Potential-based Algorithms in On-line Prediction and Game Theory *

Nicolò Cesa-Bianchi (cesa-bianchi@dti.unimi.it)
Department of Information Technologies,
University of Milan
Via Bramante 65,
26013 Crema, Italy

Gábor Lugosi (lugosi@upf.es)
Department of Economics,
Pompeu Fabra University
Ramon Trias Fargas 25-27,
08005 Barcelona, Spain

August 2, 2002

Abstract. In this paper we show that several known algorithms for sequential prediction problems (including Weighted Majority and the quasi-additive family of Grove, Littlestone, and Schuurmans), for playing iterated games (including Freund and Schapire's Hedge and MW, as well as the Λ-strategies of Hart and Mas-Colell), and for boosting (including AdaBoost) are special cases of a general decision strategy based on the notion of potential. By analyzing this strategy we derive known performance bounds, as well as new bounds, as simple corollaries of a single general theorem. Besides offering a new and unified view on a large family of algorithms, we establish a connection between potential-based analysis in learning and their counterparts independently developed in game theory. By exploiting this connection, we show that certain learning problems are instances of more general game-theoretic problems. In particular, we describe a notion of generalized regret and show its applications in learning theory.

Keywords: universal prediction, on-line learning, Blackwell's strategy, Perceptron algorithm, weighted average predictors, internal regret, boosting

1. Introduction

We begin by describing an abstract sequential decision problem and a general strategy to solve it. As we will see in detail in the subsequent sections, several previously known algorithms for more specific decision problems turn out to be special cases of this strategy.

The problem is parametrized by a *decision space* $\mathcal{X}$, by an *outcome space* $\mathcal{Y}$, and by a convex and twice differentiable *potential function* Φ :

* An extended abstract appeared in the *Proceedings of the 14th Annual Conference on Computational Learning Theory and the 5th European Conference on Computational Learning Theory*, Springer, 2001.

$\mathbb{R}^N \to \mathbb{R}^+$. At each step $t = 1, 2, \ldots$, the current state is represented by a point $\boldsymbol{R}_{t-1} \in \mathbb{R}^N$, where $\boldsymbol{R}_0 = \mathbf{0}$. The decision maker observes a vector-valued *drift function* $\boldsymbol{r}_t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^N$ and selects an element $\widehat{y}_t$ from the decision space $\mathcal{X}$. In return, an outcome $y_t \in \mathcal{Y}$ is received, and the new state of the problem is the "drifted point" $\boldsymbol{R}_t = \boldsymbol{R}_{t-1} + \boldsymbol{r}_t(\widehat{y}_t, y_t)$. The goal of the decision maker is to minimize the potential $\Phi(\boldsymbol{R}_t)$ for a given t (which might be known or unknown to the decision maker).

One of the main goals of this paper is to point out that many seemingly unrelated problems fit in the framework of the abstract sequential decision problem described above, and that their analysis may be summarized in some general simple theorems. These problems include on-line prediction problems in the "experts" model, perceptron-like classification algorithms, methods of learning in repeated game playing, etc. We usually think of $\boldsymbol{r}_t$ as the vector of "regrets" the decision maker suffers at time t and $\boldsymbol{R}_t$ is the corresponding "cumulative regret" vector. The decision maker's goal is to keep, in some sense, the cumulative regret vector close to the origin. In the applications described below, the decision maker is free to choose the potential function Φ. To fill the abstract problem described above with meaning, next we describe a propotype example which is detailed in Section 3.

Example. Consider an on-line prediction problem in the experts' framework of Cesa-Bianchi et al. (1997). Here, the decision maker is a predictor whose goal is to forecast a hidden sequence $y_1, y_2, \ldots$ of elements in the outcome space $\mathcal{Y}$. At each time t, the predictor computes its guess $\widehat{y}_t \in \mathcal{X}$ for the next outcome y_t. This guess is based on the advice $f_{1,t}, \ldots, f_{N,t} \in \mathcal{X}$ of N reference predictors, or *experts* from a fixed pool. The guesses of the predictor and the experts are then individually scored using a *loss function* $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The predictor's goal is to keep as small as possible the *cumulative regret* with respect to each expert. This quantity is defined, for expert i, by the sum

$$\sum_{s=1}^{t} \left(\ell(\widehat{y}_s, y_s) - \ell(f_{i,s}, y_s) \right) \ .$$

This can be easily modeled within our abstract decision problem by associating a coordinate to each expert and by defining the components $r_{i,t}$ of the drift function $\boldsymbol{r}_t$ by $r_{i,t}(\widehat{y}_t, y_t) = \ell(\widehat{y}_t, y_t) - \ell(f_{i,t}, y_t)$ for $i = 1, \ldots, N$.

The role of the potential function Φ in the prediction-with-experts framework is to provide a generalized way to measure the size (or

distance from the origin) of the regret $\boldsymbol{R}_t$. This distance information can then be used by the predictor to control the regret. Below, we introduce a class of predictors that use the potential information to keep the drift $\boldsymbol{r}_t$ in the same halfspace where the negative gradient of $\Phi(\boldsymbol{R}_t)$ resides. To guarantee the existence of such predictors we need to constrain our abstract decision problem by making two assumptions which will be naturally satisfied by all of our applications. The notation $\boldsymbol{u} \cdot \boldsymbol{v}$ stands for the inner product of two vectors defined by $\boldsymbol{u} \cdot \boldsymbol{v} = u_1 v_1 + \ldots + u_N v_N$.

1. **Generalized Blackwell's condition.** At each time t, a decision $\widehat{y}_t \in \mathcal{X}$ exists such that

$$\sup_{y_t \in \mathcal{Y}} \nabla\Phi(\boldsymbol{R}_{t-1}) \cdot \boldsymbol{r}_t(\widehat{y}_t, y_t) \leq 0 \ , \tag{1}$$

2. **Additive potential.** The potential Φ can be written as $\Phi(\boldsymbol{u}) = \sum_{i=1}^{N} \phi(u_i)$ for all $\boldsymbol{u} = (u_1, \ldots, u_N) \in I\!R^N$, where $\phi : I\!R \to I\!R^+$ is a nonnegative function of one variable. Typically, ϕ will be monotonically increasing and convex on $I\!R$.

Remark. Strategies satisfying condition (1) tend to keep the point $\boldsymbol{R}_t$ as close as possible to the minimum of the potential by forcing the drift vector to point away from the gradient of the current potential. This gradient descent approach to sequential decision problems is not new. A prominent example of a decision strategy of this type is the one used by Blackwell to prove his celebrated approachability theorem (Blackwell, 1956), generalizing to vector-valued payoffs von Neumann's minimax theorem. The application of Blackwell's strategy to sequential decision problems, and its generalization to arbitrary potentials, is due to a series of papers by Hart and Mas-Colell (2000, 2001), where condition (1) was first introduced (though in a somewhat more restricted context). Condition (1) has been independently introduced by Grove et al. (2001), who used it to define and analyze a new family of algorithms for solving on-line binary classification problems. This family includes, as special cases, the Perceptron (Rosenblatt, 1962) and the zero-threshold Winnow algorithm (Littlestone, 1989). Finally, our abstract decision problem bears some similarities with Schapire's drifting game (Schapire, 2001).

The rest of the paper is organized as follows. In Section 2 a general result is derived for the performance of sequential decision strategies satisfying condition (1), and the special cases of the most important

types of potential functions (i.e., exponential and polynomial) are discussed in detail. In Section 3 we return to the problem of prediction with expert advice, and recover several well-known results by the main result of Section 2. The purpose of Section 4 is to show that many variants of the perceptron algorithm for on-line linear classification (including winnow and the p-norm perceptron) are again special cases of the general problem and that it is a simple matter to re-derive several well-known mistake bounds using the general framework. In Section 5 boosting is revisited with a similar purpose. Section 6 is dedicated to problems of learning in repeated game playing. Here we discuss a family of Hannan consistent methods, and fit an algorithm of Freund and Schapire in the general framework. Finally, we discuss a very general notion of regret, and derive performance bounds for a generalization of a method of adaptive game playing due to Hart and Mas-Colell.

2. General bounds

In this section we describe a general upper bound on the potential of the location reached by the drifting point when the decision maker uses a strategy satisfying condition (1). This result is inspired by, and partially builds on, Hart and Mas-Colell's analysis of their Λ-strategies (Hart and Mas-Colell, 2001) for playing iterated games and the analysis of quasi-additive algorithms for binary classification by Grove, Littlestone and Schuurmans (1997).

