Improved Second-Order Bounds for Prediction with Expert Advice*

Nicolò Cesa-Bianchi[1], Yishay Mansour[2,**], and Gilles Stoltz[3]

[1] DSI, Università di Milano, via Comelico 39, 20135 Milano, Italy
cesa-bianchi@dsi.unimi.it
[2] School of computer Science, Tel-Aviv University, Tel Aviv, Israel
mansour@cs.tau.ac.il
[3] DMA, Ecole Normale Supérieure, 45, rue d'Ulm, 75005 Paris, France
gilles.stoltz@ens.fr

Abstract. This work studies external regret in sequential prediction games with arbitrary payoffs (nonnegative or non-positive). External regret measures the difference between the payoff obtained by the forecasting strategy and the payoff of the best action. We focus on two important parameters: M, the largest absolute value of any payoff, and Q^*, the sum of squared payoffs of the best action. Given these parameters we derive first a simple and new forecasting strategy with regret at most order of $\sqrt{Q^*(\ln N)} + M \ln N$, where N is the number of actions. We extend the results to the case where the parameters are unknown and derive similar bounds. We then devise a refined analysis of the weighted majority forecaster, which yields bounds of the same flavour. The proof techniques we develop are finally applied to the adversarial multi-armed bandit setting, and we prove bounds on the performance of an online algorithm in the case where there is no lower bound on the probability of each action.

1 Introduction

The study of online forecasting strategies in adversarial settings has received considerable attention in the last few years in the computational learning literature and elsewhere. The main focus has been on deriving simple online algorithms that have low external regret. The external regret of an online algorithm is the difference between its expected payoff and the best payoff achievable using some strategy from a given class. Usually, this class includes a strategy, for each action, which always plays that action. In a nutshell, one can show that the average external regret per time step vanishes, and much of the research has been to both

* The work of all authors was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.
** The work was done while the author was a fellow in the Institute of Advance studies, Hebrew University. His work was also supported by a grant no. 1079/04 from the Israel Science Foundation and an IBM faculty award.

P. Auer and R. Meir (Eds.): COLT 2005, LNAI 3559, pp. 217–232, 2005.

improve and refine the bounds. Ideally, in an adversarial setting one should be able to show that the regret with respect to any action only depends on the variance of the observed payoffs for that action. In a stochastic setting such a result seems like the most natural bound, and deriving its analogue in an adversarial setting would be a fundamental result. We believe that our results make a significant step toward this goal, although, unfortunately, fall short of completely achieving it.

In order to describe our results we first set up our model and notations, and relate them to previous works. In this paper we consider the following game-theoretic version of the prediction-with-expert-advice framework [5, 11, 13]. A forecaster repeatedly assigns probabilities to a fixed set of actions. After each assignment, the real payoff associated to each action is revealed and new payoffs are set for the next round. The forecaster's reward on each round is the average payoff of actions for that round, where the average is computed according to the forecaster's current probability assignment. The goal of the forecaster is to achieve, on any sequence of payoffs, a cumulative reward close to X^*, the highest cumulative payoff among all actions. As usual, we call regret the difference between X^* and the cumulative reward achieved by the forecaster on the same payoff sequence.

The special case of "one-sided games", when all payoffs have the same sign (they are either always non-positive or always nonnegative) has been considered by Freund and Schapire [9], and by Auer et al. [3] in a related context. These papers show that Littlestone and Warmuth's weighted majority algorithm [11] can be used as a basic ingredient to construct a forecasting strategy achieving a regret of $O(\sqrt{M|X^*|\ln N})$ in one-sided games, where N is the number of actions and M is a known upper bound on the size of payoffs. (If all payoffs are non-positive, then the absolute value of each payoff is called *loss* and $|X^*|$ is the cumulative loss of the best action.) By a simple rescaling of payoffs, it is possible to reduce the more general "signed game", in which each payoff might have an arbitrary sign, to either one of the one-sided games (note that this reduction assumes knowledge of M). However, the regret becomes $O(M\sqrt{n \ln N})$, where n is the number of game rounds. Recently, Allenberg and Neeman [2] proposed a direct analysis of the signed game avoiding this reduction. Before describing their results, we introduce some convenient notation and terminology.

