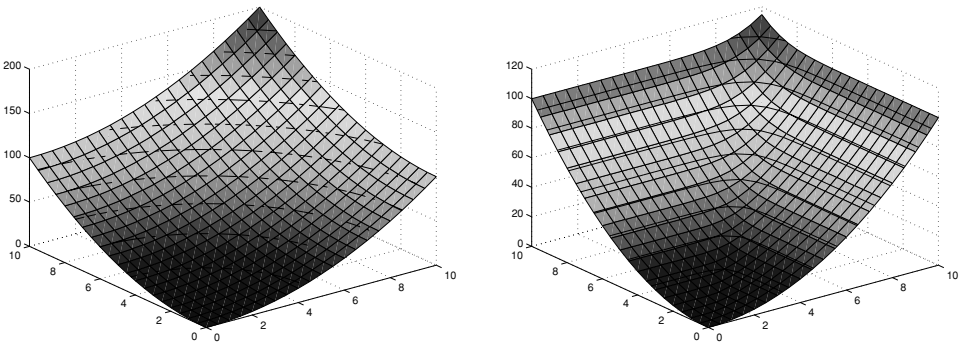

Figure 2.2. Plots of the polynomial potential function $\Phi_p(\mathbf{u})$ for $N = 2$ experts with exponents $p = 2$ and $p = 10$.

We leave it as an exercise to work out a bound similar to that of Corollary 2.1 based on this choice.

Exponentially Weighted Average Forecaster

Our second main example is the *exponentially weighted average forecaster* based on the potential

$$\Phi_\eta(\mathbf{u}) = \frac{1}{\eta} \ln \left(\sum_{i=1}^{N} e^{\eta u_i} \right) \qquad \text{(exponential potential)},$$

where η is a positive parameter. In this case, the weigths assigned to the experts are of the form

$$w_{i,t-1} = \nabla \Phi_\eta(\mathbf{R}_{t-1})_i = \frac{e^{\eta R_{i,t-1}}}{\sum_{j=1}^{N} e^{\eta R_{j,t-1}}},$$

and the weighted average forecaster simplifies to

$$\widehat{p}_t = \frac{\sum_{i=1}^{N} \exp\left(\eta \left(\widehat{L}_{t-1} - L_{i,t-1}\right)\right) f_{i,t}}{\sum_{j=1}^{N} \exp\left(\eta \left(\widehat{L}_{t-1} - L_{j,t-1}\right)\right)} = \frac{\sum_{i=1}^{N} e^{-\eta L_{i,t-1}} f_{i,t}}{\sum_{j=1}^{N} e^{-\eta L_{j,t-1}}}.$$

The beauty of the exponentially weighted average forecaster 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. Furthermore, the weights that the forecaster assigns to the experts are computable in a simple incremental way: let $w_{1,t-1}, \ldots, w_{N,t-1}$ be the weights used at round t to compute the prediction $\widehat{p}_t = \sum_{i=1}^{N} w_{i,t-1} f_{i,t}$. Then, as one can easily verify,

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

A simple application of Theorem 2.1 reveals the following performance bound for the exponentially weighted average forecaster.

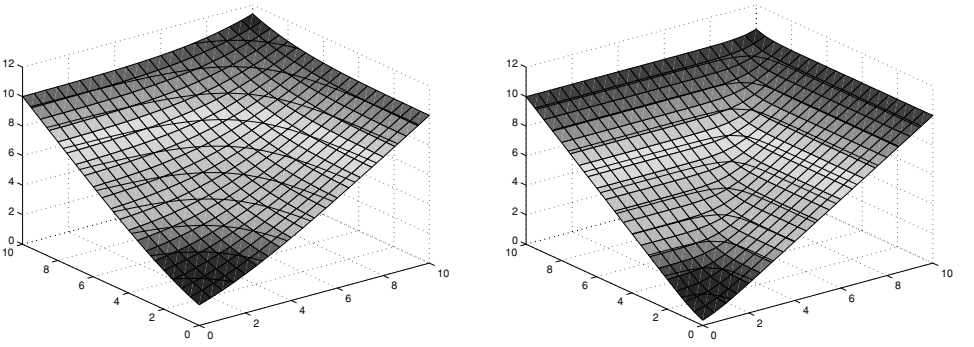

Figure 2.3. Plots of the exponential potential function $\Phi_\eta(\mathbf{u})$ for $N = 2$ experts with $\eta = 0.5$ and $\eta = 2$.

***Corollary* 2.2.** *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 regret of the exponentially weighted average forecaster satisfies*

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \le \frac{\ln N}{\eta} + \frac{n\eta}{2}.$$

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 $p = 2 \ln N$. In the next section we improve the bound of Corollary 2.2 by a direct analysis. 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 two sections we describe versions of the exponentially weighted average forecaster that do not suffer from this drawback.

***Proof of Corollary* 2.2.** Apply Theorem 2.1 using the exponential potential. Then $\phi(x) = e^{\eta x}$, $\psi(x) = (1/\eta) \ln x$, and

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

Using $\Phi_\eta(\mathbf{0}) = (\ln N)/\eta$, Theorem 2.1 implies that

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

as desired. ■

2.2 An Optimal Bound

The purpose of this section is to show that, even for general convex loss functions, the bound of Corollary 2.2 may be improved for the exponentially weighted average forecaster. The following result improves Corollary 2.2 by a constant factor. In Section 3.7 we see that the bound obtained here cannot be improved further.

***Theorem* 2.2.** *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 $y_1, \dots, y_n \in \mathcal{Y}$, the regret of the exponentially weighted average forecaster satisfies*

$$\widehat{L}_n - \min_{i=1,\dots,N} L_{i,n} \le \frac{\ln N}{\eta} + \frac{n\eta}{8}.$$

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 2.2, but now, instead of bounding the evolution of $(1/\eta) \ln \left(\sum_i e^{\eta R_{i,t}} \right)$, we bound the related quantities $(1/\eta) \ln(W_t / W_{t-1})$, where

$$W_t = \sum_{i=1}^{N} w_{i,t} = \sum_{i=1}^{N} e^{-\eta L_{i,t}}$$

for $t \ge 1$, and $W_0 = N$. In the proof we use the following classical inequality due to Hoeffding [161].

***Lemma* 2.2.** *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 Section A.1 of the Appendix.

***Proof of Theorem* 2.2**. First observe that

$$\begin{aligned} \ln \frac{W_n}{W_0} &= \ln \left(\sum_{i=1}^{N} e^{-\eta L_{i,n}} \right) - \ln N \\ &\ge \ln \left(\max_{i=1,\dots,N} e^{-\eta L_{i,n}} \right) - \ln N \\ &= -\eta \min_{i=1,\dots,N} L_{i,n} - \ln N. \end{aligned} \tag{2.1}$$

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

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

Now using Lemma 2.2, we observe that the quantity above may be upper bounded by

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

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

$$\ln \frac{W_n}{W_0} \le -\eta \widehat{L}_n + \frac{\eta^2}{8} n.$$

Combining this with the lower bound (2.1) and solving for $\widehat{L}_n$, we find that

$$\widehat{L}_n \le \min_{i=1,\ldots,N} L_{i,n} + \frac{\ln N}{\eta} + \frac{\eta}{8} n$$

as desired. ■

2.3 Bounds That Hold Uniformly over Time

As we pointed out in the previous section, the exponentially weighted average forecaster has the disadvantage that the regret bound of Corollary 2.2 does not hold uniformly over sequences of any length, but only for sequences of a given length n, where n is the value used to choose the parameter η. To fix this problem one can use the so-called "doubling trick." The idea is to partition time into periods of exponentially increasing lengths. In each period, the weighted average forecaster is used with a parameter η chosen optimally for the length of the interval. When the period ends, the weighted average forecaster is reset and then is started again in the next period with a new value for η. If the doubling trick is used with the exponentially weighted average forecaster, then it achieves, for any sequence $y_1, y_2, \ldots \in \mathcal{Y}$ of outcomes and for any $n \ge 1$,

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \le \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{\frac{n}{2} \ln N}$$

(see Exercise 2.8). This bound is worse than that of Theorem 2.2 by a factor of $\sqrt{2}/(\sqrt{2}-1)$, which is about 3.41.

Considering that the doubling trick resets the weights of the underlying forecaster after each period, one may wonder whether a better bound could be obtained by a more direct argument. In fact, we can avoid the doubling trick altogether by using the weighted average forecaster with a time-varying potential. That is, we let the parameter η of the exponential potential depend on the round number t. As the best nonuniform bounds for the exponential potential are obtained by choosing $\eta = \sqrt{8(\ln N)/n}$, a natural choice for a time-varying exponential potential is thus $\eta_t = \sqrt{8(\ln N)/t}$. By adapting the approach used to prove Theorem 2.2, we obtain for this choice of η_t a regret bound whose main term is $2\sqrt{(n/2)\ln N}$ and is therefore better than the doubling trick bound. More precisely, we prove the following result.

***Theorem* 2.3.** *Assume that the loss function ℓ is convex in its first argument and takes values in $[0, 1]$. For all $n \ge 1$ and for all $y_1, \ldots, y_n \in \mathcal{Y}$, the regret of the exponentially weighted average forecaster with time-varying parameter $\eta_t = \sqrt{8(\ln N)/t}$ satisfies*

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \le 2\sqrt{\frac{n}{2} \ln N} + \sqrt{\frac{\ln N}{8}}.$$

The exponentially weighted average forecaster with time-varying potential predicts with $\widehat{p}_t = \sum_{i=1}^N f_{i,t} w_{i,t-1} / W_{t-1}$, where $W_{t-1} = \sum_{j=1}^N w_{j,t-1}$ and $w_{i,t-1} = e^{-\eta_t L_{i,t-1}}$. The potential parameter is chosen as $\eta_t = \sqrt{a(\ln N)/t}$, where $a > 0$ is determined by the analysis. We use $w'_{i,t-1} = e^{-\eta_{t-1} L_{i,t-1}}$ to denote the weight $w_{i,t-1}$, where the parameter η_t is replaced by η_{t-1}. Finally, we use k_t to denote the expert whose loss after the first t rounds is the lowest (ties are broken by choosing the expert with smallest index). That is, $L_{k_t,t} = \min_{i \le N} L_{i,t}$. In the proof of the theorem, we also make use of the following technical lemma.

