1

Introduction

1.1 Prediction

Prediction, as we understand it in this book, is concerned with guessing the short-term evolution of certain phenomena. Examples of prediction problems are forecasting tomorrow's temperature at a given location or guessing which asset will achieve the best performance over the next month. Despite their different nature, these tasks look similar at an abstract level: one must predict the next element of an unknown sequence given some knowledge about the past elements and possibly other available information. In this book we develop a formal theory of this general prediction problem. To properly address the diversity of potential applications without sacrificing mathematical rigor, the theory will be able to accommodate different formalizations of the entities involved in a forecasting task, such as the elements forming the sequence, the criterion used to measure the quality of a forecast, the protocol specifying how the predictor receives feedback about the sequence, and any possible side information provided to the predictor.

In the most basic version of the sequential prediction problem, the predictor – or forecaster – observes one after another the elements of a sequence $y_1, y_2, \ldots$ of symbols. At each time $t = 1, 2, \ldots$, before the tth symbol of the sequence is revealed, the forecaster guesses its value y_t on the basis of the previous $t - 1$ observations.

In the classical statistical theory of sequential prediction, the sequence of elements, which we call outcomes, is assumed to be a realization of a stationary stochastic process. Under this hypothesis, statistical properties of the process may be estimated on the basis of the sequence of past observations, and effective prediction rules can be derived from these estimates. In such a setup, the *risk* of a prediction rule may be defined as the expected value of some *loss function* measuring the discrepancy between predicted value and true outcome, and different rules are compared based on the behavior of their risk.

This book looks at prediction from a quite different angle. We abandon the basic assumption that the outcomes are generated by an underlying stochastic process and view the sequence $y_1, y_2, \ldots$ as the product of some unknown and unspecified mechanism (which could be deterministic, stochastic, or even adversarially adaptive to our own behavior). To contrast it with stochastic modeling, this approach has often been referred to as prediction of *individual sequences*.

Without a probabilistic model, the notion of risk cannot be defined, and it is not immediately obvious how the goals of prediction should be set up formally. Indeed, several possibilities exist, many of which are discussed in this book. In our basic model, the performance of the forecaster is measured by the loss accumulated during many rounds of prediction, where loss is scored by some fixed loss function. Since we want to avoid any assumption on the way the sequence to be predicted is generated, there is no obvious baseline against which to measure the forecaster's performance. To provide such a baseline, we introduce a class of *reference forecasters*, also called *experts*. These experts make their prediction available to the forecaster before the next outcome is revealed. The forecaster can then make his own prediction depend on the experts' "advice" in order to keep his cumulative loss close to that of the best reference forecaster in the class.

The difference between the forecaster's accumulated loss and that of an expert is called *regret*, as it measures how much the forecaster regrets, in hindsight, of not having followed the advice of this particular expert. Regret is a basic notion of this book, and a lot of attention is payed to constructing forecasting strategies that guarantee a small regret with respect to *all* experts in the class. As it turns out, the possibility of keeping the regrets small depends largely on the size and structure of the class of experts, and on the loss function. This model of prediction using expert advice is defined formally in Chapter 2 and serves as a basis for a large part of the book.

The abstract notion of an "expert" can be interpreted in different ways, also depending on the specific application that is being considered. In some cases it is possible to view an expert as a black box of unknown computational power, possibly with access to private sources of side information. In other applications, the class of experts is collectively regarded as a statistical model, where each expert in the class represents an optimal forecaster for some given "state of nature." With respect to this last interpretation, the goal of minimizing regret on arbitrary sequences may be thought of as a robustness requirement. Indeed, a small regret guarantees that, even when the model does not describe perfectly the state of nature, the forecaster does almost as well as the best element in the model fitted to the particular sequence of outcomes. In Chapters 2 and 3 we explore the basic possibilities and limitations of forecasters in this framework.