Our forecasting game is played in rounds. At each time step $t = 1, 2, \ldots$ the forecaster computes an assignment $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{N,t})$ of probabilities over the N actions. Then the payoff vector $\boldsymbol{x}_t = (x_{1,t}, \ldots, x_{N,t}) \in \mathbb{R}^N$ for time t is revealed and the forecaster's reward is $\widehat{x}_t = x_{1,t}p_{1,t} + \ldots + x_{N,t}p_{N,t}$. We define the cumulative reward of the forecaster by $\widehat{X}_n = \widehat{x}_1 + \ldots + \widehat{x}_n$ and the cumulative payoff of action i by $X_{i,n} = x_{i,1} + \ldots + x_{i,n}$. For all n, let $X_n^* = \max_{i=1,\ldots,N} X_{i,n}$ be the cumulative payoff of the best action up to time n. The forecaster's goal is to keep the *regret* $X_n^* - \widehat{X}_n$ as small as possible uniformly over n.

The one-sided games, mentioned above, are the *loss game*, where $x_{i,t} \leq 0$ for all i and t, and the *gain game*, where $x_{i,t} \geq 0$ for all i and t. We call *signed game* the setup in which no assumptions are made on the sign of the

payoffs. For the signed game, Allenberg and Neeman [2] show that weighted majority (used in conjunction with a doubling trick) achieves the following: on any sequence of payoffs there exists an action j such that the regret is at most of order $\sqrt{M(\ln N)\sum_{t=1}^{n}|x_{j,t}|}$, where $M = \max_{i,t}|x_{i,t}|$ is a known upper bound on the size of payoffs. Note that this bound does not relate the regret to the sum $|x_1^*| + \ldots + |x_n^*|$ of payoff sizes for the optimal action (i.e., the one achieving X_n^*). In particular, the bound $O(\sqrt{M|X_n^*|\ln N})$ for the one-sided games is only obtained if an estimate of X_n^* is available in advance.

In this paper we show new regret bounds for the signed game. Our analysis has two main advantages: first, no preliminary knowledge of the payoff size M or about the best cumulative payoff X_n^* is needed; second, our bounds are expressed in terms of sums of squared payoffs, such as $x_{i,1}^2 + \ldots + x_{i,n}^2$ and related forms. These quantities replace the larger terms $M(|x_{i,1}| + \ldots + |x_{i,n}|)$ appearing in the previous bounds. As an application of our results we obtain, without any preliminary knowledge on the payoff sequence, an improved regret bound for the one-sided games of the order of $\sqrt{(Mn - |X_n^*|)(|X_n^*|/n)(\ln N)}$.

Expressions involving squared payoffs are at the core of many analyses in the framework of prediction with expert advice, especially in the presence of limited feedback. (See, for instance, the bandit problem [3] and more generally prediction under partial monitoring [6, 7, 12]). However, to the best of our knowledge, our bounds are the first ones to explicitely include second-order information extracted from the payoff sequence. In particular, our bounds are stable under many transformations of the payoff sequence, and therefore are in some sense more "fundamental".

Some of our bounds are achieved using forecasters based on weighted majority run with a dynamic learning rate. However, we are able to obtain second-order bounds of a different flavour using a new forecaster that does not use the exponential probability assignments of weighted majority. In particular, unlike virtually all previously known forecasting schemes, the weights of this forecaster can not be represented as the gradient of an additive potential [8].

In bandit problems and, more generally, in all incomplete information problems like label-efficient prediction or prediction with partial monitoring, a crucial point is to estimate the unobserved losses. In such settings, a probability distribution is formed by using weighted averages of the cumulative estimated losses, and a common practice is to mix this probability distribution, so that the resulting distribution have all the probabilities above a certain value. Technically, this is important since it is common to divide by the probabilities (see [3, 6, 7, 10, 12]). We show that, for the algorithm of [3], using our proof technique one can simply use the original probability distribution computed with the estimates without any adjustments.

2 A New Algorithm for Sequential Prediction

We introduce a new forecasting strategy for the signed game. In Theorem 3, the main result of this section, we show that, without any preliminary knowledge of

the sequence of payoffs, the regret of a variant of this strategy is bounded by a quantity defined in terms of the sums $Q_{i,n} = x_{i,1}^2 + \ldots + x_{i,n}^2$. Since $Q_{i,n} \leq M(|x_{i,1}| + \ldots + |x_{i,n}|)$, such second-order bounds are generally better than the previously known bounds (see Section 4).