***Lemma* 2.3.** *For all $N \ge 2$, for all $\beta \ge \alpha \ge 0$, and for all $d_1, \ldots, d_N \ge 0$ such that $\sum_{i=1}^N e^{-\alpha d_i} \ge 1$,*

$$\ln \frac{\sum_{i=1}^N e^{-\alpha d_i}}{\sum_{j=1}^N e^{-\beta d_j}} \le \frac{\beta - \alpha}{\alpha} \ln N.$$

Proof. We begin by writing

$$\ln \frac{\sum_{i=1}^N e^{-\alpha d_i}}{\sum_{j=1}^N e^{-\beta d_j}} = \ln \frac{\sum_{i=1}^N e^{-\alpha d_i}}{\sum_{j=1}^N e^{(\alpha-\beta) d_j} e^{-\alpha d_j}} = -\ln \mathbb{E}\left[e^{(\alpha-\beta)D}\right] \le (\beta - \alpha)\mathbb{E}\, D$$

by Jensen's inequality, where D is a random variable taking value d_i with probability $e^{-\alpha d_i} / \sum_{j=1}^N e^{-\alpha d_j}$ for each $i = 1, \ldots, N$. Because D takes at most N distinct values, its entropy $H(D)$ is at most $\ln N$ (see Section A.2 in the Appendix). Therefore,

$$\begin{aligned}
\ln N &\ge H(D) \\
&= \sum_{i=1}^N e^{-\alpha d_i} \left(\alpha d_i + \ln \sum_{k=1}^N e^{-\alpha d_k} \right) \frac{1}{\sum_{j=1}^N e^{-\alpha d_j}} \\
&= \alpha \mathbb{E}\, D + \ln \sum_{k=1}^N e^{-\alpha d_k} \\
&\ge \alpha \mathbb{E}\, D,
\end{aligned}$$

where the last inequality holds because $\sum_{i=1}^N e^{-\alpha d_i} \ge 1$. Hence $\mathbb{E}\, D \le (\ln N)/\alpha$. As $\beta > \alpha$ by hypothesis, we can substitute the upper bound on $\mathbb{E}\, D$ in the first derivation above and conclude the proof. ■

We are now ready to prove the main theorem.

***Proof of Theorem* 2.3**. As in the proof of Theorem 2.2, we study the evolution of $\ln(W_t / W_{t-1})$. However, here we need to couple this with $\ln(w_{k_{t-1},t-1} / w_{k_t,t})$, including in both terms the time-varying parameter η_t. Keeping track of the currently best expert, k_t is used to lower bound the weight $\ln(w_{k_t,t} / W_t)$. In fact, the weight of the overall best expert (after n rounds) could get arbitrarily small during the prediction process. We thus write the

following:

$$\frac{1}{\eta_t}\ln\frac{w_{k_{t-1},t-1}}{W_{t-1}} - \frac{1}{\eta_{t+1}}\ln\frac{w_{k_t,t}}{W_t}$$
$$= \underbrace{\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\ln\frac{W_t}{w_{k_t,t}}}_{(A)} + \underbrace{\frac{1}{\eta_t}\ln\frac{w'_{k_t,t}/W'_t}{w_{k_t,t}/W_t}}_{(B)} + \underbrace{\frac{1}{\eta_t}\ln\frac{w_{k_{t-1},t-1}/W_{t-1}}{w'_{k_t,t}/W'_t}}_{(C)}.$$

We now bound separately the three terms on the right-hand side. The term (A) is easily bounded by noting that $\eta_{t+1} < \eta_t$ and using the fact that k_t is the index of the expert with the smallest loss after the first t rounds. Therefore, $w_{k_t,t}/W_t$ must be at least $1/N$. Thus we have

$$(A) = \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\ln\frac{W_t}{w_{k_t,t}} \le \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\ln N.$$

We proceed to bounding the term (B) as follows:

$$(B) = \frac{1}{\eta_t}\ln\frac{w'_{k_t,t}/W'_t}{w_{k_t,t}/W_t} = \frac{1}{\eta_t}\ln\frac{\sum_{i=1}^N e^{-\eta_{t+1}(L_{i,t}-L_{k_t,t})}}{\sum_{j=1}^N e^{-\eta_t(L_{j,t}-L_{k_t,t})}}$$
$$\le \frac{\eta_t - \eta_{t+1}}{\eta_t\eta_{t+1}}\ln N = \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\ln N,$$

where the inequality is proven by applying Lemma 2.3 with $d_i = L_{i,t} - L_{k_{t+1},t}$. Note that $d_i \ge 0$ because k_t is the index of the expert with the smallest loss after the first t rounds and $\sum_{i=1}^N e^{-\eta_{t+1}d_i} \ge 1$ as for $i = k_{t+1}$ we have $d_i = 0$. The term (C) is first split as follows:

$$(C) = \frac{1}{\eta_t}\ln\frac{w_{k_{t-1},t-1}/W_{t-1}}{w'_{k_t,t}/W'_t} = \frac{1}{\eta_t}\ln\frac{w_{k_{t-1},t-1}}{w'_{k_t,t}} + \frac{1}{\eta_t}\ln\frac{W'_t}{W_{t-1}}.$$

We treat separately each one of the two subterms on the right-hand side. For the first one, we have

$$\frac{1}{\eta_t}\ln\frac{w_{k_{t-1},t-1}}{w'_{k_t,t}} = \frac{1}{\eta_t}\ln\frac{e^{-\eta_t L_{k_{t-1},t-1}}}{e^{-\eta_t L_{k_t,t}}} = L_{k_t,t} - L_{k_{t-1},t-1}.$$

For the second subterm, we proceed similarly to the proof of Theorem 2.2 by applying Hoeffding's bound (Lemma 2.2) to the random variable Z that takes the value $\ell(f_{i,t}, y_t)$ with probability $w_{i,t-1}/W_{t-1}$ for each $i = 1, \ldots, N$:

$$\frac{1}{\eta_t}\ln\frac{W'_t}{W_{t-1}} = \frac{1}{\eta_t}\ln\sum_{i=1}^N \frac{w_{i,t-1}}{W_{t-1}}e^{-\eta_t\ell(f_{i,t},y_t)}$$
$$\le -\sum_{i=1}^N \frac{w_{i,t-1}}{W_{t-1}}\ell(f_{i,t}, y_t) + \frac{\eta_t}{8}$$
$$\le -\ell(\widehat{p}_t, y_t) + \frac{\eta_t}{8},$$

where in the last step we used the convexity of the loss ℓ. Finally, we substitute back in the main equation the bounds on the first two terms (A) and (B), and the bounds on the two

subterms of the term (C). Solving for $\ell(\widehat{p}_t, y_t)$, we obtain

$$\begin{aligned}\ell(\widehat{p}_t, y_t) \le \left(L_{k_t,t} - L_{k_{t-1},t-1}\right) + \frac{\sqrt{a \ln N}}{8}\frac{1}{\sqrt{t}} \\ + \frac{1}{\eta_{t+1}} \ln \frac{w_{k_t,t}}{W_t} - \frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t-1}}{W_{t-1}} \\ + 2\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \ln N.\end{aligned}$$

We apply the above inequality to each $t = 1, \ldots, n$ and sum up using $\sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) = \widehat{L}_n$, $\sum_{t=1}^{n} \left(L_{k_t,t} - L_{k_{t-1},t-1}\right) = \min_{i=1,\ldots,N} L_{i,n}$, $\sum_{t=1}^{n} 1/\sqrt{t} \le 2\sqrt{n}$, and

$$\sum_{t=1}^{n} \left(\frac{1}{\eta_{t+1}} \ln \frac{w_{k_t,t}}{W_t} - \frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t-1}}{W_{t-1}}\right) \le -\frac{1}{\eta_1} \ln \frac{w_{k_0,0}}{W_0} = \sqrt{\frac{\ln N}{a}}$$

$$\sum_{t=1}^{n} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) = \sqrt{\frac{n+1}{a(\ln N)}} - \sqrt{\frac{1}{a(\ln N)}}.$$

Thus, we can write the bound

$$\widehat{L}_n \le \min_{i=1,\ldots,N} L_{i,n} + \frac{\sqrt{an \ln N}}{4} + 2\sqrt{\frac{(n+1)\ln N}{a}} - \sqrt{\frac{\ln N}{a}}.$$

Finally, by overapproximating and choosing $a = 8$ to trade off the two main terms, we get

$$\widehat{L}_n \le \min_{i=1,\ldots,N} L_{i,n} + 2\sqrt{\frac{n}{2} \ln N} + \sqrt{\frac{\ln N}{8}}$$

as desired. ■

2.4 An Improvement for Small Losses

The regret bound for the exponentially weighted average forecaster shown in Theorem 2.2 may be improved significantly whenever it is known beforehand that the cumulative loss of the best expert will be small. In some cases, as we will see, this improvement may even be achieved without any prior knowledge.

To understand why one can hope for better bounds for the regret when the cumulative loss of the best expert is small, recall the simple example described in the introduction when $\mathcal{Y} = \mathcal{D} = \{0, 1\}$ and $\ell(\widehat{p}, y) = |\widehat{p} - y| \in \{0, 1\}$ (this example violates our assumption that $\mathcal{D}$ is a convex set but helps understand the basic phenomenon). If the forecaster knows in advance that one of the N experts will suffer zero loss, that is, $\min_{i=1,\ldots,N} L_{i,n} = 0$, then he may predict using the following simple "majority vote." 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 remaining experts predict 0, discard all incorrectly predicting experts after y_2 is revealed, and so on. Hence, each time the forecaster makes a mistake, at least half of the surviving experts are discarded (because the forecaster always votes according to the majority of the remaining experts). If only one expert remains, the forecaster does not make any further

mistake. Thus, the total number of mistakes of the forecaster (which, in this case, coincides with his regret) is at most $\lfloor \log_2 N \rfloor$.

In this section we show that regret bounds of the form $O(\ln N)$ are possible for all bounded and convex losses on convex decision spaces whenever the forecaster is given the information that the best expert will suffer no loss. Note that such bounds are significantly better than $\sqrt{(n/2)\ln N}$, which holds independently of the loss of the best expert.

For simplicity, we write $L_n^* = \min_{i=1,\dots,N} L_{i,n}$. We now show that whenever L_n^* is known beforehand, the parameter η of the exponentially weighted average forecaster can be chosen so that his regret is bounded by $\sqrt{2L_n^* \ln N} + \ln N$, which equals to $\ln N$ when $L_n^* = 0$ and is of order $\sqrt{n \ln N}$ when L_n^* is of order n. Our main tool is the following refinement of Theorem 2.2.

***Theorem* 2.4.** *Assume that the loss function ℓ is convex in its first argument and that it takes values in $[0, 1]$. Then for any $\eta > 0$ the regret of the exponentially weighted average forecaster satisfies*

$$\widehat{L}_n \le \frac{\eta L_n^* + \ln N}{1 - e^{-\eta}}.$$

It is easy to see that, in some cases, an uninformed choice of η can still lead to a good regret bound.