Theorem 1. Let Φ be a twice differentiable additive potential function and let $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots \in \mathbb{R}^N$ be such that

$$\nabla\Phi(\boldsymbol{R}_{t-1}) \cdot \boldsymbol{r}_t \le 0$$

for all $t \ge 1$, where $\boldsymbol{R}_t = \boldsymbol{r}_1 + \ldots + \boldsymbol{r}_t$. Let $f : \mathbb{R}^+ \to \mathbb{R}^+$ be an increasing, concave, and twice differentiable auxiliary function such that, for all $t = 1, 2, \ldots$,

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

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

$$f\left(\Phi(\boldsymbol{R}_t)\right) \le f\left(\Phi(\boldsymbol{0})\right) + \frac{1}{2} \sum_{s=1}^{t} C(\boldsymbol{r}_s) \ .$$

Remark. At a first sight it not obvious how to interpret this result. Yet, as we will see below, it may be used to derive useful bounds very easily in a large variety of special cases. At this point we simply point out that in most interesting applications, one finds a *bounded* function C satisfying the assumption. In such cases one obtains, for some constant c, $f\left(\Phi(\boldsymbol{R}_t)\right) \leq f\left(\Phi(\boldsymbol{0})\right) + ct$. Now if $f(\Phi(\boldsymbol{u}))$ has a superlinear growth in some norm of $\boldsymbol{u}$ (e.g., if $f \circ \Phi$ is strictly convex) then this is sufficient to conclude that $\boldsymbol{R}_t/t \to \boldsymbol{0}$ as $t \to \infty$, independently of the outcome sequence. In the examples below we use the theorem to derive nonasymptotic inequalities of this spirit.

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

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

where the inequality follows by (1) and the fact that $f' \geq 0$. Since Φ is additive, straightforward calculation shows that

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

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

In what follows, we will often write $\boldsymbol{r}_t$ instead of $\boldsymbol{r}_t(\widehat{y}_t, y_t)$ when $\widehat{y}_t$ and y_t are taken as arbitrary elements of, respectively, $\mathcal{X}$ and $\mathcal{Y}$. Moreover, we will always use $\boldsymbol{R}_t$ to denote $\boldsymbol{r}_1(\widehat{y}_1, y_1) + \ldots + \boldsymbol{r}_t(\widehat{y}_t, y_t)$.

We now review two simple applications of Theorem 1. The first is for polynomial potential functions. For $p \geq 1$, define the p-norm of a vector $\boldsymbol{u}$ by

$$\|\boldsymbol{u}\|_p = \left(\sum_{i=1}^{N} |u_i|^p \right)^{1/p}$$

and further let a_+ denote $\max\{0, a\}$.

Corollary 1. Assume that a prediction algorithm satisfies (1) with the potential function

$$\Phi(\boldsymbol{u}) = \sum_{i=1}^{N} (u_i)_+^p \ , \tag{2}$$

where $p \geq 2$. Then

$$\Phi(\boldsymbol{R}_t)^{2/p} \leq (p-1) \sum_{s=1}^{t} \|\boldsymbol{r}_s\|_p^2 \quad \text{and} \quad \max_{1 \leq i \leq N} R_{i,t} \leq \sqrt{(p-1) \sum_{s=1}^{t} \|\boldsymbol{r}_s\|_p^2} \ .$$

Proof. Apply Theorem 1 with $f(x) = x^{2/p}$ and $\phi(x) = (x)_+^p$. By straightforward calculation,

$$f'(x) = \frac{2}{p x^{(p-2)/p}} \ .$$

On the other hand, since $\phi''(x) = p(p-1)(x)_+^{p-2}$, by Hölder's inequality,

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

Thus,

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

The conditions of Theorem 1 are then satisfied with the choice $C(\boldsymbol{r}_t) = 2(p-1)\,\|\boldsymbol{r}_t\|_p^2$. Since $\Phi(\boldsymbol{0}) = 0$, Theorem 1 implies the first statement. The second follows from the first simply because

$$\max_{1\le i\le N} R_{i,t} \le \left(\sum_{i=1}^{N} R_{i,t}^p\right)^{1/p} = \Phi(\boldsymbol{R}_t)^{1/p} \ .$$

□

Another simple and important choice for the potential function is the *exponential potential*, treated in the next corollary.

Corollary 2. Assume that a prediction algorithm satisfies (1) with the potential function

$$\Phi(\boldsymbol{u}) = \sum_{i=1}^{N} e^{\eta u_i} \ , \tag{3}$$

where $\eta > 0$ is a parameter. Then

$$\ln \Phi(\boldsymbol{R}_t) \le \ln N + \frac{\eta^2}{2} \sum_{s=1}^{t} \max_{1\le i\le N} r_{i,s}^2$$

and, in particular,

$$\max_{1\le i\le N} R_{i,t} \le \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{s=1}^{t} \max_{1\le i\le N} r_{i,s}^2 \ .$$

Proof. Choosing $f(x) = (1/\eta)\ln x$ and $\phi(x) = e^{\eta x}$, the conditions of Theorem 1 are satisfied with $C(\boldsymbol{r}_t) = \eta \max_{1\le i\le N} r_{i,t}^2$. Using $\Phi(\boldsymbol{0}) = N$ then yields the result. □

Remark. The polynomial potential was considered by Hart and Mas-Colell (2001) and, in the context of binary classification, by Grove et al. (1997), where it was used to define the p-norm Perceptron. The exponential potential is also reminiscent of the smooth fictitious play approach used in game theory (Fudenberg and Levine, 1995) (in fictitious play, the player chooses the pure strategy that is best given the past distribution of the adversary's plays; smoothing this choice amounts to introducing randomization). In learning theory, algorithms based on the exponential potential have been intensively studied and applied to a variety of problems — see, e.g., (Cesa-Bianchi et al., 1997; Freund and Schapire, 1997; Littlestone and Warmuth, 1994; Vovk, 1990; Vovk, 1998).

If $\boldsymbol{r}_t \in [-1,1]^N$ for all t, then the choice $p = 2\ln N$ for the polynomial potential yields the bound

$$\begin{aligned}\max_{1\le i\le N} R_{i,t} &\le \sqrt{(2\ln N - 1)\sum_{s=1}^{t}\left(\sum_{i=1}^{N}|r_{i,s}|^{2\ln N}\right)^{1/\ln N}} \\ &\le \sqrt{(2\ln N-1)N^{1/\ln N}t} = \sqrt{(2\ln N-1)et}\ .\end{aligned}$$

(This choice of p was also suggested by Gentile (2001) in the context of p-norm perceptron algorithms.) A similar bound can be obtained, under the same assumption on the $\boldsymbol{r}_t$'s, by setting $\eta = \sqrt{2\ln N/t}$ in the exponential potential. Note that this tuning of η requires knowledge of the horizon t.

3. Weighted average predictors

In this section, we consider one of the main applications of the potential-based strategy induced by the generalized Blackwell condition, that is, the experts' framework mentioned in Section 1. Recall that, in this framework, the i-th component of the drift vector at time t takes the form of a regret

$$r_{i,t}(\widehat{y}_t, y_t) = \ell(\widehat{y}_t, y_t) - \ell(f_{i,t}, y_t)$$

where $\ell(\widehat{y}_t, y_t)$ is the loss of the predictor and $\ell(f_{i,t}, y_t)$ is the loss of the i-th expert. Denote $\partial\Phi(\boldsymbol{u})/\partial u_i$ by $\nabla_i\Phi(\boldsymbol{u})$ and assume $\nabla_i\Phi(\boldsymbol{u}) \ge 0$ for all $\boldsymbol{u} \in I\!R^N$. A remarkable fact in this application is that, if $\mathcal{X}$ is a convex subset of a vector space and the loss function ℓ is convex in its first component, then a predictor satisfying condition (1) is always obtained by averaging the experts' predictions weighted by the normalized potential gradient. Indeed, note that condition (1) is equivalent to

$$(\forall y \in \mathcal{Y}) \qquad \ell(\widehat{y}_t, y) \le \frac{\sum_{i=1}^{N}\nabla_i\Phi(\boldsymbol{R}_{t-1})\ell(f_{i,t}, y)}{\sum_{j=1}^{N}\nabla_j\Phi(\boldsymbol{R}_{t-1})} \tag{4}$$

Now by convexity of ℓ, we have that (4) is implied by

$$(\forall y \in \mathcal{Y}) \qquad \ell(\widehat{y}_t, y) \le \ell\left(\frac{\sum_{i=1}^{N}\nabla_i\Phi(\boldsymbol{R}_{t-1})f_{i,t}}{\sum_{j=1}^{N}\nabla_j\Phi(\boldsymbol{R}_{t-1})}, y\right)$$

which is clearly satisfied by choosing

$$\widehat{y}_t = \frac{\sum_{i=1}^{N}\nabla_i\Phi(\boldsymbol{R}_{t-1})f_{i,t}}{\sum_{j=1}^{N}\nabla_j\Phi(\boldsymbol{R}_{t-1})}\ .$$

Example. Consider the exponential potential function of Corollary 2. In this case, the weighted average predictor described above simplifies to

$$\begin{aligned}\widehat{y}_t &= \frac{\sum_{i=1}^N \exp\left(\eta \sum_{s=1}^{t-1} \left(\ell(\widehat{y}_s, y_s) - \ell(f_{i,s}, y_s)\right)\right) f_{i,t}}{\sum_{i=1}^N \exp\left(\eta \sum_{s=1}^{t-1} \left(\ell(\widehat{y}_s, y_s) - \ell(f_{i,s}, y_s)\right)\right)} \\ &= \frac{\sum_{i=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \ell(f_{i,s}, y_s)\right) f_{i,t}}{\sum_{i=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \ell(f_{i,s}, y_s)\right)} \ . \end{aligned} \tag{5}$$

This is the well-known Weighted Majority predictor of Littlestone and Warmuth (1994), and Corollary 2 recovers, up to constant factors, previously known performance bounds — see, e.g., (Cesa-Bianchi, 1999). Similarly, Corollary 1 may be used to derive performance bounds for the predictor

$$\widehat{y}_t = \frac{\sum_{i=1}^N \left(\sum_{s=1}^{t-1} \left(\ell(\widehat{y}_s, y_s) - \ell(f_{i,s}, y_s)\right)\right)_+^{p-1} f_{i,t}}{\sum_{i=1}^N \left(\sum_{s=1}^{t-1} \left(\ell(\widehat{y}_s, y_s) - \ell(f_{i,s}, y_s)\right)\right)_+^{p-1}} \tag{6}$$

based on the polynomial potential (2).