Our basic forecasting strategy, which we call $\texttt{prod}(\eta)$, has an input parameter $\eta > 0$ and maintains a set of N weights. At time $t = 1$ the weights are initialized with $w_{i,1} = 1$ for $i = 1, \ldots, N$. At each time $t = 1, 2, \ldots$, $\texttt{prod}(\eta)$ computes the probability assignment $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{N,t})$, where $p_{i,t} = w_{i,t}/W_t$. After the payoff vector $\boldsymbol{x}_t$ is revealed, the weights are updated using the rule $w_{i,t+1} = w_{i,t}(1 + \eta x_{i,t})$. We use the notation $W_t = w_{1,t} + \ldots + w_{N,t}$. The following simple fact, whose proof is omitted, plays a key role in our analysis.

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

Lemma 2. *Assume there exists $M > 0$ such that the payoffs satisfy $x_{i,t} \geq -M$ for $t = 1, \ldots, n$ and $i = 1, \ldots, N$. For any sequence of payoffs, for any action k, for any $\eta \leq 1/(2M)$, and for any $n \geq 1$, the cumulative reward of $\texttt{prod}(\eta)$ is lower bounded as*

$$\widehat{X}_n \geq X_{k,n} - \frac{\ln N}{\eta} - \eta\, Q_{k,n} \ .$$

Proof. For any $k = 1, \ldots, N$, note that $x_{k,t} \geq -M$ and $\eta \leq 1/(2M)$ imply $\eta x_{k,t} \geq -1/2$. Hence, we can apply Lemma 1 to $\eta x_{k,t}$ and get

$$\begin{aligned} \ln \frac{W_{n+1}}{W_1} &= -\ln N + \ln \prod_{t=1}^{n} (1 + \eta x_{k,t}) = -\ln N + \sum_{t=1}^{n} \ln(1 + \eta x_{k,t}) \\ &\geq -\ln N + \sum_{t=1}^{n} \left(\eta x_{k,t} - \eta^2 x_{k,t}^2\right) = -\ln N + \eta X_{k,n} - \eta^2 Q_{k,n} \ . \quad (1) \end{aligned}$$

On the other hand,

$$\ln \frac{W_{n+1}}{W_1} = \sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t} = \sum_{t=1}^{n} \ln \left(\sum_{i=1}^{N} p_{i,t} \left(1 + \eta x_{i,t}\right) \right) \leq \eta \widehat{X}_n \qquad (2)$$

where in the last step we used $\ln(1 + z_t) \leq z_t$ for all $z_t = \eta \sum_{i=1}^{N} x_{i,t} p_{i,t} \geq -1/2$. Combining (1) and (2), and dividing by $\eta > 0$, we get

$$\widehat{X}_n \geq -\frac{\ln N}{\eta} + X_{k,n} - \eta\, Q_{k,n}$$

which completes the proof of the lemma. □

By choosing η appropriately, we can optimize the bound as follows.

Theorem 1. *Assume there exists $M > 0$ such that the payoffs satisfy $x_{i,t} \geq -M$ for $t = 1, \ldots, n$ and $i = 1, \ldots, N$. For any $Q > 0$, if $\texttt{prod}(\eta)$ is run with*

$$\eta = \min \left\{ 1/(2M), \sqrt{(\ln N)/Q} \right\}$$

then for any sequence of payoffs, for any action k, and for any $n \geq 1$ such that $Q_{k,n} \leq Q$,

$$\widehat{X}_n \geq X_{k,n} - \max\left\{2\sqrt{Q \ln N}, 4M \ln N\right\} \ .$$

To achieve the bound stated in Theorem 1, the parameter η must be tuned using preliminary knowledge of a lower bound on the payoffs and an upper bound on the quantities $Q_{k,n}$. The next two results remove these requirements one by one. We start by introducing a new algorithm that, using a doubling trick over `prod`, avoids any preliminary knowledge of a lower bound on the payoffs.

Let `prod-M`(Q) be the prediction algorithm that receives a number $Q > 0$ as input parameter and repeatedly runs `prod`(η_r), where $\eta_r = 1/(2M_r)$ and M_r is defined below. We call epoch r the sequence of time steps when `prod-M` is running `prod`(η_r). At the beginning, $r = 0$ and `prod-M`(Q) runs `prod`(η_0), where

$$M_0 = \sqrt{Q/(4 \ln N)} \qquad \text{and} \quad \eta_0 = 1/(2M_0) = \sqrt{(\ln N)/Q} \ .$$

The last step of epoch $r \geq 0$ is the time step $t = t_r$ when $\max_{i=1,\dots,N} |x_{i,t}| > M_r$ happens for the first time. When a new epoch $r+1$ begins, `prod` is restarted with parameter $\eta_{r+1} = 1/(2M_{r+1})$, where $M_{r+1} = \max_i 2^{\lceil \log_2 |x_{i,t_r}| \rceil}$. Note that $M_1 \geq M_0$ and, for each $r \geq 1$, $M_{r+1} \geq 2M_r$.