***Corollary* 2.3.** *Assume that the exponentially weighted average forecaster is used with $\eta = 1$. Then, under the conditions of Theorem 2.4,*

$$\widehat{L}_n \le \frac{e}{e-1}\left(L_n^* + \ln N\right).$$

This bound is much better than the general bound of Theorem 2.2 if $L_n^* \ll \sqrt{n}$, but it may be much worse otherwise.

We now derive a new bound by tuning η in Theorem 2.4 in terms of the total loss L_n^* of the best expert.

***Corollary* 2.4.** *Assume the exponentially weighted average forecaster is used with $\eta = \ln\left(1 + \sqrt{(2\ln N)/L_n^*}\right)$, where $L_n^* > 0$ is supposed to be known in advance. Then, under the conditions of Theorem 2.4,*

$$\widehat{L}_n - L_n^* \le \sqrt{2L_n^* \ln N} + \ln N.$$

Proof. Using Theorem 2.4, we just need to show that, for our choice of η,

$$\frac{\ln N + \eta L_n^*}{1 - e^{-\eta}} \le L_n^* + \ln N + \sqrt{2L_n^* \ln N}. \tag{2.2}$$

We start from the elementary inequality $(e^{\eta} - e^{-\eta})/2 = \sinh(\eta) \ge \eta$, which holds for all $\eta \ge 0$. Replacing the η in the numerator of the left-hand side of (2.2) with this upper bound, we find that (2.2) is implied by

$$\frac{\ln N}{1 - e^{-\eta}} + \frac{1 + e^{-\eta}}{2e^{-\eta}} L_n^* \le L_n^* + \ln N + \sqrt{2L_n^* \ln N}.$$

The proof is concluded by noting that the above inequality holds with equality for our choice of η. ■

Of course, the quantity L_n^* is only available after the nth round of prediction. The lack of this information may be compensated by letting η change according to the loss of the currently best expert, similarly to the way shown in Section 2.3 (see Exercise 2.10). The regret bound that is obtainable via this approach is of the form $2\sqrt{2L_n^* \ln N} + c \ln N$, where $c > 1$ is a constant. Note that, similarly to Theorem 2.3, the use of a time-varying parameter η_t leads to a bound whose leading term is twice the one obtained when η is fixed and chosen optimally on the basis of either the horizon n (as in Theorem 2.2) or the loss L_n^* of the best expert (as in Corollary 2.4).

***Proof of Theorem* 2.4.** The proof is a simple modification of that of Theorem 2.2. The only difference is that Hoeffding's inequality is now replaced by Lemma A.3 (see the Appendix). Recall from the proof of Theorem 2.2 that

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

We apply Lemma A.3 to the random variable X_t that takes value $\ell(f_{i,t}, y_t)$ with probability $w_{i,t-1}/W_{t-1}$ for each $i = 1, \ldots, N$. Note that by convexity of the loss function and Jensen's inequality, $\mathbb{E}\, X_t \ge \ell(\widehat{p}_t, y_t)$ and therefore, by Lemma A.3,

$$\ln \frac{\sum_{i=1}^{N} w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^{N} w_{j,t-1}} = \ln \mathbb{E}\left[e^{-\eta X_t}\right] \le \left(e^{-\eta} - 1\right) \mathbb{E}\, X_t \le \left(e^{-\eta} - 1\right) \ell(\widehat{p}_t, y_t).$$

Thus,

$$-\eta L_n^* - \ln N \le \sum_{t=1}^{n} \ln \frac{\sum_{i=1}^{N} w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^{N} w_{j,t-1}} \le \left(e^{-\eta} - 1\right) \widehat{L}_n.$$

Solving for $\widehat{L}_n$ yields the result. ■

2.5 Forecasters Using the Gradient of the Loss

Consider again the exponentially weighted average forecaster whose predictions are defined by

$$\widehat{p}_t = \frac{\sum_{i=1}^{N} w_{i,t-1} f_{i,t}}{\sum_{j=1}^{N} w_{j,t-1}},$$

where the weight $w_{i,t-1}$ for expert i at round t is defined by $w_{i,t-1} = e^{-\eta L_{i,t-1}}$. In this section we introduce and analyze a different exponentially weighted average forecaster in which the cumulative loss $L_{i,t-1}$ appearing at the exponent of $w_{i,t-1}$ is replaced by the gradient of the loss summed up to time $t-1$. This new class of forecasters will be generalized in Chapter 11, where we also provide extensive analysis and motivation.

Recall that the decision space $\mathcal{D}$ is a convex subset of a linear space. Throughout this section, we also assume that $\mathcal{D}$ is finite dimensional, though this assumption can be relaxed easily. If ℓ is differentiable, we use $\nabla \ell(\widehat{p}, y)$ to denote its gradient with respect to the first argument $\widehat{p} \in \mathcal{D}$.

Define the *gradient-based exponentially weighted average forecaster* by

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

To understand the intuition behind this predictor, note that the weight assigned to expert i is small if the sum of the inner products $\nabla \ell(\widehat{p}_s, y_s) \cdot f_{i,s}$ has been large in the past. This inner product is large if expert i's advice $f_{i,s}$ points in the direction of the largest increase of the loss function. Such a large value means that having assigned a little bit larger weight to this expert would have increased the loss suffered at time s. According to the philosophy of the gradient-based exponentially weighted average forecaster, the weight of such an expert has to be decreased.

The predictions of this forecaster are, of course, generally different from those of the standard exponentially weighted average forecaster. However, note that in the special case of binary prediction with absolute loss (i.e., if $\mathcal{D} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$ and $\ell(x, y) = |x - y|$), a setup that we study in detail in Chapter 8, the predictions of the two forecasters are identical (see the exercises).

We now show that, under suitable conditions on the norm of the gradient of the loss, the regret of the new forecaster can be bounded by the same quantity that was used to bound the regret of the standard exponentially weighted average forecaster in Corollary 2.2.

***Corollary* 2.5.** *Assume that the decision space $\mathcal{D}$ is a convex subset of the euclidean unit ball $\{q \in \mathbb{R}^d : \|q\| \leq 1\}$, the loss function ℓ is convex in its first argument and that its gradient $\nabla \ell$ exists and satisfies $\|\nabla \ell\| \leq 1$. For any n and $\eta > 0$, and for all $y_1, \ldots, y_n \in \mathcal{Y}$, the regret of the gradient-based exponentially weighted average forecaster satisfies*

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq \frac{\ln N}{\eta} + \frac{n\eta}{2}.$$

Proof. The weight vector $\mathbf{w}_{t-1} = (w_{1,t-1}, \ldots, w_{N,t-1})$ used by this forecaster has components

$$w_{i,t-1} = \exp\left(-\eta \sum_{s=1}^{t-1} \nabla \ell(\widehat{p}_s, y_s) \cdot f_{i,s}\right).$$

Observe that these weights correspond to the exponentially weighted average forecaster based on the loss function ℓ', defined at time t by

$$\ell'(q, y_t) = q \cdot \nabla \ell(\widehat{p}_t, y_t), \quad q \in \mathcal{D}.$$

By assumption, ℓ' takes values in $[-1, 1]$. Applying Theorem 2.2 after rescaling ℓ' in $[0, 1]$ (see Section 2.6), we get

$$\begin{aligned}\max_{i=1,\ldots,N} \sum_{i=1}^n (\widehat{p}_t - f_{i,t}) \cdot \nabla \ell(\widehat{p}_t, y_t) &= \sum_{t=1}^n \ell'(\widehat{p}_t, y_t) - \min_{i=1,\ldots,N} \sum_{t=1}^n \ell'(f_{i,t}, y_t) \\ &\leq \frac{\ln N}{\eta} + \frac{n\eta}{2}.\end{aligned}$$

The proof is completed by expanding $\ell(f_{i,t}, y_t)$ around $\ell(\widehat{p}_t, y_t)$ as follows:

$$\ell(\widehat{p}_t, y_t) - \ell(f_{i,t}, y_t) \le (\widehat{p}_t - f_{i,t}) \cdot \nabla \ell(\widehat{p}_t, y_t),$$

which implies that

$$\widehat{L}_n - \min_{i=1,\dots,N} L_{i,n} \le \sum_{t=1}^{n} \ell'(\widehat{p}_t, y_t) - \min_{i=1,\dots,N} \sum_{t=1}^{n} \ell'(f_{i,t}, y_t). \quad \blacksquare$$

2.6 Scaled Losses and Signed Games

Up to this point we have always assumed that the range of the loss function ℓ is the unit interval $[0, 1]$. We now investigate how scalings and translations of this range affect forecasting strategies and their performance.

Consider first the case of a loss function $\ell \in [0, M]$. If M is known, we can run the weighted average forecaster on the scaled losses ℓ/M and apply without any modification the analysis developed for the $[0, 1]$ case. For instance, in the case of the exponentially weighted average forecaster, Corollary 2.4, applied to these scaled losses, yields the regret bound

$$\widehat{L}_n - L_n^* \le \sqrt{2L_n^* M \ln N} + M \ln N.$$

The additive term $M \ln N$ is necessary. Indeed, if ℓ is such that for all $p \in \mathcal{D}$ there exist $p' \in \mathcal{D}$ and $y \in \mathcal{Y}$ such that $\ell(p, y) = M$ and $\ell(p', y) = 0$, then the expert advice can be chosen so that any forecaster incurs a cumulative loss of at least $M \log N$ on some outcome sequence with $L_n^* = 0$.

Consider now the translated range $[-M, 0]$. If we interpret negative losses as gains, we may introduce the regret $G_n^* - \widehat{G}_n$ measuring the difference between the cumulative gain $G_n^* = -L_n^* = \max_{i=1,\dots,N}(-L_{i,n})$ of the best expert and the cumulative gain $\widehat{G}_n = -\widehat{L}_n$ of the forecaster. As before, if M is known we can run the weighted average forecaster on the scaled gains $(-\ell)/M$ and apply the analysis developed for $[0, 1]$-valued loss functions. Adapting Corollary 2.4 we get a bound of the form

$$G_n^* - \widehat{G}_n \le \sqrt{2G_n^* M \ln N} + M \ln N.$$

Note that the regret now scales with the largest cumulative gain G_n^*.