These results are summarized as follows.

Corollary 3. Assume that the decision space $\mathcal{X}$ is a convex subset of a vector space and let ℓ be a loss function which is convex in its first component and bounded between 0 and 1. Then the exponential weighted average predictor (5) with parameter $\eta = \sqrt{2 \ln N / t}$ satisfies, for all sequences $y_1, y_2, \ldots$,

$$\sum_{s=1}^t \ell(\widehat{y}_s, y_s) \le \min_{i=1,\ldots,N} \sum_{s=1}^t \ell(f_{i,s}, y_s) + \sqrt{2t \ln N} \ ,$$

and the polynomial weighted average predictor (6) with parameter $p = 2 \ln N$ satisfies, for all sequences $y_1, y_2, \ldots$,

$$\sum_{s=1}^t \ell(\widehat{y}_s, y_s) \le \min_{i=1,\ldots,N} \sum_{s=1}^t \ell(f_{i,s}, y_s) + \sqrt{te(2 \ln N - 1)} \ .$$

The beauty of the Weighted Majority predictor of Corollary 3 is that it only depends on the past performance of the experts, whereas

the predictions made using polynomial (and other general) potentials depend on the past predictions $\widehat{y}_s$, $s < t$ as well.

Remark. In some cases Theorem 1 gives suboptimal bounds. In fact, the arguments of Theorem 1 use Taylor's theorem to bound the increase of the potential function. However, in some situations the value of the potential function is actually nonincreasing. The following property is proven by repeating an argument of Kivinen and Warmuth (1999).

Proposition 1. Consider the weighted majority predictor (5). If the loss function ℓ is such that the function $F(z) = e^{-\eta\ell(z,y)}$ is concave for all $y \in \mathcal{Y}$, then for all $t \geq 1$, $\Phi(\boldsymbol{R}_t) \leq \Phi(\mathbf{0})$ where Φ is the exponential potential function (3). In particular, since $\Phi(\mathbf{0}) = N$, we have $\max_{i=1,\ldots,N} R_{i,t} \leq \ln(N)/\eta$.

Proof. It suffices to show that $\Phi(\boldsymbol{R}_t) \leq \Phi(\boldsymbol{R}_{t-1})$ or, equivalently, that

$$\sum_{i=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell(f_{i,s}, y_s)\right) e^{\eta(\ell(\hat{y}_t, y_t) - \ell(f_{i,t}, y_t))} \leq \sum_{i=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell(f_{i,s}, y_s)\right) ,$$

which, denoting $w_{i,t-1} = \exp\left(-\eta \sum_{s=1}^{t-1} \ell(f_{i,s}, y_s)\right)$, may be written as

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

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

Simple and common examples of loss functions satisfying the concavity assumption of the proposition include the square loss $\ell(z, y) = (z - y)^2$ for $\mathcal{X} = \mathcal{Y} = [0, 1]$ with $\eta = 1/2$ and the logarithmic loss $\ell(z, y) = y \ln(y/z) + (1 - y) \ln((1 - y)/(1 - z))$ with $\eta = 1$. For more information on this type of prediction problems we refer to (Vovk, 2001; Haussler et al., 1998; Kivinen and Warmuth, 1999). Observe that the proof of the proposition does not make explicit use of the generalized Blackwell condition.

We close this section by mentioning that classification algorithms based on time-varying potentials or nonadditive potential functions

have been defined and analyzed in (Auer et al., 2002; Cesa-Bianchi et al., 2002).

4. The quasi-additive algorithm

In this section, we show that the quasi-additive algorithm of Grove, Littlestone and Schuurmans (whose specific instances are the p-norm Perceptron (Gentile, 2001; Grove et al., 1997), the classical Perceptron (Block, 1962; Novikoff, 1962; Rosenblatt, 1962), and the zero-threshold Winnow algorithm (Littlestone, 1989)) is a special case of our general decision strategy. Then, we derive performance bounds as corollaries of Theorem 1.

We recall that the quasi-additive algorithm performs binary classification of *attribute vectors* $\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathbb{R}^N$ by incrementally adjusting a vector $\boldsymbol{w} \in \mathbb{R}^N$ of *weights*. If $\boldsymbol{w}_t$ is the weight vector before observing the t-th attribute vector $\boldsymbol{x}_t$, then the quasi-additive algorithm predicts the unknown label $y_t \in \{-1, 1\}$ of $\boldsymbol{x}_t$ with the thresholded linear function $\widehat{y}_t = \textsc{sgn}(\boldsymbol{x}_t \cdot \boldsymbol{w}_t)$. If the correct label y_t is different from $\widehat{y}_t$, then the weight vector is updated, and the precise way this update occurs distinguishes the various instances of the quasi-additive algorithm.

To fit and analyze the quasi-additive algorithm in our framework, we specialize the abstract decision problem of Section 1 as follows. The decision space $\mathcal{X}$ and the outcome space $\mathcal{Y}$ are both set equal to $\{-1, 1\}$. The drift vector at time t is the function $\boldsymbol{r}_t(\widehat{y}_t, y_t) = \mathbf{I}_{\{y_t \neq \widehat{y}_t\}} y_t \boldsymbol{x}_t$ where $\mathbf{I}_{\{E\}}$ is the indicator function of event E. Instances of the quasi-additive algorithm are parametrized by a potential function Φ and use the gradient of the current potential as weight vector, that is, $\boldsymbol{w}_t = \nabla\Phi(\boldsymbol{R}_{t-1})$. Hence, the weight update is defined by

$$\boldsymbol{w}_{t+1} = \nabla\Phi\left((\nabla\Phi)^{-1}(\boldsymbol{w}_t) + \boldsymbol{r}_t\right)$$

where $(\nabla\Phi)^{-1}$ is the functional inverse of $\nabla\Phi$ (as we will show in Section 4.3, this inverse always exists for the potentials considered here). We now check that condition (1) is satisfied. If $\widehat{y}_t = y_t$, then $\boldsymbol{r}_t(\widehat{y}_t, y_t) = \mathbf{0}$ and the condition is satisfied. Otherwise, $\boldsymbol{r}_t \cdot \nabla\Phi(\boldsymbol{R}_{t-1}) = \mathbf{I}_{\{y_t \neq \widehat{y}_t\}} y_t \boldsymbol{x}_t \cdot \boldsymbol{w}_t \leq 0$, and the condition is satisfied in this case as well.

In the rest of this section, we denote by $M_t = \sum_{s=1}^{t} \mathbf{I}_{\{y_t \neq \widehat{y}_t\}}$ the total number of mistakes made by the specific quasi-additive algorithm being considered.

4.1. The p-norm Perceptron

As defined in (Grove et al., 1997), the p-norm Perceptron uses the potential based on $\phi(u) = |u|^p$, which is just a slight modification of our polynomial potential (2). We now derive a generalization of the Perceptron convergence theorem (Block, 1962; Novikoff, 1962). A version somewhat stronger than ours was proven by Gentile (2001).

For an arbitrary sequence $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_t, y_t)$ of labeled attribute vectors, let $D_t = \sum_{s=1}^{t} \max\{0, \gamma - y_s \, \boldsymbol{x}_s \cdot \boldsymbol{v}_0\}$ be the total *deviation* (Freund and Schapire, 1999b; Gentile, 2001; Gentile and Warmuth, 1999) of $\boldsymbol{v}_0 \in \mathbb{R}^N$ with respect to a given margin $\gamma > 0$. Each term in the sum defining D_t tells whether, and by how much, the linear threshold classifier based on weight vector $\boldsymbol{v}_0$ missed to classify, to within a certain margin, the corresponding example. Thus D_t measures a notion of loss, called *hinge loss* in (Gentile and Warmuth, 1999), different from the number of misclassifications, associated to the weight vector $\boldsymbol{v}_0$.