Theorem 2. *For any sequence of payoffs, for any action k, and for any $n \geq 1$ such that $Q_{k,n} \leq Q$, the cumulative reward of algorithm* `prod-M`(Q) *is lower bounded as*

$$\widehat{X}_n \geq X_{k,n} - 2\sqrt{Q \ln N} - 4M\,(2 + 3 \ln N)$$

where $M = \max_{1 \leq i \leq N} \max_{1 \leq t \leq n} |x_{i,t}|$.

Proof. We denote by R the index of the last epoch and let $t_R = n$. If we have only one epoch, then the theorem follows from Theorem 1 applied with a lower bound of $-M_0$ on the payoffs. Therefore, for the rest of the proof we assume $R \geq 1$. Let

$$X_k^r = \textstyle\sum_{s=t_{r-1}+1}^{t_r-1} x_{k,s}, \quad Q_k^r = \sum_{s=t_{r-1}+1}^{t_r-1} x_{k,s}^2, \quad \widehat{X}^r = \sum_{s=t_{r-1}+1}^{t_r-1} \widehat{x}_s \ ,$$

where the sums are over all the time steps t in epoch r except the last one, t_r. (Here t_{-1} is conventionally set to 0.) Applying Lemma 1 to each epoch $r = 0, \dots, R$ we get that $\widehat{X}_n - X_{k,n}$ is equal to

$$\sum_{r=0}^{R} \left(\widehat{X}^r - X_k^r\right) + \sum_{r=0}^{R-1} (\widehat{x}_{t_r} - x_{k,t_r}) \geq -\sum_{r=0}^{R} \frac{\ln N}{\eta_r} - \sum_{r=0}^{R} \eta_r Q_k^r + \sum_{r=0}^{R-1} (\widehat{x}_{t_r} - x_{k,t_r}) \ .$$

We bound each sum separately. For the first sum note that

$$\sum_{r=0}^{R} \frac{\ln N}{\eta_r} = \sum_{r=0}^{R} 2M_r \ln N \leq 6M_R \ln N$$

since $M_R \geq 2^{R-r} M_r$ for each $r \geq 1$ and $M_0 \leq M_R$. For the second sum, using that the η_r decrease, we have

$$\sum_{r=0}^{R} \eta_r Q_k^r \leq \eta_0 \sum_{r=0}^{R} Q_k^r \leq \eta_0 Q_{k,n} \leq \sqrt{\frac{\ln N}{Q}}\, Q = \sqrt{Q \ln N}\ .$$

Finally,

$$\sum_{r=0}^{R-1} |\widehat{x}_{t_r} - x_{k,t_r}| \leq \sum_{r=1}^{R} 2\, M_r \leq 4\, M_R\ .$$

The resulting lower bound $2M_R(2+3\ln N) + \sqrt{Q \ln N}$ implies the one stated in the theorem by noting that, when $R \geq 1$, $M_R \leq 2M$. □

We now show a regret bound for the case when M and the $Q_{k,n}$ are both unknown. Let k_t^* be the index of the best action up to time t; that is, $k_t^* \in \operatorname{argmax}_k X_{k,t}$ (ties are broken by choosing the action k with minimal associated $Q_{k,t}$). We denote the associated quadratic penalty by

$$Q_t^* = Q_{k_t^*}^* = \textstyle\sum_{s=1}^{t} x_{k_t^*,s}^2\ .$$

Ideally, our final regret bound should depend on Q_n^*. However, note that the sequence $Q_1^*, Q_2^*, \ldots$ is not necessarily monotone, as Q_t^* and Q_{t+1}^* cannot be possibly related when the actions achieving the largest cumulative payoffs at rounds t and $t+1$ are different. Therefore, we cannot use a straightforward doubling trick, as this only applies to monotone sequences. Our solution is to express the bound in terms of the smallest nondecreasing sequence that upper bounds the original sequence $(Q_t^*)_{t \geq 1}$. This is a general trick to handle situations where the penalty terms are not monotone. Allenberg and Neeman [2] faced a similar situation, and we improve their results.

We define a new (parameterless) prediction algorithm `prod-MQ` in the following way. The algorithm runs in epochs using `prod-M`(Q) as a subroutine. The last step of epoch r is the time step $t = t_r$ when $Q_t^* > 4^r$ happens for the first time. At the beginning of each new epoch $r = 0, 1, \ldots$, algorithm `prod-M`(Q) is restarted with parameter $Q = 4^r$.