We now turn to the general case in which the forecasters are scored using a generic *payoff function* $h : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$, concave in its first argument. The goal of the forecaster is to maximize his cumulative payoff. The corresponding regret is defined by

$$\max_{i=1,\dots,N} \sum_{t=1}^{n} h(f_{i,t}, y_t) - \sum_{t=1}^{n} h(\widehat{p}_t, y_t) = H_n^* - \widehat{H}_n.$$

If the payoff function h has range $[0, M]$, then it is a gain function and the forecaster plays a *gain game*. Similarly, if h has range in $[-M, 0]$, then it is the negative of a loss function and the forecaster plays a *loss game*. Finally, if the range of h includes a neighborhood of 0, then the game played by the forecaster is a *signed game*.

Translated to this terminology, the arguments proposed at the beginning of this section say that in any unsigned game (i.e., any loss game $[-M, 0]$ or gain game $[0, M]$),

rescaling of the payoffs yields a regret bound of order $\sqrt{|H_n^*|M \ln N}$ whenever M is known. In the case of signed games, however, scaling is not sufficient. Indeed, if $h \in [-M, M]$, then the reduction to the $[0, 1]$ case is obtained by the linear transformation $h \mapsto (h + M)/(2M)$. Applying this to the analysis leading to Corollary 2.4, we get the regret bound

$$H_n^* - \widehat{H}_n \le \sqrt{4(H_n^* + Mn)(M \ln N)} + 2M \ln N = O\left(M\sqrt{n \ln N}\right).$$

This shows that, for signed games, reducing to the $[0, 1]$ case might not be the best thing to do. Ideally, we would like to replace the factor n in the leading term with something like $|h(f_{i,1}, y_1)| + \cdots + |h(f_{i,n}, y_n)|$ for an arbitrary expert i. In the next sections we show that, in certain cases, we can do even better than that.

2.7 The Multilinear Forecaster

Potential functions offer a convenient tool to derive weighted average forecasters. However, good forecasters for signed games can also be designed without using potentials, as shown in this section.

Fix a signed game with payoff function $h : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$, and consider the weighted average forecaster that predicts, at time t,

$$\widehat{p}_t = \frac{\sum_{i=1}^N w_{i,t-1} f_{i,t}}{W_{t-1}},$$

where $W_{t-1} = \sum_{j=1}^N w_{j,t-1}$. The weights $w_{i,t-1}$ of this forecaster are recursively defined as follows

$$w_{i,t} = \begin{cases} 1 & \text{if } t = 0 \\ w_{i,t-1}\big(1 + \eta h(f_{i,t}, y_t)\big) & \text{otherwise,} \end{cases}$$

where $\eta > 0$ is a parameter of the forecaster. Because $w_{i,t}$ is a multilinear form of the payoffs, we call this the *multilinear forecaster*.

Note that the weights $w_{1,t}, \ldots, w_{N,t}$ cannot be expressed as functions of the regret $\mathbf{R}_t$ of components $H_{i,n} - \widehat{H}_n$. On the other hand, since $(1 + \eta h) \approx e^{\eta h}$, the regret of the multilinear forecaster can be bounded via a technique similar to the one used in the proof of Theorem 2.2 for the exponentially weighted average forecaster. We just need the following simple lemma (proof is left as exercise).

***Lemma* 2.4.** *For all $z \ge -1/2$, $\ln(1 + z) \ge z - z^2$.*

The next result shows that the regret of the multilinear forecaster is naturally expressed in terms of the squared sum of the payoffs of an arbitrary expert.

***Theorem* 2.5.** *Assume that the payoff function h is concave in its first argument and satisfies $h \in [-M, \infty)$. For any n and $0 < \eta < 1/(2M)$, and for all $y_1, \ldots, y_n \in \mathcal{Y}$, the regret of the multilinear forecaster satisfies*

$$H_{i,n} - \widehat{H}_n \le \frac{\ln N}{\eta} + \eta \sum_{t=1}^n h(f_{i,t}, y_t)^2 \quad \text{for each } i = 1, \ldots, N.$$

Proof. For any $i = 1, \dots, N$, note that $h(f_{i,t}, y_t) \geq -M$ and $\eta \leq 1/(2M)$ imply that $\eta h(f_{i,t}, y_t) \geq -1/2$. Hence, we can apply Lemma 2.4 to $\eta h(f_{i,t}, y_t)$ and get

$$\begin{aligned}
\ln \frac{W_n}{W_0} &= -\ln N + \ln \prod_{t=1}^{n} \bigl(1 + \eta h(f_{i,t}, y_t)\bigr) \\
&= -\ln N + \sum_{t=1}^{n} \ln\bigl(1 + \eta h(f_{i,t}, y_t)\bigr) \\
&\geq -\ln N + \sum_{t=1}^{n} \bigl(\eta h(f_{i,t}, y_t) - \eta^2 h(f_{i,t}, y_t)^2\bigr) \\
&= -\ln N + \eta H_{i,n} - \eta^2 \sum_{t=1}^{n} h(f_{i,t}, y_t)^2.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\ln \frac{W_n}{W_0} &= \sum_{t=1}^{n} \ln \frac{W_t}{W_{t-1}} \\
&= \sum_{t=1}^{n} \ln \left(\sum_{i=1}^{N} \widehat{p}_{i,t} \left(1 + \eta h(f_{i,t}, y_t)\right) \right) \\
&\leq \eta \sum_{t=1}^{n} \sum_{i=1}^{N} \widehat{p}_{i,t}\, h(f_{i,t}, y_t) \quad \text{(since } \ln(1+x) \leq x \text{ for } x > -1\text{)} \\
&\leq \eta \widehat{H}_n \quad \text{(since } h(\cdot, y) \text{ is concave)}.
\end{aligned}$$

Combining the upper and lower bounds of $\ln(W_n/W_0)$, and dividing by $\eta > 0$, we get the claimed bound. ■

Let $Q_n^* = h(f_{k,1}, y_1)^2 + \cdots + h(f_{k,n}, y_n)^2$ where k is such that $H_{k,n} = H_n^* = \max_{i=1,\dots,N} H_{i,n}$. If η is chosen using Q_n^*, then Theorem 2.5 directly implies the following.

***Corollary* 2.6.** *Assume the multilinear forecaster is run with*

$$\eta = \min \left\{ \frac{1}{2M}, \sqrt{\frac{\ln N}{Q_n^*}} \right\},$$

where $Q_n^* > 0$ is supposed to be known in advance. Then, under the conditions of Theorem 2.5,

$$H_n^* - \widehat{H} \leq 2\sqrt{Q_n^* \ln N} + 4M \ln N.$$

To appreciate this result, consider a loss game with $h \in [-M, 0]$ and let $L_n^* = -\max_i H_{i,n}$. As $Q_n^* \leq M L_n^*$, the performance guarantee of the multilinear forecaster is at most a factor of $\sqrt{2}$ larger than that of the exponentially weighted average forecaster, whose regret in this case has the leading term $\sqrt{2 L_n^* M \ln N}$ (see Section 2.4). However, in some cases Q_n^* may be significantly smaller than $M L_n^*$, so that the bound of Corollary 2.6 presents a real improvement. In Section 2.8, we show that a more careful analysis of the exponentially

weighted average forecaster yields similar (though noncomparable) second-order regret bounds.

It is still an open problem to obtain regret bounds of order $\sqrt{Q_n^* \ln N}$ without exploiting some prior knowledge on the sequence $y_1, \ldots, y_n$ (see Exercise 2.14). In fact, the analysis of adaptive tuning techniques, such as the doubling trick or the time-varying η, rely on the monotonicity of the quantity whose evolution determines the tuning strategy. On the other hand, the sequence $Q_1^*, Q_2^*, \ldots$ is not necessarily monotone as Q_t^* and Q_{t+1}^* cannot generally be related when the experts achieving the largest cumulative payoffs at rounds t and $t+1$ are different.

2.8 The Exponential Forecaster for Signed Games

A slight modification of our previous analysis is sufficient to show that the exponentially weighted average forecaster also is able to achieve a small regret in signed games. Like the multilinear forecaster of Section 2.7, this new bound is expressed in terms of sums of quadratic terms that are related to the variance of the experts' losses with respect to the distribution induced by the forecaster's weights. Furthermore, the use of a time-varying potential allows us to dispense with the need of any preliminary knowledge of the best cumulative payoff H_n^*.

We start by redefining, for the setup where payoff functions are used, the exponentially weighted average forecaster with time-varying potential introduced in Section 2.3. Given a payoff function h, this forecaster predicts with $\widehat{p}_t = \sum_{i=1}^N f_{i,t} w_{i,t-1} / W_{t-1}$, where $W_{t-1} = \sum_{j=1}^N w_{j,t-1}$, $w_{i,t-1} = e^{\eta_t H_{i,t-1}}$, $H_{i,t-1} = h(f_{i,1}, y_1) + \cdots + h(f_{i,t-1}, y_{t-1})$, and we assume that the sequence $\eta_1, \eta_2, \ldots$ of parameters is positive. Note that the value of η_1 is immaterial because $H_{i,0} = 0$ for all i.

Let $X_1, X_2, \ldots$ be random variables such that $X_t = h(f_{i,t}, y_t)$ with probability $w_{i,t-1}/W_{t-1}$ for all i and t. The next result, whose proof is left as an exercise, bounds the regret of the exponential forecaster for any nonincreasing sequence of potential parameters in terms of the process $X_1, \ldots, X_n$. Note that this lemma does not assume any boundedness condition on the payoff function.

***Lemma* 2.5.** *Let h be a payoff function concave in its first argument. The exponentially weighted average forecaster, run with any nonincreasing sequence $\eta_1, \eta_2, \ldots$ of parameters satisfies, for any $n \geq 1$ and for any sequence $y_1, \ldots, y_n$ of outcomes,*

$$H_n^* - \widehat{H}_n \leq \left(\frac{2}{\eta_{n+1}} - \frac{1}{\eta_1} \right) \ln N + \sum_{t=1}^n \frac{1}{\eta_t} \ln \mathbb{E}\left[e^{\eta_t (X_t - \mathbb{E} X_t)} \right].$$

Let

$$V_t = \sum_{s=1}^t \operatorname{var}(X_s) = \sum_{s=1}^t \mathbb{E}\left[\left(X_s - \mathbb{E}\, X_s \right)^2 \right].$$

Our next result shows that, with an appropriate choice of the sequence η_t, the regret of the exponential forecaster at time n is at most of order $\sqrt{V_n \ln N}$. Note, however, that the bound is not in closed form as V_n depends on the forecaster's weights $w_{i,t}$ for all i and t.