Models of prediction of individual sequences arose in disparate areas motivated by problems as different as playing repeated games, compressing data, or gambling. Because of this diversity, it is not easy to trace back the first appearance of such a study. But it is now recognized that Blackwell, Hannan, Robbins, and the others who, as early as in the 1950s, studied the so-called *sequential compound decision* problem were the pioneering contributors in the field. Indeed, many of the basic ideas appear in these early works, including the use of randomization as a powerful tool of achieving a small regret when it would otherwise be impossible. The model of randomized prediction is introduced in Chapter 4. In Chapter 6 several variants of the basic problem of randomized prediction are considered in which the information available to the forecaster is limited in some way.

Another area in which prediction of individual sequences appeared naturally and found numerous applications is information theory. The influential work of Cover, Davisson, Lempel, Rissanen, Shtarkov, Ziv, and others gave the information-theoretic foundations of sequential prediction, first motivated by applications for data compression and "universal" coding, and later extended to models of sequential gambling and investment. This theory mostly concentrates on a particular loss function, the so-called *logarithmic* or *self-information* loss, as it has a natural interpretation in the framework of *sequential probability assignment*. In this version of the prediction problem, studied in Chapters 9 and 10, at each time instance the forecaster determines a probability distribution over the set of possible outcomes. The total likelihood assigned to the entire sequence of outcomes is then used to score the forecaster. Sequential probability assignment has been studied in different closely related models in statistics, including bayesian frameworks and the problem of *calibration* in various forms. Dawid's "prequential" statistics is also close in spirit to some of the problems discussed here.

In computer science, algorithms that receive their input sequentially are said to operate in an *online* modality. Typical application areas of online algorithms include tasks that involve sequences of decisions, like when one chooses how to serve each incoming request in a stream. The similarity between decision problems and prediction problems, and the fact that online algorithms are typically analyzed on arbitrary sequences of inputs, has resulted in a fruitful exchange of ideas and techniques between the two fields. However, some crucial features of sequential decision problems that are missing in the prediction framework (like the presence of states to model the interaction between the decision maker and the mechanism generating the stream of requests) has so far prevented the derivation of a general theory allowing a unified analysis of both types of problems.

1.2 Learning

Prediction of individual sequences has also been a main topic of research in the theory of machine learning, more concretely in the area of *online learning*. In fact, in the late 1980s–early 1990s the paradigm of prediction with expert advice was first introduced as a model of online learning in the pioneering papers of De Santis, Markowski, and Wegman; Littlestone and Warmuth; and Vovk, and it has been intensively investigated ever since. An interesting extension of the model allows the forecaster to consider other information apart from the past outcomes of the sequence to be predicted. By considering *side information* taking values in a vector space, and experts that are linear functions of the side information vector, one obtains classical models of online pattern recognition. For example, Rosenblatt's Perceptron algorithm, the Widrow-Hoff rule, and ridge regression can be naturally cast in this framework. Chapters 11 and 12 are devoted to the study of such online learning algorithms.

Researchers in machine learning and information theory have also been interested in the computational aspects of prediction. This becomes a particularly important problem when very large classes of reference forecasters are considered, and various tricks need to be invented to make predictors feasible for practical applications. Chapter 5 gathers some of these basic tricks illustrated on a few prototypical examples.

1.3 Games

The online prediction model studied in this book has an intimate connection with game theory. First of all, the model is most naturally defined in terms of a repeated game played between the forecaster and the "environment" generating the outcome sequence, thus offering a convenient way of describing variants of the basic theme. However, the connection is much deeper. For example, in Chapter 7 we show that classical minimax theorems of game theory can be recovered as simple applications of some basic bounds for the performance of sequential prediction algorithms. On the other hand, certain generalized minimax theorems, most notably *Blackwell's approachability theorem* can be used to define forecasters with good performance on individual sequences.

Perhaps surprisingly, the connection goes even deeper. It turns out that if all players in a repeated normal form game play according to certain simple regret-minimizing prediction strategies, then the induced dynamics leads to equilibrium in a certain sense. This interesting line of research has been gaining terrain in game theory, based on the pioneering work of Foster, Vohra, Hart, Mas-Colell, and others. In Chapter 7 we discuss the possibilities and limitations of strategies based on regret minimizing forecasting algorithms that lead to various notions of equilibria.