Theorem 3. *For any sequence of payoffs and for any $n \geq 1$, the cumulative reward of algorithm* `prod-MQ` *satisfies*

$$\widehat{X}_n \geq X_n^* - 8\sqrt{(\ln N) \max\left\{1,\, \max_{s \leq n} Q_s^*\right\}} - 12M\left(2 + \log_4 \max_{s \leq n} Q_s^*\right)(1 + \ln N)$$

where $M = \max_{1 \leq i \leq N} \max_{1 \leq t \leq n} |x_{i,t}|$.

Proof. We denote by R the index of the last epoch and let $t_R = n$. Assume that $R \geq 1$ (otherwise the proof is concluded by Theorem 2). Similarly to the proof of Theorem 2, for all epochs r and actions k introduce

$$X_k^r = \textstyle\sum_{s=t_{r-1}+1}^{t_r - 1} x_{k,s}\ , \quad Q_k^r = \sum_{s=t_{r-1}+1}^{t_r - 1} x_{k,s}^2\ , \quad \widehat{X}^r = \sum_{s=t_{r-1}+1}^{t_r - 1} \widehat{x}_s$$

where $t_{-1} = 0$. We also denote $k_r = k^*_{t_r-1}$ the index of the best overall expert up to time $t_r - 1$ (one time step before the end of epoch r). We have that $Q^r_{k_r} \leq Q_{k_r,t_r-1} = Q^*_{t_r-1}$. Now, by definition of the algorithm, $Q^*_{t_r-1} \leq 4^r$. Theorem 2 (applied to time steps $t_{r-1}+1, \ldots, t_r - 1$) shows that $\widehat{X}^r \geq X^r_{k_r} - \Phi\left(M, 4^r\right)$, where $\Phi(M, x) = 2\sqrt{x \ln N} + 4M(2 + 3 \ln N)$. Summing over $r = 0, \ldots, R$ we get

$$\widehat{X}_n = \sum_{r=0}^{R} \widehat{X}^r + \widehat{x}_{k_r,t_r} \geq \sum_{r=0}^{R} \left(\widehat{x}_{k_r,t_r} + X^r_{k_r} - \Phi\left(M, 4^r\right)\right) \ . \tag{3}$$

Now, since k_1 is the index of the expert with largest payoff up to time $t_1 - 1$, we have that $X_{k_2,t_2-1} = X^1_{k_2} + x_{k_2,t_1} + X^2_{k_2} \leq X^1_{k_1} + X^2_{k_2} + M$. By a simple induction, we in fact get

$$X_{k_R,t_R-1} \leq \sum_{r=0}^{R-1} \left(X^r_{k_r} + M\right) + X^R_{k_R} \ . \tag{4}$$

As, in addition, X_{k_R,t_R-1} and $X_{k^*_n,n}$ may only differ by at most M, combining (3) and (4) we have indeed proven that

$$\widehat{X}_n \geq X_{k^*_n,n} - \left(2(1+R)M + \sum_{r=0}^{R} \Phi\left(M, 4^r\right)\right) \ .$$

The sum over r is now bounded as follows

$$\sum_{r=0}^{R} \Phi\left(M, 4^r\right) \leq 4M(1+R)\left(2 + 3\ln N\right) + 2^{R+1}\left(2\sqrt{\ln N}\right) \ .$$

The proof is concluded by noting that, as $R \geq 1$, $\sup_{s \leq n} Q^*_s \geq 4^{R-1}$ by definition of the algorithm. □

3 Second-Order Bounds for Weighted Majority

In this section we derive new regret bounds for the weighted majority forecaster of Littlestone and Warmuth [11] using a time-varying learning rate. This allows us to avoid the doubling trick of Section 2 and keep the assumption that no knowledge on the payoff sequence is available to the forecaster beforehand.

Similarly to the results of Section 2, the main term in the new bounds depends on second-order quantities associated to the sequence of payoffs. However, the precise definition of these quantities makes the bounds of this section generally not comparable to the bounds obtained in Section 2.