Theorem 2.6. *Let h be a $[-M, M]$-valued payoff function concave in its first argument. Suppose the exponentially weighted average forecaster is run with*

$$\eta_t = \min\left\{\frac{1}{2M}, \sqrt{\frac{2\left(\sqrt{2}-1\right)}{e-2}}\sqrt{\frac{\ln N}{V_{t-1}}}\right\}, \qquad t = 2, 3, \ldots.$$

Then, for any $n \geq 1$ and for any sequence $y_1, \ldots, y_n$ of outcomes,

$$H_n^* - \widehat{H}_n \leq 4\sqrt{V_n \ln N} + 4M \ln N + (e-2)M.$$

Proof. For brevity, write

$$C = \sqrt{\frac{2\left(\sqrt{2}-1\right)}{e-2}}.$$

We start by applying Lemma 2.5 (with, say, $\eta_1 = \eta_2$)

$$\begin{aligned} H_n^* - \widehat{H}_n &\leq \left(\frac{2}{\eta_{n+1}} - \frac{1}{\eta_1}\right)\ln N + \sum_{t=1}^{n}\frac{1}{\eta_t}\ln \mathbb{E}\left[e^{\eta_t(X_t - \mathbb{E}X_t)}\right] \\ &\leq 2\max\left\{2M\ln N,\ \frac{1}{C}\sqrt{V_n \ln N}\right\} + \sum_{t=1}^{n}\frac{1}{\eta_t}\ln \mathbb{E}\left[e^{\eta_t(X_t - \mathbb{E}X_t)}\right]. \end{aligned}$$

Since $\eta_t \leq 1/(2M)$, $\eta_t(X_t - \mathbb{E}\,X_t) \leq 1$ and we may apply the inequality $e^x \leq 1 + x + (e-2)x^2$ for all $x \leq 1$. We thus find that

$$H_n^* - \widehat{H}_n \leq 2\max\left\{2M\ln N,\ \frac{1}{C}\sqrt{V_n \ln N}\right\} + (e-2)\sum_{t=1}^{n}\eta_t \operatorname{var}(X_t).$$

Now denote by T the first time step t when $V_t > M^2$. Using $\eta_t \leq 1/(2M)$ for all t and $V_T \leq 2M^2$, we get

$$\sum_{t=1}^{n}\eta_t \operatorname{var}(X_t) \leq M + \sum_{t=T+1}^{n}\eta_t \operatorname{var}(X_t).$$

We bound the sum using $\eta_t \leq C\sqrt{(\ln N)/V_{t-1}}$ for $t \geq 2$ (note that, for $t > T$, $V_{t-1} \geq V_T > M^2 > 0$). This yields

$$\sum_{t=T+1}^{n}\eta_t \operatorname{var}(X_t) \leq C\sqrt{\ln N}\sum_{t=T+1}^{n}\frac{V_t - V_{t-1}}{\sqrt{V_{t-1}}}.$$

Let $v_t = \operatorname{var}(X_t) = V_t - V_{t-1}$. Since $V_t \leq V_{t-1} + M^2$ and $V_{t-1} \geq M^2$, we have

$$\frac{v_t}{\sqrt{V_{t-1}}} = \frac{\sqrt{V_t} + \sqrt{V_{t-1}}}{\sqrt{V_{t-1}}}\left(\sqrt{V_t} - \sqrt{V_{t-1}}\right) \leq \left(\sqrt{2}+1\right)\left(\sqrt{V_t} - \sqrt{V_{t-1}}\right).$$

Therefore,

$$\sum_{t=T+1}^{n}\eta_t \operatorname{var}(X_t) \leq \frac{C\sqrt{\ln N}}{\sqrt{2}-1}\left(\sqrt{V_n} - \sqrt{V_T}\right) \leq \frac{C}{\sqrt{2}-1}\sqrt{V_n \ln N}.$$

Substituting our choice of C and performing trivial overapproximations concludes the proof. ■

***Remark* 2.2.** The analyses proposed by Theorem 2.5, Corollary 2.6, and Theorem 2.6 show that the multilinear forecaster and the exponentially weighted average forecaster work, with no need of translating payoffs, in both unsigned and signed games. In addition, the regret bounds shown in these results are potentially much better than the invariant bound $M\sqrt{n \ln N}$ obtained via the explicit payoff transformation $h \mapsto (h + M)/(2M)$ from signed to unsigned games (see Section 2.6). However, none of these bounds applies to the case when no preliminary information is available on the sequence of observed payoffs.

The main term of the bound stated in Theorem 2.6 contains V_n. This quantity is smaller than all quantities of the form

$$\sum_{t=1}^{n} \sum_{i=1}^{N} \frac{w_{i,t-1}}{W_{t-1}} \big(h(f_{i,t}, y_t) - \mu_t\big)^2$$

where $\mu_1, \mu_2, \ldots$ is any sequence of real numbers that may be chosen in *hindsight*, as it is not required for the definition of the forecaster. This gives us a whole family of upper bounds, and we may choose for the analysis the most convenient sequence of μ_t.

To provide a concrete example, denote the effective range of the payoffs at time t by $R_t = \max_{i=1,\ldots,N} h(f_{i,t}, y_t) - \min_{j=1,\ldots,N} h(f_{j,t}, y_t)$ and consider the choice $\mu_t = \min_{j=1,\ldots,N} h(f_{j,t}, y_t) + R_t/2$.

***Corollary* 2.7.** *Under the same assumptions as in Theorem 2.6,*

$$H_n^* - \widehat{H}_n \le 2 \sqrt{(\ln N) \sum_{t=1}^{n} R_t^2 + 4M \ln N + (e-2)M}.$$

In a loss game, where h has the range $[-M, 0]$, Corollary 2.7 states that the regret is bounded by a quantity dominated by the term $2M\sqrt{n \ln N}$. A comparison with the bound of Theorem 2.3 shows that we have only lost a factor $\sqrt{2}$ to obtain a much more general result.

2.9 Simulatable Experts

So far, we have viewed experts as unspecified entities generating, at each round t, an advice to which the forecaster has access. A different setup is when the experts themselves are accessible to the forecaster, who can make arbitrary experiments to reveal their future behavior. In this scenario we may define an expert E using a sequence of functions $f_{E,t} : \mathcal{Y}^{t-1} \to \mathcal{D}$, $t = 1, 2, \ldots$ such that, at every time instant t, the expert predicts according to $f_{E,t}(y^{t-1})$. We also assume that the forecaster has access to these functions and therefore can, at any moment, hypothesize future outcomes and compute all experts' future predictions for that specific sequence of outcomes. Thus, the forecaster can "simulate" the experts' future reactions, and we call such experts *simulatable*. For example, a simulatable expert for a prediction problem where $\mathcal{D} = \mathcal{Y}$ is the expert E such

that $f_{E,t}(y^{t-1}) = (y_1 + \cdots + y_{t-1})/(t-1)$. Note that we have assumed that at time t the prediction of a simulatable expert only depends on the sequence y^{t-1} of past observed outcomes. This is not true for the more general type of experts, whose advice might depend on arbitrary sources of information also hidden from the forecaster. More importantly, while the prediction of a general expert, at time t, may depend on the past moves $\widehat{p}_1, \ldots, \widehat{p}_{t-1}$ of the forecaster (just recall the protocol of the game of prediction with expert advice), the values a simulatable expert outputs only depend on the past sequence of outcomes. Because here we are not concerned with computational issues, we allow $f_{E,t}$ to be an arbitrary function of y^{t-1} and assume that the forecaster may always compute such a function.

A special type of simulatable expert is a *static* expert. An expert E is static when its predictions $f_{E,t}$ only depend on the round index t and not on y^{t-1}. In other words, the functions $f_{E,t}$ are all constant valued. Thus, a static expert E is completely described by the sequence $f_{E,1}, f_{E,2}, \ldots$ of its predictions at each round t. This sequence is fixed irrespective of the actual observed outcomes. For this reason we use $f = (f_1, f_2, \ldots)$ to denote an arbitrary static expert. Abusing notation, we use f also to denote simulatable experts.

Simulatable and static experts provide the forecaster with additional power. It is then interesting to consider whether this additional power could be exploited to reduce the forecaster's regret. This is investigated in depth for specific loss functions in Chapters 8 and 9.

2.10 Minimax Regret

In the model of prediction with expert advice, the best regret bound obtained so far, which holds for all $[0, 1]$-valued convex losses, is $\sqrt{(n/2)\ln N}$. This is achieved (for any fixed $n \geq 1$) by the exponentially weighted average forecaster. Is this the best possible uniform bound? Which type of forecaster achieves the best regret bound for each specific loss? To address these questions in a rigorous way we introduce the notion of minimax regret. Fix a loss function ℓ and consider N general experts. Define the *minimax regret at horizon* n by

$$V_n^{(N)} = \sup_{(f_{1,1},\ldots,f_{N,1})\in\mathcal{D}^N} \inf_{\widehat{p}_1\in\mathcal{D}} \sup_{y_1\in\mathcal{Y}} \sup_{(f_{1,2},\ldots,f_{N,2})\in\mathcal{D}^N} \inf_{\widehat{p}_2\in\mathcal{D}} \sup_{y_2\in\mathcal{Y}} \cdots \sup_{(f_{1,n},\ldots,f_{N,n})\in\mathcal{D}^N} \inf_{\widehat{p}_n\in\mathcal{D}} \sup_{y_n\in\mathcal{Y}} \left(\sum_{t=1}^{n} \ell(\widehat{p}_t, y_t) - \min_{i=1,\ldots,N} \sum_{t=1}^{n} \ell(f_{i,t}, y_t) \right).$$

An equivalent, but simpler, definition of minimax regret can be given using static experts. Define a strategy for the forecaster as a prescription for computing, at each round t, the prediction $\widehat{p}_t$ given the past $t-1$ outcomes $y_1, \ldots, y_{t-1}$ and the expert advice $(f_{1,s}, \ldots, f_{N,s})$ for $s = 1, \ldots, t$. Formally, a forecasting strategy P is a sequence $\widehat{p}_1, \widehat{p}_2, \ldots$ of functions

$$\widehat{p}_t : \mathcal{Y}^{t-1} \times \left(\mathcal{D}^N\right)^t \to \mathcal{D}.$$

Now fix any class $\mathcal{F}$ of N static experts and let $\widehat{L}_n(P, \mathcal{F}, y^n)$ be the cumulative loss on the sequence y^n of the forecasting strategy P using the advice of the experts in $\mathcal{F}$. Then the

minimax regret $V_n^{(N)}$ can be equivalently defined as

$$V_n^{(N)} = \inf_P \sup_{\{\mathcal{F}:|\mathcal{F}|=N\}} \sup_{y^n \in \mathcal{Y}^n} \max_{i=1,\dots,N} \left(\widehat{L}_n(P, \mathcal{F}, y^n) - \sum_{t=1}^n \ell(f_{i,t}, y_t) \right),$$

where the infimum is over all forecasting strategies P and the first supremum is over all possible classes of N static experts (see Exercise 2.18).