1.4 A Gentle Start

To introduce the reader to the spirit of the results contained in this book, we now describe in detail a simple example of a forecasting procedure and then analyze its performance on an arbitrary sequence of outcomes.

Consider the problem of predicting an unknown sequence $y_1, y_2, \ldots$ of bits $y_t \in \{0, 1\}$. At each time t the forecaster first makes his guess $\widehat{p}_t \in \{0, 1\}$ for y_t. Then the true bit y_t is revealed and the forecaster finds out whether his prediction was correct. To compute $\widehat{p}_t$ the forecaster listens to the advice of N experts. This advice takes the form of a binary vector $(f_{1,t}, \ldots, f_{N,t})$, where $f_{i,t} \in \{0, 1\}$ is the prediction that expert i makes for the next bit y_t. Our goal is to bound the number of time steps t in which $\widehat{p}_t \neq y_t$, that is, to bound the number of mistakes made by the forecaster.

To start with an even simpler case, assume we are told in advance that, on this particular sequence of outcomes, there is some expert i that makes no mistakes. That is, we know that $f_{i,t} = y_t$ for some i and for all t, but we do not know for which i this holds. Using this information, it is not hard to devise a forecasting strategy that makes at most $\lfloor \log_2 N \rfloor$ mistakes on the sequence. To see this, consider the forecaster that starts by assigning a weight $w_j = 1$ to each expert $j = 1, \ldots, N$. At every time step t, the forecaster predicts with $\widehat{p}_t = 1$ if and only if the number of experts j with $w_j = 1$ and such that $f_{j,t} = 1$ is bigger than those with $w_j = 1$ and such that $f_{j,t} = 0$. After y_t is revealed, if $\widehat{p}_t \neq y_t$, then the forecaster performs the assignment $w_k \leftarrow 0$ on the weight of all experts k such that $f_{k,t} \neq y_t$. In words, this forecaster keeps track of which experts make a mistake and predicts according to the majority of the experts that have been always correct.

The analysis is immediate. Let W_m be the sum of the weights of all experts after the forecaster has made m mistakes. Initially, $m = 0$ and $W_0 = N$. When the forecaster makes his mth mistake, at least half of the experts that have been always correct so far make their first mistake. This implies that $W_m \leq W_{m-1}/2$, since those experts that were incorrect for the first time have their weight zeroed by the forecaster. Since the above inequality holds for all $m \geq 1$, we have $W_m \leq W_0/2^m$. Recalling that expert i never makes a mistake, we know that $w_i = 1$, which implies that $W_m \geq 1$. Using this together with $W_0 = N$, we thus find that $1 \leq N/2^m$. Solving for m (which must be an integer) gives the claimed inequality $m \leq \lfloor \log_2 N \rfloor$.

We now move on to analyze the general case, in which the forecaster does not have any preliminary information on the number of mistakes the experts will make on the sequence. Our goal now is to relate the number of mistakes made by the forecaster to the number of mistakes made by the best expert, irrespective of which sequence is being predicted.

Looking back at the previous forecasting strategy, it is clear that setting the weight of an incorrect expert to zero makes sense only if we are sure that some expert will never make a mistake. Without this guarantee, a safer choice could be performing the assignment $w_k \leftarrow \beta\, w_k$ every time expert k makes a mistake, where $0 < \beta < 1$ is a free parameter. In other words, every time an expert is incorrect, instead of zeroing its weight we shrink it by a constant factor. This is the only modification we make to the old forecaster, and this makes its analysis almost as easy as the previous one. More precisely, the new forecaster compares the total weight of the experts that recommend predicting 1 with those that recommend 0 and predicts according to the weighted majority. As before, at the time the forecaster makes his mth mistake, the overall weight of the incorrect experts must be at least $W_{m-1}/2$. The weight of these experts is then multiplied by β, and the weight of the other experts, which is at most $W_{m-1}/2$, is left unchanged. Hence, we have $W_m \le W_{m-1}/2 + \beta\, W_{m-1}/2$. As this holds for all $m \ge 1$, we get $W_m \le W_0(1+\beta)^m/2^m$. Now let k be the expert that has made the fewest mistakes when the forecaster made his mth mistake. Denote this minimal number of mistakes by m^*. Then the current weight of this expert is $w_k = \beta^{m^*}$, and thus we have $W_m \ge \beta^{m^*}$. This provides the inequality $\beta^{m^*} \le W_0(1+\beta)^m/2^m$. Using this, together with $W_0 = N$, we get the final bound