Corollary 4. Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots \in \mathbb{R}^N \times \{-1, 1\}$ be any sequence of labeled attribute vectors. Then the number M_t of mistakes made by the p-norm Perceptron on a prefix of arbitrary length t of this sequence such that $\|\boldsymbol{x}_s\|_p \le X_p$ for some X_p and for all $s \le t$ is at most

$$\begin{aligned} M_t &\le \frac{D_t}{\gamma} + \frac{p-1}{2}\left(\frac{X_p}{\gamma}\right)^2 + \sqrt{\frac{(p-1)^2 X_p^4 + 4(p-1)\gamma D_t X_p^2}{4\gamma^4}} \\ &\le \frac{D_t}{\gamma} + (p-1)\left(\frac{X_p}{\gamma}\right)^2 + \sqrt{(p-1)\left(\frac{X_p}{\gamma}\right)^2 \frac{D_t}{\gamma}} \end{aligned}$$

where D_t is the hinge loss of $\boldsymbol{v}_0$ with respect to margin γ for any $\boldsymbol{v}_0$ of unit q-norm (q being the dual norm of p) and any $\gamma > 0$.

Proof. Adapting the proof of Corollary 1 to the potential based on $\phi(u) = |u|^p$, and using the bound on $\|\boldsymbol{x}_t\|_p$, we find that $\|\boldsymbol{R}_t\|_p^2 \le (p-1) X_p^2 M_t$. On the other hand, let $\boldsymbol{v}_0 \in \mathbb{R}^N$ be any vector such that $\|\boldsymbol{v}_0\|_q = 1$. Then

$$\begin{aligned} \|\boldsymbol{R}_t\|_p &\ge \boldsymbol{R}_t \cdot \boldsymbol{v}_0 \quad \text{(by Hölder's inequality)} \\ &= \boldsymbol{R}_{t-1} \cdot \boldsymbol{v}_0 + \mathbf{I}_{\{y_t \neq \widehat{y}_t\}} y_t \boldsymbol{x}_t \cdot \boldsymbol{v}_0 \\ &\ge \boldsymbol{R}_{t-1} \cdot \boldsymbol{v}_0 + \mathbf{I}_{\{y_t \neq \widehat{y}_t\}} (\gamma - d_t) \\ &= \cdots \quad \ge \gamma M_t - D_t \ . \end{aligned}$$

Piecing together the two inequalities, and solving the resulting inequality for M_t, yields the desired result. □

4.2. Zero-threshold Winnow

The zero-threshold Winnow algorithm is based on the exponential potential (3). As we did for the p-norm Perceptron, we derive as a corollary a robust version of the bound shown by Grove et al. (1997). Let D_t be the same as in Corollary 4.

Corollary 5. Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots \in \mathbb{R}^N \times \{-1, 1\}$ be any sequence of labeled attribute vectors. On a prefix of arbitrary length t of this sequence such that

$$\begin{aligned} \|\boldsymbol{x}_s\|_\infty \le X_\infty \quad & \text{for some } X_\infty \text{ and for all } s \le t, \\ L \ge D_t/\gamma \quad & \text{for some probability vector } \boldsymbol{v}_0 \text{ and for some } L, \gamma > 0, \end{aligned}$$

the number M_t of mistakes made by zero-threshold Winnow tuned with

$$\eta = \begin{cases} \gamma/(X_\infty^2) & \text{if } L < 2(X_\infty/\gamma)^2 \ln N \\ \sqrt{\dfrac{2 \ln N}{X_\infty^2 L}} & \text{otherwise} \end{cases}$$

is at most $6\,(X_\infty/\gamma)^2 \ln N$ if $L < 2(X_\infty/\gamma)^2 \ln N$, and at most

$$\frac{D_t}{\gamma} + \sqrt{2L \left(\frac{X_\infty}{\gamma}\right)^2 \ln N} + 2 \left(\frac{X_\infty}{\gamma}\right)^2 \ln N \ .$$

otherwise.

Proof. Corollary 2 implies that $\ln \Phi(\boldsymbol{R}_t) \le \ln N + (\eta^2/2) X_\infty^2 M_t$. To obtain a lower bound on $\ln \Phi(\boldsymbol{R}_t)$, consider any vector $\boldsymbol{v}_0$ of convex coefficients. Then we use the well-known "log-sum inequality" — see (Cover and Thomas, 1991) — which implies that, for any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^N$ of nonnegative numbers with $\sum_{i=1}^N v_i = 1$,

$$\ln \sum_{i=1}^N u_i \ge \sum_{i=1}^N v_i \ln u_i + H(\boldsymbol{v}) \ ,$$

where $H(\boldsymbol{v}) = -\sum_{i=1}^N v_i \ln v_i$ is the *entropy* of $\boldsymbol{v}$. Therefore, for any vector $\boldsymbol{v}_0$ of convex coefficients such that $y_s \boldsymbol{v}_0 \cdot \boldsymbol{x}_s \ge \gamma$ for all $s = 1, \ldots, t$,

$$\ln \Phi(\boldsymbol{R}_t) = \ln \sum_{i=1}^N e^{\eta R_{i,t}} \ge \eta \boldsymbol{R}_t \cdot \boldsymbol{v}_0 + H(\boldsymbol{v}_0) \ge \eta\,(\gamma M_t - D_t) + H(\boldsymbol{v}_0)$$

where in the last step we proceeded just like in the proof of Corollary 4. Putting the upper and lower bounds for $\ln \Phi(\boldsymbol{R}_t)$ together we obtain

$$\eta \left(\gamma M_t - D_t\right) + H(\boldsymbol{v}_0) \leq \ln N + (\eta^2/2) X_\infty^2 M_t$$

which, dropping the positive term $H(\boldsymbol{v}_0)$, implies

$$M_t \left(1 - \frac{\eta}{2} \frac{X_\infty^2}{\gamma}\right) \leq \frac{D_t}{\gamma} + \frac{\ln N}{\eta\gamma} \ . \tag{7}$$

We show the proof only for the case $L \geq 2(X_\infty/\gamma)^2 \ln N$. Letting $\beta = (\eta X_\infty^2)/(2\gamma)$, and verifying that $\beta < 1$, we may rearrange (7) as follows

$$\begin{aligned}
M_t &\leq \frac{1}{1-\beta} \left(\frac{D_t}{\gamma} + \frac{1}{2\beta} \left(\frac{X_\infty}{\gamma}\right)^2 \ln N\right) \\
&\leq \frac{D_t}{\gamma} + \frac{1}{1-\beta} \left(\beta \frac{D_t}{\gamma} + \frac{1}{\beta} \frac{A}{2}\right) \quad \text{where we set } A = (X_\infty/\gamma)^2 \ln N \\
&\leq \frac{D_t}{\gamma} + \frac{1}{1-\beta} \left(\beta L + \frac{1}{\beta} \frac{A}{2}\right) \quad \text{since } L \geq D_t/\gamma \text{ by hypothesis} \\
&\leq \frac{D_t}{\gamma} + \frac{\sqrt{2AL}}{1 - \sqrt{A/(2L)}} \quad \text{since } \beta = \sqrt{A/(2L)} \text{ by our choice of } \eta \\
&\leq \frac{D_t}{\gamma} + \sqrt{2AL} + 2A
\end{aligned}$$

whenever $L \geq 2A$, which holds by hypothesis.

4.3. Potentials and Bregman divergences

An alternative analysis for the quasi-additive algorithm, which was proposed in (Warmuth and Jagota, 1997; Kivinen and Warmuth, 2001), leads to mistake bounds essentially equivalent to ours. This alternative analysis is based on the notion on Bregman divergences (Bregman, 1967). In this section we re-derive a (very general) form of these mistake bounds starting from our notion of potential. Concrete bounds for particular choices of the potential functions can be also derived just as we did for our potential-based analysis.

Fix a differentiable strictly convex nonegative additive potential Φ on $\mathbb{R}^N$. For any pair of vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^N$, the Bregman divergence from $\boldsymbol{u}$ to $\boldsymbol{v}$ is defined by

$$\Delta_\Phi \left(\boldsymbol{u}, \boldsymbol{v}\right) = \Phi(\boldsymbol{u}) - \Phi(\boldsymbol{v}) - (\boldsymbol{u} - \boldsymbol{v}) \nabla \Phi(\boldsymbol{v}) \ .$$

Hence, $\Delta_\Phi(\boldsymbol{u}, \boldsymbol{v})$ is the error of the first-order Taylor approximation of the convex potential $\Phi(\boldsymbol{u})$ around $\boldsymbol{v}$.

The only property of Bregman divergences we use is the following trivial fact.

Fact 1. For all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in I\!R^N$,

$$\Delta_\Phi(\boldsymbol{u}, \boldsymbol{v}) + \Delta_\Phi(\boldsymbol{v}, \boldsymbol{w}) = \Delta_\Phi(\boldsymbol{u}, \boldsymbol{w}) + (\boldsymbol{u} - \boldsymbol{v})\,(\nabla\Phi(\boldsymbol{w}) - \nabla\Phi(\boldsymbol{v})) \ .$$

The Bregman divergence $\Delta_\Phi(\cdot, \cdot)$, defined directly on the additive potential $\Phi(\boldsymbol{u}) = \sum_i \phi(u_i)$ on which the quasi-additive algorithm is defined, turns out to be the wrong quantity for deriving mistake bounds. Instead, we will use the related potential $\widetilde{\Phi}(\boldsymbol{u}) = \sum_{i=1}^N \widetilde{\phi}(u_i)$, where $\widetilde{\phi} : I\!R \to I\!R$ has the form

$$\widetilde{\phi}(u) = \int_{-\infty}^{u} (\phi')^{-1}(s)\, ds \ .$$

Note that the inverse $(\phi')^{-1}$ exists since we assumed $\phi'' > 0$.