The minimax regret measures the best possible performance guarantee one can have for a forecasting algorithm that holds for all possible classes of N experts and all outcome sequences of length n. An upper bound on $V_n^{(N)}$ establishes the existence of a forecasting strategy achieving a regret not larger than the upper bound, regardless of what the class of experts and the outcome sequence are. On the other hand, a lower bound on $V_n^{(N)}$ shows that for any forecasting strategy there exists a class of N experts and an outcome sequence such that the regret of the forecaster is at least as large as the lower bound.

In this chapter and in the next, we derive minimax regret upper bounds for several losses, including $V_n^{(N)} \le \ln N$ for the logarithmic loss $\ell(x, y) = -\mathbb{I}_{\{y=1\}} \ln x - \mathbb{I}_{\{y=0\}} \ln(1-x)$, where $x \in [0,1]$ and $y \in \{0,1\}$, and $V_n^{(N)} \le \sqrt{(n/2)\ln N}$ for all $[0,1]$-valued convex losses, both achieved by the exponentially weighted average forecaster. In Chapters 3, 8, and 9 we complement these results by proving, among other related results, that $V_n^{(N)} = \ln N$ for the logarithmic loss provided that $n \ge \log_2 N$ and that the minimax regret for the absolute loss $\ell(x, y) = |x - y|$ is asymptotically $\sqrt{(n/2)\ln N}$, matching the upper bound we derived for convex losses. This entails that the exponentially weighted average forecaster is minimax optimal, in an asymptotic sense, for both the logarithmic and absolute losses.

The notion of minimax regret defined above is based on the performance of any forecaster in the case of the worst possible class of experts. However, often one is interested in the best possible performance a forecaster can achieve compared with the best expert in a *fixed* class. This leads to the definition of minimax regret for a fixed class of (simulatable) experts as follows. Fix some loss function ℓ and let $\mathcal{F}$ be a (not necessarily finite) class of simulatable experts. A forecasting strategy P based on $\mathcal{F}$ is now just a sequence $\widehat{p}_1, \widehat{p}_2, \dots$ of functions $\widehat{p}_t : \mathcal{Y}^{t-1} \to \mathcal{D}$. (Note that $\widehat{p}_t$ implicitly depends on $\mathcal{F}$, which is fixed. Therefore, as the experts in $\mathcal{F}$ are simulatable, $\widehat{p}_t$ need not depend explicitly on the expert advice.) The *minimax regret with respect to* $\mathcal{F}$ at horizon n is then defined by

$$V_n(\mathcal{F}) = \inf_P \sup_{y^n \in \mathcal{Y}^n} \left(\sum_{t=1}^n \ell\big(\widehat{p}_t(y^{t-1}), y_t\big) - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \ell\big(f_t(y^{t-1}), y_t\big) \right).$$

This notion of regret is studied for specific losses in Chapters 8 and 9.

Given a class $\mathcal{F}$ of simulatable experts, one may also define the closely related quantity

$$U_n(\mathcal{F}) = \sup_Q \inf_P \int_{\mathcal{Y}^n} \left(\sum_{t=1}^n \ell\big(\widehat{p}_t(y^{t-1}), y_t\big) - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \ell\big(f_t(y^{t-1}), y_t\big) \right) \mathrm{d}Q(y^n),$$

where the supremum is taken over all probability measures over the set $\mathcal{Y}^n$ of sequences of outcomes of length n. $U_n(\mathcal{F})$ is called the *maximin regret with respect to* $\mathcal{F}$. Of course, to define probability measures over $\mathcal{Y}^n$, the set $\mathcal{Y}$ of outcomes should satisfy certain regularity properties. For simplicity, assume that $\mathcal{Y}$ is a compact subset of $\mathbb{R}^d$. This assumption is satisfied for most examples that appear in this book and can be significantly weakened if necessary. A general minimax theorem, proved in Chapter 7, implies that if the decision

space $\mathcal{D}$ is convex and the loss function ℓ is convex and continuous in its first argument, then

$$V_n(\mathcal{F}) = U_n(\mathcal{F}).$$

This equality follows simply by the fact that the function

$$F(P, Q) = \int_{\mathcal{Y}^n} \left(\sum_{t=1}^{n} \ell\big(\widehat{p}_t(y^{t-1}), y_t\big) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell\big(f_t(y^{t-1}), y_t\big) \right) \mathrm{d}Q(y^n)$$

is convex in its first argument and concave (actually linear) in the second. Here we define a convex combination $\lambda P^{(1)} + (1-\lambda)P^{(2)}$ of two forecasting strategies $P^{(1)} = (\widehat{p}_1^{(1)}, \widehat{p}_2^{(1)}, \ldots)$ and $P^{(2)} = (\widehat{p}_1^{(2)}, \widehat{p}_2^{(2)}, \ldots)$ by a forecaster that predicts, at time t, according to

$$\lambda \widehat{p}_t^{(1)}(y^{t-1}) + (1-\lambda)\widehat{p}_t^{(2)}(y^{t-1}).$$

We leave the details of checking the conditions of Theorem 7.1 to the reader (see Exercise 2.19).

2.11 Discounted Regret

In several applications it is reasonable to assume that losses in the past are less significant than recently suffered losses. Thus, one may consider *discounted regrets* of the form

$$\rho_{i,n} = \sum_{t=1}^{n} \beta_{n-t} r_{i,t},$$

where the discount factors β_t are typically decreasing with t and $r_{i,t} = \ell(\widehat{p}_t, y_t) - \ell(f_{i,t}, y_t)$ is the instantaneous regret with respect to expert i at round t. In particular, we assume that $\beta_0 \geq \beta_1 \geq \beta_2 \geq \cdots$ is a nonincreasing sequence and, without loss of generality, we let $\beta_0 = 1$. Thus, at time $t = n$, the actual regret $r_{i,t}$ has full weight while regrets suffered in the past have smaller weight; the more distant the past, the less its weight.

In this setup the goal of the forecaster is to ensure that, regardless of the sequence of outcomes, the average discounted cumulative regret

$$\max_{i=1,\ldots,N} \frac{\sum_{t=1}^{n} \beta_{n-t} r_{i,t}}{\sum_{t=1}^{n} \beta_{n-t}}$$

is as small as possible. More precisely, one would like to bound the average discounted regret by a function of n that converges to zero as $n \to \infty$. The purpose of this section is to explore for what sequences of discount factors it is possible to achieve this goal. The case when $\beta_t = 1$ for all t corresponds to the case studied in the rest of this chapter. Other natural choices include the exponential discount sequence $\beta_t = a^{-t}$ for some $a > 1$ or sequences of the form $\beta_t = (t+1)^{-a}$ with $a > 0$.

First we observe that if the discount sequence decreases too quickly, then, except for trivial cases, there is no hope to prove any meaningful bound.

***Theorem* 2.7.** *Assume that there is a positive constant c such that for each n there exist outcomes $y_1, y_2 \in \mathcal{Y}$ and two experts $i \neq i'$ such that $i = \operatorname{argmin}_j \ell(f_{j,n}, y_1)$,*

$i' = \operatorname{argmin}_j \ell(f_{j,n}, y_2)$, *and* $\min_{y=y_1,y_2} |\ell(f_{i,n}, y) - \ell(f_{i',n}, y)| \geq c$. *If* $\sum_{t=0}^{\infty} \beta_t < \infty$, *then there exists a constant* C *such that, for any forecasting strategy, there is a sequence of outcomes such that*

$$\max_{i=1,\ldots,N} \frac{\sum_{t=1}^{n} \beta_{n-t} r_{i,t}}{\sum_{t=1}^{n} \beta_{n-t}} \geq C$$

for all n.

Proof. The lower bound follows simply by observing that the weight of the regrets at the last time step $t = n$ is too large: it is comparable with the total weight of the whole past. Formally,

$$\max_{i=1,\ldots,N} \sum_{t=1}^{n} \beta_{n-t} r_{i,t} \geq \max_{i=1,\ldots,N} \beta_0 r_{i,n} = \ell(\widehat{p}_n, y_n) - \min_{i=1,\ldots,N} \ell(f_{i,n}, y_n).$$

Thus,

$$\begin{aligned} &\sup_{y^n \in \mathcal{Y}^n} \max_{i=1,\ldots,N} \frac{\sum_{t=1}^{n} \beta_{n-t} r_{i,t}}{\sum_{t=1}^{n} \beta_{n-t}} \\ &\quad \geq \frac{\sup_{y \in \mathcal{Y}} \bigl(\ell(\widehat{p}_n, y) - \min_{i=1,\ldots,N} \ell(f_{i,n}, y)\bigr)}{\sum_{t=0}^{\infty} \beta_t} \\ &\quad \geq \frac{C}{2 \sum_{t=0}^{\infty} \beta_t}. \quad \blacksquare \end{aligned}$$

Next we contrast the result by showing that whenever the discount factors decrease sufficiently slowly such that $\sum_{t=0}^{\infty} \beta_t = \infty$, it is possible to make the average discounted regret vanish for large n. This follows from an easy application of Theorem 2.1. We may define weighted average strategies on the basis of the discounted regrets simply by replacing $r_{i,t}$ by $\widetilde{r}_{i,t} = \beta_{n-t} r_{i,t}$ in the definition of the weighted average forecaster. Of course, to use such a predictor, one needs to know the time horizon n in advance. We obtain the following.

