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* $n = 1, 2, \ldots,$

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

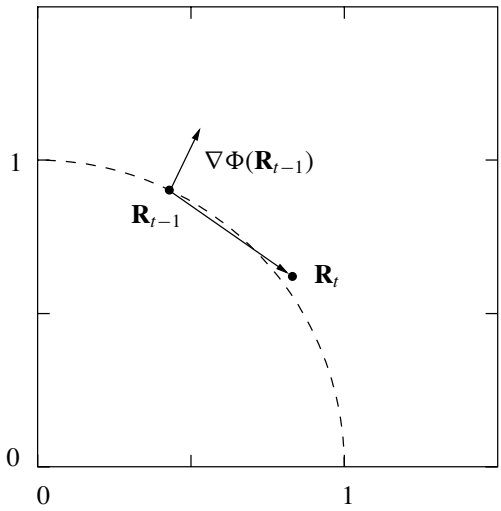

Figure 2.1. An illustration of the Blackwell condition with $N = 2$. The dashed line shows the points in regret space with potential equal to 1. The prediction at time t changed the potential from $\Phi(\mathbf{R}_{t-1}) = 1$ to $\Phi(\mathbf{R}_t) = \Phi(\mathbf{R}_{t-1} + \mathbf{r}_t)$. Though $\Phi(\mathbf{R}_t)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

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

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

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

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.