The weighted majority forecaster using the sequence $\eta_2, \eta_3, \ldots > 0$ of learning rates assigns at time t a probability distribution $\boldsymbol{p}_t$ over the N experts defined by $\boldsymbol{p}_1 = (1/N, \ldots, 1/N)$ and

$$p_{i,t} = \frac{e^{\eta_t X_{i,t-1}}}{\sum_{j=1}^{N} e^{\eta_t X_{j,t-1}}} \qquad \text{for } i = 1, \ldots, N \text{ and } t \geq 2. \tag{5}$$

Note that the quantities $\eta_t > 0$ may depend on the past payoffs $x_{i,s}$, $i = 1, \ldots, N$ and $s = 1, \ldots, t-1$. The analysis of Auer, Cesa-Bianchi, and Gentile [4], for a related variant of weighted majority, is at the core of the proof of the following lemma (proof omitted from this extended abstract).

Lemma 3. *Consider any nonincreasing sequence $\eta_2, \eta_3, \ldots$ of positive learning rates and any sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \in \mathbb{R}^N$ of payoff vectors. Define the nonnegative function Φ by*

$$\Phi(\boldsymbol{p}_t,\, \eta_t,\, \boldsymbol{x}_t) = -\sum_{i=1}^{N} p_{i,t} x_{i,t} + \frac{1}{\eta_t} \ln \sum_{i=1}^{N} p_{i,t} e^{\eta_t x_{i,t}} = \frac{1}{\eta_t} \ln \left(\sum_{i=1}^{N} p_{i,t} e^{\eta_t (x_{i,t} - \widehat{x}_t)} \right)$$

Then the weighted majority forecaster (5) run with the sequence $\eta_2, \eta_3, \ldots$ satisfies, for any $n \geq 1$ and for any $\eta_1 \geq \eta_2$,

$$\widehat{X}_n - X_n^* \geq -\left(\frac{2}{\eta_{n+1}} - \frac{1}{\eta_1} \right) \ln N - \sum_{t=1}^{n} \Phi(\boldsymbol{p}_t,\, \eta_t,\, \boldsymbol{x}_t) \ .$$

Let Z_t be the random variable with range $\{x_{1,t}, \ldots, x_{N,t}\}$ and law $\boldsymbol{p}_t$. Note that $\mathbb{E} Z_t$ is the expected payoff $\widehat{x}_t$ of the forecaster using distribution $\boldsymbol{p}_t$ at time t. Introduce

$$\operatorname{Var} Z_t = \mathbb{E} Z_t^2 - \mathbb{E}^2 Z_t = \sum_{i=1}^{N} p_{i,t} x_{i,t}^2 - \left(\sum_{i=1}^{N} p_{i,t} x_{i,t} \right)^2 \ .$$

Hence $\operatorname{Var} Z_t$ is the variance of the payoffs at time t under the distribution $\boldsymbol{p}_t$ and the cumulative variance $V_n = \operatorname{Var} Z_1 + \ldots \operatorname{Var} Z_n$ is the main second-order quantity used in this section. The next result bounds $\Phi(\boldsymbol{p}_t,\, \eta_t,\, \boldsymbol{x}_t)$ in terms of $\operatorname{Var} Z_t$.

Lemma 4. *For all payoff vectors $\boldsymbol{x}_t = (x_{1,t}, \ldots, x_{N,t})$, all probability distributions $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{N,t})$, and all learning rates $\eta_t \geq 0$, we have*

$$\Phi(\boldsymbol{p}_t,\, \eta_t,\, \boldsymbol{x}_t) \leq 2\, M$$

where M is such that $|x_{i,t}| \leq M$ for all i. If, in addition, $0 \leq \eta_t |x_{i,t}| \leq 1/2$ for all $i = 1, \ldots, N$, then

$$\Phi(\boldsymbol{p}_t,\, \eta_t,\, \boldsymbol{x}_t) \leq (e-2) \eta_t \operatorname{Var} Z_t \ .$$

Proof. The first inequality is straightforward. To prove the second one we use $e^a \leq 1 + a + (e-2)\, a^2$ for $|a| \leq 1$. Consequently, noting that $\eta_t |x_{i,t} - \widehat{x}_t| \leq 1$ for all i by assumption, we have that

$$\begin{aligned} \Phi(\boldsymbol{p}_t,\, \eta_t,\, \boldsymbol{x}_t) &= \frac{1}{\eta_t} \ln \left(\sum_{i=1}^{N} p_{i,t} e^{\eta_t (x_{i,t} - \widehat{x}_t)} \right) \\ &\leq \frac{1}{\eta_t} \ln \left(\sum_{i=1}^{N} p_{i,t} \left(1 + \eta_t (x_{i,t} - \widehat{x}_t) + (e-2) \eta_t^2 (x_{i,t} - \widehat{x}_t)^2 \right) \right) \ . \end{aligned}$$

Using $\ln(1+a) \le a$ for all $a \ge -1$ and some simple algebra concludes the proof of the second inequality. □

In [3] a very similar result is proven, except that there the variance is further bounded (up to a multiplicative factor) by the expectation $\widehat{x}_t$ of Z_t.