The additive potential $\widetilde{\Phi}$ has the following key property.

Fact 2. $\nabla\widetilde{\Phi} = (\nabla\Phi)^{-1}$.

Proof. Pick any $\boldsymbol{R} \in I\!R^N$ and let $\boldsymbol{w} = \nabla\Phi(\boldsymbol{R})$. Then

$$\nabla\widetilde{\Phi}(\boldsymbol{w})_i = (\phi')^{-1}(w_i) = (\phi')^{-1}\left(\phi'(R_i)\right) = R_i$$

and thus $\nabla\widetilde{\Phi}(\boldsymbol{w}) = \boldsymbol{R}$. □

We are now ready to derive a bound on the cumulative hinge loss, or total deviation, of the quasi-additive algorithm. As the hinge loss upper bounds the number of mistakes, this will also serve as a mistake bound for the same algorithm. Our derivation is taken from (Warmuth and Jagota, 1997; Kivinen and Warmuth, 2001); we just change their notation to the one used here.

Theorem 2. Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots \in I\!R^N \times \{-1, 1\}$ be any sequence of labeled attribute vectors. If the quasi-additive algorithm is run with potential Φ then its cumulative hinge loss $\sum_{s=1}^t (\gamma - y_s\, \boldsymbol{w}_s \cdot \boldsymbol{x}_s)_+$ on a prefix of arbitrary length t of this sequence is at most

$$D_t + \Delta_{\widetilde{\Phi}}(\boldsymbol{v}_0, \boldsymbol{0}) + \sum_{s=1}^{t} \Delta_{\widetilde{\Phi}}(\boldsymbol{w}_s, \boldsymbol{w}_{s+1})$$

where D_t is the cumulative hinge loss of $\boldsymbol{v}_0$ at margin γ, for any $\boldsymbol{v}_0 \in I\!R^N$ and any $\gamma > 0$.

Proof. For all $y \in \{-1,1\}$, all $\gamma > 0$, and all $\boldsymbol{u}, \boldsymbol{x} \in \mathbb{R}^N$ let $\ell(\boldsymbol{u}\cdot\boldsymbol{x}, y_s)$ be the hinge loss $(\gamma - y\,\boldsymbol{u}\cdot\boldsymbol{x})_+$. Let also

$$\nabla\ell(\boldsymbol{u}\cdot\boldsymbol{x}, y) = \left(\frac{\partial\ell(\boldsymbol{u}\cdot\boldsymbol{x}, y)}{\partial u_1}, \ldots, \frac{\partial\ell(\boldsymbol{u}\cdot\boldsymbol{x}, y)}{\partial u_N}\right)$$

and note that when $\gamma > y\,\boldsymbol{u}\cdot\boldsymbol{x}$, then $-\nabla\ell(\boldsymbol{u}\cdot\boldsymbol{x}, y) = y\,\boldsymbol{x}$ is the drift vector $\boldsymbol{r}(\boldsymbol{u}\cdot\boldsymbol{x}, y)$. Now fix $\gamma > 0$ and $\boldsymbol{v}_0 \in \mathbb{R}^N$. Then for every $s = 1, \ldots, t$ such that $\ell(\boldsymbol{w}_s\cdot\boldsymbol{x}_s, y_s)$ is positive, we have

$$\begin{aligned}
&\ell(\boldsymbol{w}_s\cdot\boldsymbol{x}_s, y_s) - \ell(\boldsymbol{v}_0\cdot\boldsymbol{x}_s, y_s)\\
&\quad\le\ -(\boldsymbol{v}_0 - \boldsymbol{w}_s)\cdot\nabla\ell(\boldsymbol{w}_s\cdot\boldsymbol{x}_s, y_s)\\
&\qquad\text{(by Taylor's theorem and using convexity of the hinge loss)}\\
&\quad=\ (\boldsymbol{v}_0 - \boldsymbol{w}_s)\cdot(\boldsymbol{R}_s - \boldsymbol{R}_{s-1})\\
&\qquad\text{(as } -\nabla\ell(\boldsymbol{w}_s\cdot\boldsymbol{x}_s, y_s) = \boldsymbol{r}_s)\\
&\quad=\ (\boldsymbol{v}_0 - \boldsymbol{w}_s)\cdot\left(\nabla\widetilde{\Phi}(\boldsymbol{w}_{s+1}) - \nabla\widetilde{\Phi}(\boldsymbol{w}_s)\right)\\
&\qquad\text{(by Fact 2 recalling that } \boldsymbol{w}_s = \nabla\Phi(\boldsymbol{R}_{s-1}) \text{ for each } s \ge 1).\\
&\quad=\ \left(\Delta_{\widetilde{\Phi}}(\boldsymbol{v}_0, \boldsymbol{w}_s) - \Delta_{\widetilde{\Phi}}(\boldsymbol{v}_0, \boldsymbol{w}_{s+1}) + \Delta_{\widetilde{\Phi}}(\boldsymbol{w}_s, \boldsymbol{w}_{s+1})\right)\\
&\qquad\text{(by Fact 1)}
\end{aligned}$$

By summing over s, using the positivity of Bregman divergences, and recalling that $\boldsymbol{w}_1 = \boldsymbol{0}$ we get the desired result. □

Remark. Unlike our potential-based analysis, the analysis based on Bregman divergences shown above can be naturally applied not only to the binary classification problem but also to regression problems with any arbitrary convex loss. However, while the divergence-based analysis appears to be limited to predictors based on weighted averages, the potential-based analysis can handle more general predictors like those introduced in Section 6. It is therefore an interesting open problem to understand whether these more complex predictors could be analyzed using Bregman divergences (or, alternatively, to show that the problems of Section 6 can be solved by weighted average predictors).

5. Boosting

Boosting algorithms for binary classification problems receive in input a labeled sample $(v_1, \ell_1), \ldots, (v_N, \ell_N) \in \mathcal{V} \times \{-1, 1\}$, where $\mathcal{V}$ is a generic instance space, and return linear-threshold classifiers of the form $\textsc{sgn}\left(\sum_{s=1}^t \alpha_s h_s\right)$, where $\alpha_s \in \mathbb{R}$ and the functions $h_s : \mathcal{V} \to$

$[-1,1]$ belong to a fixed *hypothesis space* $\mathcal{H}$. In the *boosting by resampling* schema, the classifier is built incrementally: at each step t, the booster weighs the sample and calls an oracle (the so-called *weak learner*) that returns some $h_t \in \mathcal{H}$. Then the booster chooses α_t based on the performance of h_t on the weighted sample and adds $\alpha_t h_t$ to the linear-threshold classifier. Boosting by resampling can be easily fitted in our framework by letting, at each round t, α_t be the decision maker's choice ($\mathcal{X} = I\!R$) and h_t be the outcome ($\mathcal{Y} = \mathcal{H}$). The drift function $\boldsymbol{r}_t$ is defined by $r_{i,t}(\alpha_t, h_t) = -\alpha_t \ell_i h_t(v_i)$ for each $i = 1, \ldots, N$, and condition (1) takes the form

$$\nabla\Phi(\boldsymbol{R}_{t-1}) \cdot \boldsymbol{r}_t = -\alpha_t \sum_{i=1}^{N} \ell_i h_t(v_i) \nabla_i \Phi(\boldsymbol{R}_{t-1}) \leq 0 \ .$$

Define $\overline{m}(h_t) = \sum_{i=1}^{N} \ell_i h_t(v_i) \nabla_i \Phi(\boldsymbol{R}_{t-1})$ as the *weighted margin* of h_t. We see that (1) corresponds to $\alpha_t \overline{m}(h_t) \geq 0$. Freund and Schapire's AdaBoost (1997) is a special case of this schema: the potential is exponential and α_t is chosen in a way such that (1) is satisfied. We recover the known bound on the training accuracy of the classifier output by AdaBoost as a special case of our main result.

Corollary 6. For every training set $(v_1, \ell_1), \ldots, (v_N, \ell_N) \in \mathcal{V} \times \{-1, 1\}$, and for every sequence $h_1, h_2, \ldots$ of functions $h_t : \mathcal{V} \to [-1, 1]$, if Φ is the exponential potential (3) with $\eta = 1$, then the training error of the classifier $f = \textsc{sgn}\left(\sum_{s=1}^{t} \widetilde{\alpha}_s h_s\right)$ satisfies

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{I}_{\{f(v_i) \neq \ell_i\}} \leq \exp\left(-\frac{1}{2} \sum_{s=1}^{t} \widetilde{\alpha}_s^2\right) \ ,$$

where $\widetilde{\alpha}_s = \overline{m}(h_s) / \left(\sum_{i=1}^{N} \exp(R_{i,t-1})\right)$ is the normalized weighted margin.