***Theorem* 2.8.** *Consider a discounted polynomially weighted average forecaster defined, for* $t = 1, \ldots, n$, *by*

$$\widehat{p}_t = \frac{\sum_{i=1}^{N} \phi'\bigl(\sum_{s=1}^{t-1} \widetilde{r}_{i,s}\bigr) f_{i,s}}{\sum_{j=1}^{N} \phi'\bigl(\sum_{s=1}^{t-1} \widetilde{r}_{j,s}\bigr)} = \frac{\sum_{i=1}^{N} \phi'\bigl(\sum_{s=1}^{t-1} \beta_{n-s} r_{i,s}\bigr) f_{i,s}}{\sum_{j=1}^{N} \phi'\bigl(\sum_{s=1}^{t-1} \beta_{n-s} r_{j,s}\bigr)},$$

where $\phi'(x) = (p-1)x^p$, *with* $p = 2 \ln N$. *Then the average discounted regret satisfies*

$$\max_{i=1,\ldots,N} \frac{\sum_{t=1}^{n} \beta_{n-t} r_{i,t}}{\sum_{t=1}^{n} \beta_{n-t}} \leq \sqrt{2e \ln N} \frac{\sqrt{\sum_{t=1}^{n} \beta_{n-t}^2}}{\sum_{t=1}^{n} \beta_{n-t}}.$$

(A similar bound may be proven for the discounted exponentially weighted average forecaster as well.) In particular, if $\sum_{t=0}^{\infty} \beta_t = \infty$, *then*

$$\max_{i=1,\ldots,N} \frac{\sum_{t=1}^{n} \beta_{n-t} r_{i,t}}{\sum_{t=1}^{n} \beta_{n-t}} = o(1).$$

Proof. Clearly the forecaster satisfies the Blackwell condition, and for each t, $|\widetilde{r}_{i,t}| \le \beta_{n-t}$. Then Theorem 2.1 implies, just as in the proof of Corollary 2.1,

$$\max_{i=1,\dots,N} \rho_{i,n} \le \sqrt{2e \ln N} \sqrt{\sum_{t=1}^{n} \beta_{n-t}^2}$$

and the first statement follows. To prove the second, just note that

$$\begin{aligned}
\sqrt{2e \ln N} \frac{\sqrt{\sum_{t=1}^{n} \beta_{n-t}^2}}{\sum_{t=1}^{n} \beta_{n-t}} &= \sqrt{2e \ln N} \frac{\sqrt{\sum_{t=1}^{n} \beta_{t-1}^2}}{\sum_{t=1}^{n} \beta_{t-1}} \\
&\le \sqrt{2e \ln N} \frac{\sqrt{\sum_{t=1}^{n} \beta_{t-1}}}{\sum_{t=1}^{n} \beta_{t-1}} \\
&= \frac{\sqrt{2e \ln N}}{\sqrt{\sum_{t=1}^{n} \beta_{t-1}}} = o(1). \quad \blacksquare
\end{aligned}$$

It is instructive to consider the special case when $\beta_t = (t+1)^{-a}$ for some $0 < a \le 1$. (Recall from Theorem 2.7 that for $a > 1$, no meaningful bound can be derived.) If $a = 1$, Theorem 2.8 implies that

$$\max_{i=1,\dots,N} \frac{\sum_{t=1}^{n} \beta_{n-t} r_{i,t}}{\sum_{t=1}^{n} \beta_{n-t}} \le \frac{C}{\log n}$$

for a constant $C > 0$. This slow rate of convergence to zero is not surprising in view of Theorem 2.7, because the series $\sum_t 1/(t+1)$ is "barely nonsummable." In fact, this bound cannot be improved substantially (see Exercise 2.20). However, for $a < 1$ the convergence is faster. In fact, an easy calculation shows that the upper bound of the theorem implies that

$$\max_{i=1,\dots,N} \frac{\sum_{t=1}^{n} \beta_{n-t} r_{i,t}}{\sum_{t=1}^{n} \beta_{n-t}} = \begin{cases} O\left(1/\log n\right) & \text{if } a = 1 \\ O\left(n^{a-1}\right) & \text{if } 1/2 < a < 1 \\ O\left(\sqrt{(\log n)/n}\right) & \text{if } a = 1/2 \\ O\left(1/\sqrt{n}\right) & \text{if } a < 1/2. \end{cases}$$

Not surprisingly, the slower the discount factor decreases, the faster the average discounted cumulative regret converges to zero. However, it is interesting to observe the "phase transition" occurring at $a = 1/2$: for all $a < 1/2$, the average regret decreases at a rate $n^{-1/2}$, a behavior quantitatively similar to the case when no discounting is taken into account.

2.12 Bibliographic Remarks

Our model of sequential prediction with expert advice finds its roots in the theory of repeated games. Zero-sum repeated games with fixed loss matrix are a classical topic of game theory. In these games, the regret after n plays is defined as the excess loss of the row player with respect to the smallest loss that could be incurred had he known in advance the empirical distribution of the column player actions during the n plays. In his pioneering work, Hannan [141] devises a randomized playing strategy whose per-round

expected regret grows at rate $N\sqrt{3nm/2}$, where N is the number of rows, m is the number of columns in the loss matrix, and n is the number of plays. As shown in Chapter 7, our polynomially and exponentially weighted average forecasters can be used to play zero-sum repeated games achieving the regret $\sqrt{(n/2)\ln N}$. We obtain the same dependence on the number n of plays Compared with Hannan's regret, but we significantly improve the dependence on the dimensions, N and m, of the loss matrix. A different randomized player with a vanishing per-round regret can be also derived from the celebrated Blackwell's approachability theorem [28], generalizing von Neumann's minimax theorem to vector-valued payoffs. This result, which we re-derive in Chapter 7, is based on a mixed strategy equivalent to our polynomially weighted average forecaster with $p = 2$. Exact asymptotical constants for the minimax regret (for a special case) were first shown by Cover [68]. In our terminology, Cover investigates the problem of predicting a sequence of binary outcomes with two static experts, one always predicting 0 and the other always predicting 1. He shows that the minimax regret for the absolute loss in this special case is $(1 + o(1))\sqrt{n/(2\pi)}$.

The problem of sequential prediction, deprived of any probabilistic assumption, is deeply connected with the information-theoretic problem of compressing an individual data sequence. A pioneering research in this field was carried out by Ziv [317, 318] and Lempel and Ziv [197, 317, 319], who solved the problem of compressing an individual data sequence almost as well as the best finite-state automaton. As shown by Feder, Merhav, and Gutman [95], the Lempel–Ziv compressor can be used as a randomized forecaster (for the absolute loss) with a vanishing per-round regret against the class of all finite-state experts, a surprising result considering the rich structure of this class. In addition, Feder, Merhav, and Gutman devise, for the same expert class, a forecaster with a convergence rate better than the rate provable for the Lempel–Ziv forecaster (see also Merhav and Feder [213] for further results along these lines). In Section 9 we continue the investigation of the relationship between prediction and compression showing simple conditions under which prediction with logarithmic loss is minimax equivalent to adaptive data compression. Connections between prediction with expert advice and information content of an individual sequence have been explored by Vovk and Watkins [303], who introduced the notion of predictive complexity of a data sequence, a quantity that, for the logarithmic loss, is related to the Kolmogorov complexity of the sequence. We refer to the book of Li and Vitányi [198] for an excellent introduction to the algorithmic theory of randomness.

Approximately at the same time when Hannan and Blackwell were laying down the foundations of the game-theoretic approach to prediction, Solomonoff had the idea of formalizing the phenomenon of inductive inference in humans as a sequential prediction process. This research eventually led him to the introduction of a *universal prior probability* [273–275], to be used as a prior in bayesian inference. An important "side product" of Solomonoff's universal prior is the notion of algorithmic randomess, which he introduced independently of Kolmogorov. Though we acknowledge the key role played by Solomonoff in the field of sequential prediction theory, especially in connection with Kolmogorov complexity, in this book we look at the problem of forecasting from a different angle. Having said this, we certainly think that exploring the connections between algorithmic randomness and game theory, through the unifying notion of prediction, is a surely exciting research plan.

The field of inductive inference investigates the problem of sequential prediction when experts are functions taken from a large class, possibly including all recursive languages or all partial recursive functions, and the task is that of eventually identifying an expert that

is consistent (or nearly consistent) with an infinite sequence of observations. This learning paradigm, introduced in 1967 by Gold [130], is still actively studied. Unlike the theory described in this book, whose roots are game theoretic, the main ideas and analytical tools used in inductive inference come from recursion theory (see Odifreddi [227]).

In computer science, an area related to prediction with experts is competitive analysis of online algorithms (see the monograph of Borodin and El-Yaniv [36] for a survey). A good example of a paper exploring the use of forecasting algorithms in competitive analysis is the work by Blum and Burch [32].

The paradigm of prediction with expert advice was introduced by Littlestone and Warmuth [203] and Vovk [297], and further developed by Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth [48] and Vovk [298], although some of its main ingredients already appear in the papers of De Santis, Markowski, and Wegman [260], Littlestone [200], and Foster [103]. The use of potential functions in sequential prediction is due to Hart and Mas-Colell [146], who used Blackwell's condition in a game-theoretic context, and to Grove, Littlestone, and Schuurmans [133], who used exactly the same condition for the analysis of certain variants of the Perceptron algorithm (see Chapter 11). Our Theorem 2.1 is inspired by, and partially builds on, Hart and Mas-Colell's analysis of Λ-strategies for playing repeated games [146] and on the analysis of the quasi-additive algorithm of Grove, Littlestone, and Schuurmans [133]. The unified framework for sequential prediction based on potential functions that we describe here was introduced by Cesa-Bianchi and Lugosi [54]. Forecasting based on the exponential potential has been used in game theory as a variant of smooth fictitious play (see, e.g., the book of Fudenberg and Levine [119]). In learning theory, exponentially weighted average forecasters were introduced and analyzed by Littlestone and Warmuth [203] (the weighted majority algorithm) and by Vovk [297] (the aggregating algorithm). The trick of setting the parameter p of the polynomial potential to $2 \ln N$ is due to Gentile [123]. The analysis in Section 2.2 is based on Cesa-Bianchi's work [46]. The idea of doubling trick of Section 2.3 appears in the articles of Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth [48] and Vovk [298], whereas the analysis of Theorem 2.3 is adapted from Auer, Cesa-Bianchi, and Gentile [13]. The data-dependent bounds of Section 2.4 are based on two sources: Theorem 2.4 is from the work of Littlestone and Warmuth [203] and Corollary 2.4 is due to Freund and Schapire [112]. A more sophisticated analysis of the exponentially weighted average forecaster with time-varying η_t is due to Yaroshinski, El-Yaniv, and Seiden [315]. They show a regret bound of the order $(1 + o(1))\sqrt{2L_n^* \ln N}$, where $o(1) \to 0$ for $L_n^* \to \infty$. Hutter and Poland [165] prove a result similar to Exercise 2.10 using follow-the-perturbed-leader, a randomized forecaster that we analyze in Chapter 4.

The multilinear forecaster and the results of Section 2.8 are due to Cesa-Bianchi, Mansour, and Stoltz [57]. A weaker version of Corollary 2.7 was proven by Allenberg-Neeman and Neeman [7].

The gradient-based forecaster of Section 2.5 was introduced by Kivinen and Warmuth [181]. The proof of Corollary 2.5 is due to Cesa-Bianchi [46]. The notion of simulatable experts and worst-case regret for the experts' framework was first investigated by Cesa-Bianchi et al. [48]. Results for more general loss functions are contained in Chung's paper [60]. Fudenberg and Levine [121] consider discounted regrets is a somewhat different model than the one discussed here.