We now introduce a time-varying learning rate based on V_n. For any sequence of payoff vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ and for all $t = 1, 2, \ldots$ let $M_t = 2^k$, where k is the smallest *nonnegative* integer such that $\max_{s=1,\ldots,t} \max_{i=1,\ldots,N} |x_{i,s}| \le 2^k$. Now let the sequence $\eta_2, \eta_3, \ldots$ be defined as

$$\eta_t = \min\left\{\frac{1}{2\,M_{t-1}},\, C\sqrt{\frac{\ln N}{V_{t-1}}}\right\} \quad \text{for } t \ge 2, \text{ with} \quad C = \sqrt{\frac{2}{e-2}\left(\sqrt{2}-1\right)}\,. \tag{6}$$

Note that η_t depends on the forecaster's past predictions. This is in the same spirit as the self-confident learning rates considered in [4].

We are now ready to state and prove the main result of this section.

Theorem 4. *Consider the weighted majority forecaster using the time-varying learning rate (6). Then, for all sequences of payoffs and for all $n \ge 1$,*

$$\widehat{X}_n - X_n^* \ge -4\sqrt{V_n \ln N} - 16\,\max\{M,\,1\}\ln N - 8\,\max\{M,\,1\} - M^2$$

where $M = \max_{t=1,\ldots,n} \max_{i=1,\ldots,N} |x_{i,t}|$.

Proof. We start by applying Lemma 3 using the learning rate (6), and setting $\eta_1 = \eta_2$ for the analysis,

$$\begin{aligned}
\widehat{X}_n - X_n^* &\ge -\left(\frac{2}{\eta_{n+1}} - \frac{1}{\eta_1}\right)\ln N - \sum_{t=1}^{n} \Phi(\boldsymbol{p}_t,\, \eta_t,\, \boldsymbol{x}_t) \\
&\ge -2\max\left\{2\,M_n \ln N,\, (1/C)\sqrt{V_n \ln N}\right\} - \sum_{t=1}^{n} \Phi(\boldsymbol{p}_t,\, \eta_t,\, \boldsymbol{x}_t) \\
&= -2\max\left\{2\,M_n \ln N,\, (1/C)\sqrt{V_n \ln N}\right\} \\
&\quad - \sum_{t\in\mathcal{T}} \Phi(\boldsymbol{p}_t,\, \eta_t,\, \boldsymbol{x}_t) - \sum_{t\notin\mathcal{T}} \Phi(\boldsymbol{p}_t,\, \eta_t,\, \boldsymbol{x}_t)
\end{aligned}$$

where C is defined in (6), and $\mathcal{T}$ is the set of times rounds $t \ge 2$ when $\eta_t |x_{i,t}| \le 1/2$ for all $i = 1, \ldots, N$ (note that $1 \notin \mathcal{T}$ by definition). Using the second bound of Lemma 4 on $t \in \mathcal{T}$ and the first bound of Lemma 4 on $t \notin \mathcal{T}$, which in this case reads $\Phi(\boldsymbol{p}_t,\, \eta_t,\, \boldsymbol{x}_t) \le 2M_t$, we get

$$\begin{aligned}
\widehat{X}_n - X_n^* \ge\; & -2\max\left\{2\,M_n \ln N,\, (1/C)\sqrt{V_n \ln N}\right\} \\
& - (e-2)\sum_{t\in\mathcal{T}} \eta_t \operatorname{Var} Z_t - \sum_{t\notin\mathcal{T}} 2\,M_t
\end{aligned} \tag{7}$$

(where $2M_1$ appears in the last sum). We first note that

$$\sum_{t \notin \mathcal{T}} M_t \leq \sum_{r=0}^{\lceil \log_2 \max\{M,1\} \rceil} 2^r \leq 2^{1+\lceil \log_2 \max\{M,1\} \rceil} \leq 4 \max\{M,1\} \; .$$

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

$$\sum_{t \in \mathcal{T}} \eta_t \operatorname{Var} Z_t \leq M^2 + \sum_{t=T+1}^{n} \eta_t \operatorname{Var} Z_t \; . \tag{8}$$