Proof. The result does not follow directly from Corollary 2. We need to slightly modify the proof of Theorem 1 when the negative term $f'(\Phi(\boldsymbol{R}_{t-1}))\nabla\Phi(\boldsymbol{R}_{t-1}) \cdot \boldsymbol{r}_t$ was dropped. Here, this term takes the form

$$-\widetilde{\alpha}_t \sum_{i=1}^{N} \frac{\ell_i h_t(v_i) \exp(R_{i,t-1})}{\sum_{j=1}^{N} \exp(R_{j,t-1})} = -\frac{\widetilde{\alpha}_t \overline{m}(h_t)}{\sum_{j=1}^{N} \exp(R_{j,t-1})} = -\widetilde{\alpha}_t^2 \ .$$

We keep this term around and proceed by noting that

$$C(\boldsymbol{r}_t) = \max_{1 \leq i \leq N} r_{i,t}^2 = \max_{1 \leq i \leq N} \left(\widetilde{\alpha}_t \ell_i h_t(v_i)\right)^2 \leq \widetilde{\alpha}_t^2 \ .$$

Continuing as in the proof of Corollary 2, we obtain

$$\ln \Phi(\boldsymbol{R}_t) \leq \ln N + \sum_{s=1}^{t} \left(-\widetilde{\alpha}_s^2 + \frac{\widetilde{\alpha}_s^2}{2} \right) = \ln N - \frac{1}{2} \sum_{s=1}^{t} \widetilde{\alpha}_s^2 \ .$$

By rearranging and exponentiating we get

$$\frac{\Phi(\boldsymbol{R}_t)}{N} \leq \exp\left(-\frac{1}{2} \sum_{s=1}^{t} \widetilde{\alpha}_s^2 \right) \ .$$

As, for the exponential potential,

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{I}_{\{f(v_i) \neq \ell_i\}} \leq \frac{1}{N} \sum_{i=1}^{N} \exp\left(-\ell_i \sum_{s=1}^{t} \widetilde{\alpha}_s h_s(v_i) \right) = \frac{\Phi(\boldsymbol{R}_t)}{N}$$

we get the desired result. □

6. Potential-based algorithms in game theory

Our abstract decision problem can be applied to the problem of playing repeated games. Consider first a game between a player and an adversary. At each round of the game, the player chooses an action (or pure strategy) $i \in \{1, \ldots, m\}$ and, independently, the adversary chooses an action $y \in \mathcal{Y}$. The player's loss $L(i, y)$ is the value of a loss function $L : \{1, \ldots, m\} \times \mathcal{Y} \to [0, 1]$ for all $(i, y) \in \{1, \ldots, m\} \times \mathcal{Y}$. Now suppose that, at the t-th round of the game, the player chooses an action according to the mixed strategy (i.e., probability distribution over actions) $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{m,t})$, and suppose the adversary chooses action $y_t \in \mathcal{Y}$. Then the regret for the player is the vector $\boldsymbol{r}_t \in \mathbb{R}^m$, whose j-th component is

$$r_{j,t}(\boldsymbol{p}_t, y_t) = \sum_{k=1}^{m} p_{k,t} L(k, y_t) - L(j, y_t) = \sum_{k=1}^{m} p_{k,t} \left(L(k, y_t) - L(j, y_t) \right) \ . \tag{8}$$

The first term $\sum_{k=1}^{m} p_{k,t} L(k, y)$ is just the expected loss of the predictor, and this is compared to $L(j, y)$, the loss of playing action j. Thus, the j-th component of the drift $\boldsymbol{r}_t(\boldsymbol{p}_t, y)$ measures the expected change in the player's loss if it were to deterministically choose action j, and the adversary did not change his action. If all components of the per-round cumulative regret vector $\boldsymbol{R}_t/t = (\boldsymbol{r}_1 + \ldots + \boldsymbol{r}_t)/t$ were close to zero, then it would mean that the player has played as well as the best pure

action. This notion was made precise by Hannan as follows: A player is *Hannan consistent* if the per-round regret vector $\boldsymbol{R}_t/t$ converges to the zero vector as t grows to infinity.

Our general decision strategy can be used to play repeated games of this type by letting the decision space $\mathcal{X}$ be the set of distributions on the player set $\{1,\ldots,m\}$ of actions and the drift vector be the regret vector (8). Define now a general potential-based mixed strategy $\boldsymbol{p}_t$ by

$$p_{i,t} = \frac{\nabla_i \Phi(\boldsymbol{R}_{t-1})}{\sum_{k=1}^{m} \nabla_k \Phi(\boldsymbol{R}_{t-1})} \tag{9}$$

for $t > 1$ and $p_{i,1} = 1/m$ for $i = 1,\ldots,m$, where Φ is an appropriate twice differentiable additive potential function. It is immediate to see that for any value of the outcome y_t,

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

and therefore condition (1) is satisfied.

The Hedge algorithm (Freund and Schapire, 1997) and the strategy in Blackwell's proof of the approachability theorem are special cases of (9) for, respectively, the exponential potential (3) and the polynomial potential (2) with $p = 2$. Corollaries 1 and 2 imply the Hannan consistency of these two algorithms. Hart and Mas-Colell (2001) characterize the whole class of potentials for which condition (1) yields a Hannan consistent player.

6.1. Multiplicative algorithms for playing repeated games

Next we consider the setup discussed by Freund and Schapire (1999a) for adaptive game playing. Here the game is defined by an $m \times M$ loss matrix S of entries in $[0,1]$. In each round t the row player chooses a row of S according to a mixed strategy $\boldsymbol{p}_t = (p_{1,t},\ldots,p_{m,t})$ and the column player chooses a column of S according to the mixed strategy $\boldsymbol{q}_t = (q_{1,t},\ldots,q_{M,t})$. The row player's loss at time t is

$$S(\boldsymbol{p}_t, \boldsymbol{q}_t) = \sum_{i=1}^{m} \sum_{j=1}^{M} p_{i,t} q_{j,t} S(i,j)$$

and its goal is to achieve a cumulative loss $\sum_{s=1}^{t} S(\boldsymbol{p}_t, \boldsymbol{q}_t)$ almost as small as the cumulative loss $\min_{\boldsymbol{p}} \sum_{s=1}^{t} S(\boldsymbol{p}, \boldsymbol{q}_t)$ of the best fixed mixed strategy.

Freund and Schapire introduce an algorithm, based on a multiplicative updating of weights, defined by

$$p_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} S(i, \boldsymbol{q}_s)\right)}{\sum_{k=1}^{m} \exp\left(-\eta \sum_{s=1}^{t-1} S(k, \boldsymbol{q}_s)\right)} \qquad i = 1, \ldots, m$$

with $p_{i,1}$ set to $1/m$, where $\eta > 0$ is an appropriately chosen constant. Next, we point out that this algorithm is just a special case of the potental-based algorithm (9) defined above. To see this, define the action space $\mathcal{Y}$ of the adversary as the set of all probability distributions $\boldsymbol{q}$ over the columns, and the loss function L by

$$L(i, \boldsymbol{q}) = S(i, \boldsymbol{q}) = \sum_{j=1}^{M} q_j S(i, j) \ .$$

Then we are back to the problem described at the beginning of this section. Indeed, defining the drift vector $\boldsymbol{r}_t$ as in (8), it is immediate to see that the multiplicative weight algorithm of Freund and Schapire is just (9) with the exponential potential (3).

In view of this observation, it is now straightforward to derive performance bounds for the multiplicative update algorithm. Corollary 2 implies that

$$\ln \Phi(\boldsymbol{R}_t) = \ln \sum_{i=1}^{m} e^{\eta R_{i,t}} \leq \ln m + \frac{t\eta^2}{2} \ .$$

To obtain a lower bound for $\ln \Phi(\boldsymbol{R}_t)$ we use again the log-sum inequality (see Section 4.2) to conclude that, for any probability vector $\boldsymbol{p}$,

$$\ln \Phi(\boldsymbol{R}_t) \geq \eta \boldsymbol{R}_t \cdot \boldsymbol{p} + H(\boldsymbol{p}) = \eta \left(\sum_{s=1}^{t} S(\boldsymbol{p}_s, \boldsymbol{q}_s) - \sum_{s=1}^{t} S(\boldsymbol{p}, \boldsymbol{q}_s) \right) + H(\boldsymbol{p}) \ .$$

Comparing the upper and lower bounds for $\ln \Phi(\boldsymbol{R}_t)$ we obtain the following result.

Corollary 7. The multiplicative update algorithm defined above satisfies

$$\frac{1}{t} \sum_{s=1}^{t} S(\boldsymbol{p}_s, \boldsymbol{q}_s) \leq \min_{\boldsymbol{p}} \left(\frac{1}{t} \sum_{s=1}^{t} S(\boldsymbol{p}, \boldsymbol{q}_s) - \frac{H(\boldsymbol{p})}{t\eta} \right) + \frac{\ln m}{t\eta} + \frac{\eta}{2} \ .$$

This bound is very similar to the one derived by Freund and Schapire (1999a). By choosing $\eta = \sqrt{2 \ln m / t}$ and using the nonnegativity of the entropy, we obtain

$$\frac{1}{t} \sum_{s=1}^{t} S(\boldsymbol{p}_s, \boldsymbol{q}_s) \leq \min_{\boldsymbol{p}} \frac{1}{t} \sum_{s=1}^{t} S(\boldsymbol{p}, \boldsymbol{q}_s) + \sqrt{\frac{2 \ln m}{t}} \ ,$$

which is an insignificant improvement over Corollary 4 of Freund and Schapire (1999a).