$$m \le \left\lfloor \frac{\log_2 N + m^* \log_2(1/\beta)}{\log_2 \frac{2}{1+\beta}} \right\rfloor .$$

For any fixed value of β, this inequality establishes a linear dependence between the mistakes made by the forecaster, after any number of predictions, and the mistakes made by the expert that is the best after that same number of predictions. Note that this bound holds irrespective of the choice of the sequence of outcomes.

The fact that m and m^* are linearly related means that, in some sense, the performance of this forecaster gracefully degrades as a function of the "misfit" m^* between the experts and the outcome sequence. The bound also exhibits a mild dependence on the number of experts: the $\log_2 N$ term implies that, apart from computational considerations, doubling the number of experts causes the bound to increase by a small additive term.

Notwithstanding its simplicity, this example contains some of the main themes developed in the book, such as the idea of computing predictions using weights that are functions of the experts' past performance. In the subsequent chapters we develop this and many other ideas in a rigorous and systematic manner with the intent of offering a comprehensive view on the many facets of this fascinating subject.

1.5 A Note to the Reader

The book is addressed to researchers and students of computer science, mathematics, engineering, and economics who are interested in various aspects of prediction and learning. Even though we tried to make the text as self-contained as possible, the reader is assumed to be comfortable with some basic notions of probability, analysis, and linear algebra. To help the reader, we collect in the Appendix some technical tools used in the book. Some of this material is quite standard but may not be well known to all potential readers.

In order to minimize interruptions in the flow of the text, we gathered bibliographical references at the end of each chapter. In these references we intend to trace back the origin of the results described in the text and point to some relevant literature. We apologize for any possible omissions. Some of the material is published here for the first time. These results are not flagged. Each chapter is concluded with a list of exercises whose level of difficulty varies between distant extremes. Some of the exercises can be solved by an easy adaptation of the material described in the main text. These should help the reader in mastering the material. Some others resume difficult research results. In some cases we offer guidance to the solution, but there is no solution manual.

Figure 1.1 describes the dependence structure of the chapters of the book. This should help the reader to focus on specific topics and teachers to organize the material of various possible courses.

1	Introduction
2	Prediction with expert advice
3	Tight bounds for specific losses
4	Randomized prediction
5	Efficient forecasters for large classes of experts
6	Prediction with limited feedback
7	Prediction and playing games
8	Absolute loss
9	Logarithmic loss
10	Sequential investment
11	Linear pattern recognition
12	Linear classification

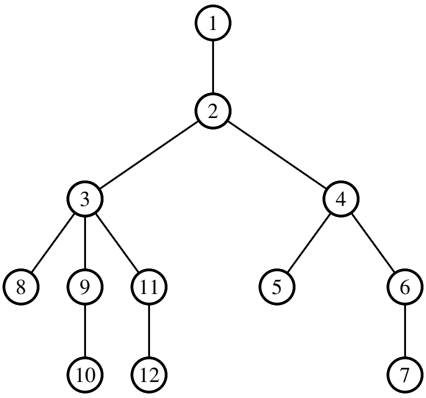

Figure 1.1. The dependence structure of the chapters.

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 $\left\{ f_{E,t} \,:\, E \in \mathcal{E} \right\}$ 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 $\left\{ f_{E,t} \,:\, E \in \mathcal{E} \right\}$ 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 $\left\{ f_{E,t} \in \mathcal{D} \,:\, E \in \mathcal{E} \right\}$; 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

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

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

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