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

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

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

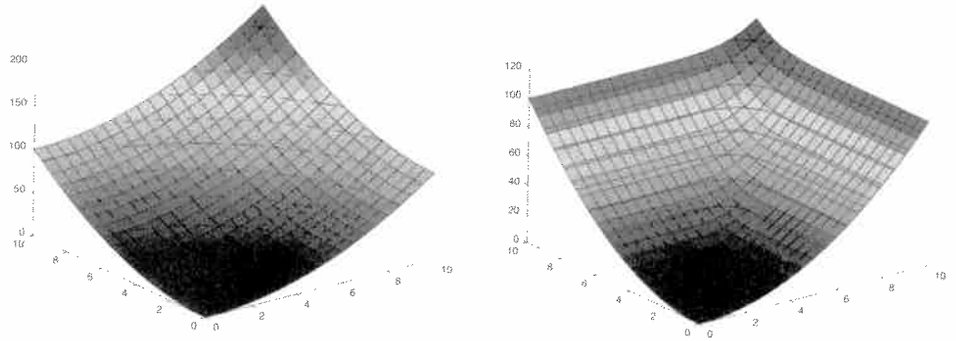

Figure 2.2. Plots of the polynomial potential function $\Phi_p(\mathbf{u})$ for $N = 2$ experts with exponents $p = 2$ and $p = 10$.

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,

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

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

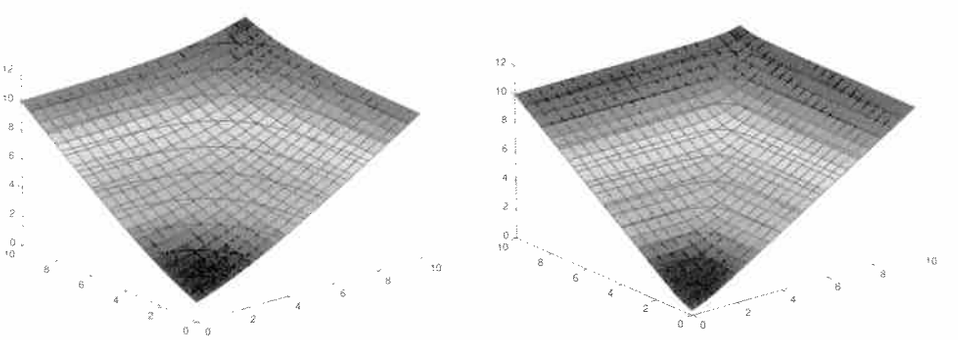

Figure 2.3. Plots of the exponential potential function $\Phi_\eta(\mathbf{u})$ for $N = 2$ experts with $\eta = 0.5$ and $\eta = 2$.

***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 $y_1, \ldots, y_n \in \mathcal{Y}$, the regret of the exponentially weighted average forecaster satisfies*

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

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

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

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

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

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*

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