Of course, potential functions different from the exponential may be used as well. By varying the potential function, we obtain a whole family of algorithms whose performance bounds are straightforward to obtain by Theorem 1.

6.2. Generalized regret in learning with experts

In this section we consider a more general notion of regret, which we call "generalized regret", introduced by Lehrer (2001). As we will see, generalized regret includes, as special cases, several other notions of regret, such as those defined by Fudenberg and Levine (1999) and by Foster and Vohra (1997). According to our definition, a repeated game can be viewed as an on-line prediction problem with a randomized predictor. Hence, we can use generalized regret to analyze such on-line prediction problems. Consider the prediction with experts framework, where $f_{1,t}, \ldots, f_{N,t} \in \{1, \ldots, m\}$ denote the predictions of the experts at time t. For each expert $i = 1, \ldots, N$, define an *activation function* $A_i : \{1, \ldots, m\} \times \mathbb{N} \to \{0, 1\}$. The activation function determines whether the corresponding expert is active based on the current step index t and, possibly, on the predictor's guess k. At each time instant t, the values $A_i(k, t)$, $i = 1, \ldots, N$, $k = 1, \ldots, m$ of the activation function are revealed to the predictor who then decides on his guess $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{m,t})$. Define the *generalized regret* of a randomized predictor with respect to expert i at round t by

$$r_{i,t}(\boldsymbol{p}_t, y_t) = \sum_{k=1}^{m} p_{k,t} \, A_i(k, t) \, (L(k, y_t) - L(f_{i,t}, y_t)) \ . \qquad (10)$$

Hence, the (instantaneous) generalized regret with respect to expert i is nonzero only when expert i is active,

To illustrate the power of this general notion of regret, next we describe some important special cases.

Example. external regret. The simplest special case is when $N = m$, $f_{i,t} = i$, and $A_i(k, t) = 1$ for all k and t. Then the problem

reduces to the one described in the previous section and the drift function becomes just (8). In game theory, this regret is sometimes called "external" (as opposed to the internal regret described below).

Example. SPECIALISTS. The general formulation permits us to consider a much wider family of prediction problems. Examples include variants of the learning with experts framework, such as "shifting experts" or the more general "specialists" (Freund et al., 1997). In the specialists framework, the activation function $A_i(k,t)$ depends arbitrarily on the round index t but not on the actual predictor's guess k. This setup may be useful to model prediction scenarios where experts are allowed to occasionally abstain from predicting (experts may want to abstain for several reasons; for instance, when they are not confident in their prediction). See (Cohen and Singer, 1999) for a practical application of the specialists framework.

Example. INTERNAL REGRET. Here we discuss in detail the special case of the problem of minimizing the so-called internal (or conditional) regret (Hart and Mas-Colell, 2000). Foster and Vohra (1999) survey this notion of a regret and its relationship with the external regret (8). Minimization of the internal regret plays a key role in the construction of adaptive game-playing strategies which achieve, asymptotically, a correlated equilibrium (see Hart and Mas-Colell Hart and Mas-Colell (2000)). The formal description is as follows: the $N = m(m-1)$ experts are labeled by pairs (i,j) for $i \neq j$. Expert (i,j) predicts always i, that is, $f_{(i,j),t} = i$ for all t, and it is active only when the predictor's guess is j, that is, $A_{(i,j)}(k,t) = \mathbf{I}_{\{k=j\}}$. Thus, component (i,j) of the generalized regret vector $\boldsymbol{r}_t(\boldsymbol{p}_t, y) \in I\!R^N$ becomes

$$r_{(i,j),t}(\boldsymbol{p}_t, y) = p_{j,t}(L(j,y) - L(i,y)) \ .$$

Hence, the cumulative internal regret with respect to expert (i,j),

$$R_{(i,j),t} = r_{(i,j),1} + \ldots + r_{(i,j),t}$$

is the total amount by which the predictor's cumulative loss would have increased, had he predicted i at every time he predicted j. Thus, $R_{(i,j),t}$ may be interpreted as the regret the predictor feels of not having predicted i each time he predicted j.

The internal regret is formally similar to the external regret (8) but there are some important differences. First of all, it is easy to see that internal regret is stronger than the usual regret (8) in the sense that if a player succeeds in keeping the internal regret small for each pairs (i,j), then every component of the per-round cumulative

external regret vector stays close to zero as well. Indeed, assume that $\max\{0, R_{(i,j),t}\} \leq a_t = o(t)$ for all possible pairs (i,j) and for some sequence $a_t \geq 0$, $t \geq 1$. Let $k \in \{1,\ldots,m\}$ be the action with minimal cumulative loss, that is, $\sum_{s=1}^{t} L(k,y_s) = \min_{1\leq i\leq m} \sum_{s=1}^{t} L(i,y_s)$. Then the cumulative regret based on (8) is just

$$\sum_{s=1}^{t}\left(\sum_{j=1}^{m} p_{j,t} L(j,y_s) - L(k,y_s)\right) = \sum_{j=1}^{m} R_{(k,j),t} \leq m a_t = o(t) \ .$$

Thus, small cumulative internal regret implies small cumulative regret of the form considered in the experts' framework. On the other hand, it is easy to show by example that, for $m \geq 3$, small cumulative regret does not imply small internal regret, and in fact, it is significantly more difficult to construct strategies which achieve a small internal regret.

The key question now is whether it is possible to define a predictor $\boldsymbol{p}_t$ satisfying condition (1) for the generalized regret. The existence of such $\boldsymbol{p}_t$ is shown in the next result. For such a predictor we may then apply Theorem 1 and its corollaries to obtain performance bounds without further work.

Theorem 3. Consider a decision problem described above with drift function (10) and potential Φ, where $\nabla\Phi \geq \mathbf{0}$. Then a randomized predictor satisfying condition (1) is defined by the unique solution to the set of m linear equations

$$p_{k,t} = \frac{\sum_{j=1}^{m} p_{j,t} \sum_{i=1}^{N} \mathbf{I}_{\{f_{i,t}=k\}} A_i(j,t) \nabla_i \Phi(\boldsymbol{R}_{t-1})}{\sum_{i=1}^{N} A_i(k,t) \nabla_i \Phi(\boldsymbol{R}_{t-1})} \qquad k = 1,\ldots,m.$$

Observe that in the special case of $N = m$ and $f_{i,t} = i$ and $A_i(k,t) = 1$, the predictor of Theorem 3 reduces to the predictor (9). The proof of the theorem, which we relegate to the appendix, is a generalization of a proof contained in (Hart and Mas-Colell, 2000).

We return now to the special case of internal regret. Hart and Mas-Colell (2000) first proved the existence of a predictor for which the maximal cumulative internal regret $\max_{j,k} R_{(j,k),t}$ is $o(t)$. Indeed, their algorithm is just the special case of the predictor of Theorem 3 for the polynomial potential with $p = 2$. For the algorithm of Hart and Mas-Colell one obtains a bound of the form $\max_{j,k} R_{(j,k),t} = O(\sqrt{tm})$. This bound may be improved significantly for large values of m by considering the predictor of Theorem 3 with other potential functions. For example, if Φ is the exponential potential (3), then a straightforward combination of Theorem 3 and Corollary 2 implies the following:

Corollary 8. If the randomized predictor of Theorem 3 is run with the exponential potential (3) and parameter $\eta = \sqrt{4 \ln m / t}$, then for all sequences $y_1, y_2, \ldots \in \mathcal{Y}$ its internal regret satisfies

$$\max_{j,k} R_{(j,k),t} \leq 2\sqrt{t \ln m} \ .$$

A similar bound may be obtained by using the polynomial potential with exponent $p = 2 \ln N$. It is an interesting open question to find the best possible bound for $\max_{j,k} R_{(j,k),t}$ in terms of m.