The model of prediction with expert advice is connected to bayesian decision theory. For instance, when the absolute loss is used, the normalized weights of the weighted average

forecaster based on the exponential potential closely approximate the posterior distribution of a simple stochastic generative model for the data sequence (see Exercise 2.7). From this viewpoint, our regret analysis shows an example where the Bayes decisions are robust in a strong sense, because their performance can be bounded not only in expectation with respect to the random draw of the sequence but also for each individual sequence.

2.13 Exercises

2.1 Assume that you have to predict a sequence $Y_1, Y_2, \ldots \in \{0, 1\}$ of i.i.d. random variables with unknown distribution, your decision space is $[0, 1]$, and the loss function is $\ell(\widehat{p}, y) = |\widehat{p} - y|$. How would you proceed? Try to estimate the cumulative loss of your forecaster and compare it to the cumulative loss of the best of the two experts, one of which always predicts 1 and the other always predicts 0. Which are the most "difficult" distributions? How does your (expected) regret compare to that of the weighted average algorithm (which does not "know" that the outcome sequence is i.i.d.)?

2.2 Consider a weighted average forecaster based on a potential function

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

Assume further that the quantity $C(\mathbf{r}_t)$ appearing in the statement of Theorem 2.1 is bounded by a constant for all values of $\mathbf{r}_t$ and that the function $\psi(\phi(u))$ is strictly convex. Show that there exists a nonnegative sequence $\varepsilon_n \to 0$ such that the cumulative regret of the forecaster satisfies, for every n and for every outcome sequence y^n,

$$\frac{1}{n}\left(\max_{i=1,\ldots,N} R_{i,n}\right) \le \varepsilon_n.$$

2.3 Analyze the polynomially weighted average forecaster using Theorem 2.1 but using the potential function $\Phi(\mathbf{u}) = \|\mathbf{u}_+\|_p$ instead of the choice $\Phi_p(\mathbf{u}) = \|\mathbf{u}_+\|_p^2$ used in the proof of Corollary 2.1. Derive a bound of the same form as in Corollary 2.1, perhaps with different constants.

2.4 Let $\mathcal{Y} = \{0, 1\}$, $\mathcal{D} = [0, 1]$, and $\ell(\widehat{p}, y) = |\widehat{p} - y|$. Prove that the cumulative loss $\widehat{L}$ of the exponentially weighted average forecaster is always at least as large as the cumulative loss $\min_{i \le N} L_i$ of the best expert. Show that for other loss functions, such as the square loss $(\widehat{p} - y)^2$, this is not necessarily so. *Hint:* Try to reverse the proof of Theorem 2.2.

2.5 ***(Nonuniform initial weights)*** By definition, the weighted average forecaster uses uniform initial weights $w_{i,0} = 1$ for all $i = 1, \ldots, N$. However, there is nothing special about this choice, and the analysis of the regret for this forecaster can be carried out using any set of nonnegative numbers for the initial weights.

Consider the exponentially weighted average forecaster run with arbitrary initial weights $w_{1,0}, \ldots, w_{N,0} > 0$, defined, for all $t = 1, 2, \ldots$, by

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

Under the same conditions as in the statement of Theorem 2.2, show that for every n and for every outcome sequence y^n,

$$\widehat{L}_n \le \min_{i=1,\ldots,N}\left(L_{i,n} + \frac{1}{\eta} \ln \frac{1}{w_{i,0}}\right) + \frac{\ln W_0}{\eta} + \frac{\eta}{8} n,$$

where $W_0 = w_{1,0} + \cdots + w_{N,0}$.

2.6 ***(Many good experts)*** Sequences of outcomes on which many experts suffer a small loss are intuitively easier to predict. Adapt the proof of Theorem 2.2 to show that the exponentially weighted forecaster satistifies the following property: for every n, for every outcome sequence y^n, and for all $L > 0$,

$$\widehat{L}_n \le L + \frac{1}{\eta} \ln \frac{N}{N_L} + \frac{\eta}{8} n,$$

where N_L is the cardinality of the set $\left\{1 \le i \le N \,:\, L_{i,n} \le L\right\}$.

2.7 ***(Random generation of the outcome sequence)*** Consider the exponentially weighted average forecaster and define the following probabilistic model for the generation of the sequence $y^n \in \{0, 1\}^n$, where we now view each bit y_t as the realization of a Bernoulli random variable Y_t. An expert I is drawn at random from the set of N experts. For each $t = 1, \ldots, n$, first $X_t \in \{0, 1\}$ is drawn so that $X_t = 1$ with probability $f_{I,t}$. Then Y_t is set to X_t with probability β and is set to $1 - X_t$ with probability $1 - \beta$, where $\beta = 1/(1 + e^{-\eta})$. Show that the forecaster weights $w_{i,t}/(w_{1,t} + \cdots + w_{N,t})$ and are equal to the posterior probability $P[I = i \mid Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}]$ that expert i is drawn given that the sequence $y_1, \ldots, y_{t-1}$ has been observed.

2.8 ***(The doubling trick)*** Consider the following forecasting strategy ("doubling trick"): time is divided in periods $(2^m, \ldots, 2^{m+1} - 1)$, where $m = 0, 1, 2, \ldots$. In period $(2^m, \ldots, 2^{m+1} - 1)$ the strategy uses the exponentially weighted average forecaster initialized at time 2^m with parameter $\eta_m = \sqrt{8(\ln N)/2^m}$. Thus, the weighted average forecaster is reset at each time instance that is an integer power of 2 and is restarted with a new value of η. Using Theorem 2.2 prove that, for any sequence $y_1, y_2, \ldots \in \mathcal{Y}$ of outcomes and for any $n \ge 1$, the regret of this forecaster is at most

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \le \frac{\sqrt{2}}{\sqrt{2} - 1} \sqrt{\frac{n}{2} \ln N}.$$

2.9 ***(The doubling trick, continued)*** In Exercise 2.8, quite arbitrarily, we divided time into periods of length 2^m, $m = 1, 2, \ldots$. Investigate what happens if instead the period lengths are of the form $\lfloor a^m \rfloor$ for some other value of $a > 0$. Which choice of a minimizes, asymptotically, the constant in the bound? How much can you gain compared with the bound given in the text?

2.10 Combine Theorem 2.4 with the doubling trick of Exercise 2.8 to construct a forecaster that, without any previous knowledge of L^*, achieves, for all n,

$$\widehat{L}_n - L_n^* \le 2\sqrt{2L_n^* \ln N} + c \ln N$$

whenever the loss function is bounded and convex in its first argument, and where c is a positive constant.

2.11 ***(Another time-varying potential)*** Consider the adaptive exponentially weighted average forecaster that, at time t, uses

$$\eta_t = c\sqrt{\frac{\ln N}{\min_{i=1,\ldots,N} L_{i,t-1}}},$$

where c is a positive constant. Show that whenever ℓ is a $[0, 1]$-valued loss function convex in its first argument, then there exists a choice of c such that

$$\widehat{L}_n - L_n^* \le 2\sqrt{2L_n^* \ln N} + \kappa \ln N,$$

where $\kappa > 0$ is an appropriate constant (Auer, Cesa-Bianchi, and Gentile [13]). *Hint:* Follow the outline of the proof of Theorem 2.3. This exercise is not easy.

2.12 Consider the prediction problem with $\mathcal{Y} = \mathcal{D} = [0, 1]$ with the absolute loss $\ell(\widehat{p}, y) = |\widehat{p} - y|$. Show that in this case the gradient-based exponentially weighted average forecaster coincides

with the exponentially weighted average forecaster. (Note that the derivative of the loss does not exist for $\widehat{p} = y$ and the definition of the gradient-based exponentially weighted average forecaster needs to be adjusted appropriately.)

2.13 Prove Lemma 2.4.

2.14 Use the doubling trick to prove a variant of Corollary 2.6 in which no knowledge about the outcome sequence is assumed to be preliminarily available (however, as in the corollary, we still assume that the payoff function h has range $[-M, M]$ and is concave in its first argument). Express the regret bound in terms of the smallest monotone upper bound on the sequence $Q_1^*, Q_2^*, \ldots$ (see Cesa-Bianchi, Mansour, and Stoltz [57]).

2.15 Prove a variant of Theorem 2.6 in which no knowledge about the range $[-M, M]$ of the payoff function is assumed to be preliminarily available (see Cesa-Bianchi, Mansour, and Stoltz [57]). *Hint:* Replace the term $1/(2M)$ in the definition of η_t with $2^{-(1+k_t)}$, where k is the smallest nonnegative integer such that $\max_{s=1,\ldots,t-1} \max_{i=1,\ldots,N} |h(f_{i,s}, y_s)| \le 2^k$.

2.16 Prove a regret bound for the multilinear forecaster using the update $w_{i,t} = w_{i,t-1}(1 + \eta r_{i,t})$, where $r_{i,t} = h(f_{i,t}, y_t) - h(\widehat{p}_t, y_t)$ is the instantaneous regret. What can you say about the evolution of the total weight $W_t = w_{1,t} + \cdots + w_{N,t}$ of the experts?

2.17 Prove Lemma 2.5. *Hint:* Adapt the proof of Theorem 2.3.

2.18 Show that the two expressions of the minimax regret $V_n^{(N)}$ in Section 2.10 are equivalent.

2.19 Consider a class $\mathcal{F}$ of simulatable experts. Assume that the set $\mathcal{Y}$ of outcomes is a compact subset of $\mathbb{R}^d$, the decision space $\mathcal{D}$ is convex, and the loss function ℓ is convex and continuous in its first argument. Show that $V_n(\mathcal{F}) = U_n(\mathcal{F})$. *Hint:* Check the conditions of Theorem 7.1.

2.20 Consider the discount factors $\beta_t = 1/(t+1)$ and assume that there is a positive constant c such that for each n there exist outcomes $y_1, y_2 \in \mathcal{Y}$ and two experts $i \neq i'$ such that $i = \operatorname{argmin}_j \ell(f_{j,n}, y_1)$, $i' = \operatorname{argmin}_j \ell(f_{j,n}, y_2)$, and $\min_{y=y_1,y_2} |\ell(f_{i,n}, y) - \ell(f_{i',n}, y)| \ge c$. Show that there exists a constant C such that for any forecasting strategy, there is a sequence of outcomes such that

$$\max_{i=1,\ldots,N} \frac{\sum_{t=1}^n \beta_{n-t} r_{i,t}}{\sum_{t=1}^n \beta_{n-t}} \ge \frac{C}{\log n}$$

for all n.