Appendix

Proof. We write condition (1) as follows:

$$\begin{aligned}
&\nabla \Phi(\boldsymbol{R}_{t-1}) \cdot \boldsymbol{r}_t \\
&\quad = \sum_{i=1}^{N} \nabla_i \Phi(\boldsymbol{R}_{t-1}) \sum_{k=1}^{m} p_{k,t} A_i(k,t) \left[L(k, y_t) - L(f_{i,t}, y_t) \right] \\
&\quad = \sum_{k=1}^{m} \sum_{j=1}^{m} \sum_{i=1}^{N} \mathbf{I}_{\{f_{i,t}=j\}} \nabla_i \Phi(\boldsymbol{R}_{t-1}) p_{k,t} A_i(k,t) \left[L(k, y_t) - L(f_{i,t}, y_t) \right] \\
&\quad = \sum_{k=1}^{m} \sum_{j=1}^{m} \sum_{i=1}^{N} \mathbf{I}_{\{f_{i,t}=j\}} \nabla_i \Phi(\boldsymbol{R}_{t-1}) p_{k,t} A_i(k,t) L(k, y_t) \\
&\qquad - \sum_{k=1}^{m} \sum_{j=1}^{m} \sum_{i=1}^{N} \mathbf{I}_{\{f_{i,t}=j\}} \nabla_i \Phi(\boldsymbol{R}_{t-1}) p_{k,t} A_i(k,t) L(f_{i,t}, y_t) \\
&\quad = \sum_{k=1}^{m} \sum_{j=1}^{m} \sum_{i=1}^{N} \mathbf{I}_{\{f_{i,t}=j\}} \nabla_i \Phi(\boldsymbol{R}_{t-1}) p_{k,t} A_i(k,t) L(k, y_t) \\
&\qquad - \sum_{k=1}^{m} \sum_{j=1}^{m} \sum_{i=1}^{N} \mathbf{I}_{\{f_{i,t}=k\}} \nabla_i \Phi(\boldsymbol{R}_{t-1}) p_{j,t} A_i(j,t) L(k, y_t) \\
&\quad = \sum_{k=1}^{m} L(k, y_t) \left[\sum_{i=1}^{N} \nabla_i \Phi(\boldsymbol{R}_{t-1}) p_{k,t} A_i(k,t) \right. \\
&\qquad \left. - \sum_{j=1}^{m} \sum_{i=1}^{N} \mathbf{I}_{\{f_{i,t}=k\}} \nabla_i \Phi(\boldsymbol{R}_{t-1}) p_{j,t} A_i(j,t) \right] \leq 0 \ .
\end{aligned}$$

Since the $L(k, y_t)$ are arbitrary and nonnegative, the above is implied by

$$\sum_{i=1}^{N} \nabla_i \Phi(\boldsymbol{R}_{t-1}) p_{k,t} A_i(k,t) - \sum_{j=1}^{m} \sum_{i=1}^{N} \mathbf{I}_{\{f_{i,t}=k\}} \nabla_i \Phi(\boldsymbol{R}_{t-1}) p_{j,t} A_i(j,t) \leq 0 \tag{11}$$

for each $k = 1, \ldots, m$. Solving for $p_{k,t}$ yields the result.

We now check that such a predictor always exists. Let M be the $(m \times m)$ matrix whose entries are

$$M_{k,j} = \frac{\sum_{i=1}^{N} \mathbf{I}_{\{f_{i,t}=k\}} \nabla_i \Phi(\boldsymbol{R}_{t-1}) A_i(j,t)}{\sum_{i=1}^{N} \nabla_i \Phi(\boldsymbol{R}_{t-1}) A_i(k,t)} \ .$$

Then condition (11) is implied by $M\boldsymbol{p} = \boldsymbol{p}$. As $\nabla\Phi \geq \mathbf{0}$, M is nonnegative, and thus the eigenvector equation $M\boldsymbol{p} = \boldsymbol{p}$ has a positive solution by the Perron-Frobenius theorem (Seneta, 1981). □

Acknowledgements

Both authors acknowledge partial support of ESPRIT Working Group EP 27150, Neural and Computational Learning II (NeuroCOLT II). The work of the second author was also supported by DGI grant BMF2000-0807.

References

Auer, P., Cesa-Bianchi, N. and Gentile, C. (2002). Adaptive and Self-Confident On-Line Learning Algorithms. *Journal of Computer and System Sciences*, 64:1, 48-75.

Blackwell, D. (1956). An Analog of the Minimax Theorem for Vector Payoffs. *Pacific Journal of Mathematics*, 6, 1–8.

Block, H. (1962). The Perceptron: A Model for Brain Functioning. *Review of Modern Physics*, 34, 123–135.

Bregman, L. (1967). The Relaxation Method of Finding the Common Point of Convex Sets and its Application to the Solutions of Problems in Convex Programming. *USSR Computational Mathematics and Physics*, 7, 200–217.

Cesa-Bianchi, N. (1999). Analysis of Two Gradient-based Algorithms for On-line Regression. *Journal of Computer and System Sciences*, 59:3, 392–411.

Cesa-Bianchi, N., Freund, Y., Helmbold, D., Haussler, D., Schapire, R., and Warmuth, M. (1997). How to Use Expert Advice. *Journal of the ACM*, 44:3, 427–485.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2002). A Second-Order Perceptron Algorithm. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*. Lecture Notes in Artificial Intelligence, Springer, 2002.

Cohen, W. and Singer, Y. (1999). Context-Sensitive Learning Methods for Text Categorization. *ACM Transactions on Information Systems*, 17:2, 141–173.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley and Sons.

Foster, D. and Vohra, R. (1997). Calibrated Learning and Correlated Equilibrium. *Games and Economic Behaviour*, 21, 40–55.

Foster, D. and Vohra, R. (1999). Regret in the On-line Decision Problem. *Games and Economic Behaviour*, 29:1/2, 7–35.

Freund, Y. and Schapire, R. (1997). A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55:1, 119–139.

Freund, Y. and Schapire, R. (1999a). Adaptive Game Playing Using Multiplicative weights. *Games and Economic Behavior*, 29, 79–103.

Freund, Y. and Schapire, R. (1999b). Large Margin Classification Using the Perceptron algorithm. *Machine Learning* 37:3, 277–296.

Freund, Y., Schapire, R., Singer, Y., and Warmuth, M. (1997). Using and Combining Predictors that Specialize. In *Proceedings of the 29th Annual ACM Symposium on the Theory of Computing*. (pp. 334–343). ACM Press.

Fudenberg, D. and Levine, D. (1995). Universal Consistency and Cautious Fictitious Play. *Journal of Economic, Dynamics, and Control*, 19, 1065–1090.

Fudenberg, D. and Levine, D. (1999). Conditional Universal Consistency. *Games and Economic Behaviour*, 29, 7–35.

Gentile, C. (2001). The Robustness of the p-norm Algorithms. An extended abstract (co-authored with N. Littlestone) appeared in the *Proceedings of the 12th Annual Conference on Computational Learning Theory*, pages 1–11. ACM Press, 1999.

Gentile, C. and Warmuth, M. (1999). Linear Hinge Loss and Average Margin. In *Advances in Neural Information Processing Systems 10*. MIT Press.

Grove, A., Littlestone, N., and Schuurmans, D. (1997). General Convergence Results for Linear Discriminant Updates. In *Proceedings of the 10th Annual Conference on Computational Learning Theory*. (pp. 171–183). ACM Press.

Grove, A., Littlestone, N., and Schuurmans, D. (2001). General Convergence Results for Linear Discriminant Updates. *Machine Learning*, 43:3, 173–210.

Hannan, J. (1957). Approximation to Bayes Risk in Repeated Play. *Contributions to the theory of games*, 3, 97–139.

Hart, S. and Mas-Colell, A. (2000). A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica*, 68, 1127–1150.

Hart, S. and Mas-Colell, A. (2001). A General Class of Adaptive Strategies. *Journal of Economic Theory*, 98:1, 26–54.

Haussler, D., Kivinen, J., and Warmuth, M. (1998). Sequential Prediction of Individual Sequences Under General Loss Functions. *IEEE Transactions on Information Theory*, 44, 1906–1925.

Kivinen, J. and Warmuth, M. (1999). Averaging Expert Predictions. In *Proceedings of the Fourth European Conference on Computational Learning Theory*. (pp. 153–167). Lecture Notes in Artificial Intelligence, Vol. 1572. Springer.

Kivinen, J. and Warmuth, M. (2001). Relative Loss Bounds for Multidimensional Regression Problems. *Machine Learning*, 45:3, 301–329.

Lehrer, E. (2001). A Wide Range No-regret Theorem. *Games and Economic Behavior*. To appear.

Littlestone, N. (1989). Mistake Bounds and Logarithmic Linear-threshold Learning Algorithms. Ph.D. thesis, University of California at Santa Cruz.

Littlestone, N. and Warmuth, M. (1994). The Weighted Majority Algorithm. *Information and Computation*, 108, 212–261.

Novikoff, A. (1962). On Convergence Proofs of Perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, Vol. XII. (pp. 615–622).

Rosenblatt, F. (1962). *Principles of Neurodynamics*. Spartan Books.

Schapire, R. (2001). Drifting Games. *Machine Learning*, 43:5, 265–291.

Seneta, E. (1981, *Non-negative Matrices and Markov Chains*. Springer.

Vovk, V. (1990). Aggregating Strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*. (pp. 372–383).

Vovk, V. (1998). A Game of Prediction with Expert Advice. *Journal of Computer and System Sciences*, 56:2, 153–173.

Vovk, V. (2001). Competitive On-line Statistics. *International Statistical Review*, 69, 213–248.

Warmuth, M. and Jagota, A. (1997). Continuous and Discrete-Time Nonlinear Gradient Descent: Relative Loss Bounds and Convergence. In *Electronic proceedings of the 5th International Symposium on Artificial Intelligence and Mathematics*.