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

$$\sum_{t=T+1}^{n} \eta_t \operatorname{Var} Z_t \leq C\sqrt{\ln N} \sum_{t=T+1}^{n} \frac{V_t - V_{t-1}}{\sqrt{V_{t-1}}} \; .$$

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

$$\frac{v_t}{\sqrt{V_{t-1}}} = \frac{\sqrt{V_t} + \sqrt{V_{t-1}}}{\sqrt{V_{t-1}}} \left(\sqrt{V_t} - \sqrt{V_{t-1}} \right) \leq (\sqrt{2}+1) \left(\sqrt{V_t} - \sqrt{V_{t-1}} \right) \; . \tag{9}$$

Therefore, using that $\sqrt{2}+1 = 1/(\sqrt{2}-1)$,

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

When $\sqrt{V_n} \geq 2CM_n\sqrt{\ln N}$, using $M_n \geq M$ we have that $\widehat{X}_n - X_n^*$ is at least

$$-\frac{2}{C}\sqrt{V_n \ln N} - \frac{C(e-2)}{\sqrt{2}-1}\sqrt{V_n \ln N} - 8 \max\{M,\,1\} - (e-2)M^2$$
$$\geq -4\sqrt{V_n \ln N} - 8 \max\{M,\,1\} - M^2$$

where we substituted the value of C and obtained a constant for the leading term equal to $2\sqrt{2(e-2)}/\sqrt{\sqrt{2}-1} \leq 3.75$. When $\sqrt{V_n} \leq 2CM_n\sqrt{\ln N}$, using $M_n \leq \max\{1,\,2M\}$ we have that $\widehat{X}_n - X_n^*$ is at least

$$-8\,M \ln N - \frac{C^2 4(e-2)}{\sqrt{2}-1} \max\{1/2,\,M\} \ln N - 8 \max\{M,\,1\} - (e-2)M^2$$
$$\geq -16 \max\{M,\,1\} \ln N - 8 \max\{M,\,1\} - M^2 \; .$$

This concludes the proof. □

4 Applications

To demonstrate the usefulness of the bounds proven in Theorems 3 and 4 we show that they lead to several improvements or extensions of earlier results.

Improvements for Loss Games. Recall the definition of quadratic penalties Q_t^* in Section 2. In case of a loss game (i.e., all payoffs are non-positive), $Q_t^* \leq ML_t^*$, where L_t^* is the cumulative loss of the best action up to time t. Therefore, $\max_{s\leq n} Q_s^* \leq ML_n^*$ and the bound of Theorem 3 is at least as good as the family of bounds called "improvements for small losses" (see, e.g., [4]), whose main term is of the form $\sqrt{ML_n^* \ln N}$. However, it is easy to exhibit examples where the new bound is far better by considering sequences of outcomes where there are some "outliers" among the $x_{i,t}$. These outliers may raise the maximum M significantly, whereas they have only little impact on the $\max_{s\leq n} Q_s^*$.

Using Translations of Payoffs. Recall that Z_t is the random variable which takes the value $x_{i,t}$ with probability $p_{i,t}$, for $i = 1, \ldots, N$. The main term of the bound stated in Theorem 4 contains $V_n = \operatorname{Var} Z_1 + \ldots + \operatorname{Var} Z_n$. Note that V_n is smaller than all quantities of the form $\sum_{t=1}^n \sum_{i=1}^N p_{i,t} \left(x_{i,t} - \mu_t\right)^2$ where $(\mu_t)_{t\geq 1}$ is any sequence of real numbers which may be chosen in *hindsight*, as it is not required for the definition of the forecaster. (The minimal value of the expression is obtained for $\mu_t = \hat{x}_t$.) This gives us a whole family of upper bounds, and we may choose for the analysis the most convenient sequence of μ_t.

To provide a concrete example, denote the effective range of the payoffs at time t by $R_t = \max_{i=1,\ldots,N} x_{i,t} - \min_{j=1,\ldots,N} x_{j,t}$ and consider the choice $\mu_t = \min_{j=1,\ldots,N} x_{j,t} + R_t/2$. The next result improves on a result of Allenberg and Neeman [2], who show a regret bound, in terms of the cumulative effective range, whose main term is $5.7\sqrt{2(\ln N)M \sum_{t=1}^n R_t}$, for a given bound M over the payoffs.

Corollary 1. *The regret of the weighted majority forecaster with variable learning rate (6) satisfies*

$$\widehat{X}_n - X_n^* \geq -2\sqrt{(\ln N)\sum_{t=1}^n R_t^2 - 16 \max\{M, 1\} \ln N - 8 \max\{M, 1\} - M^2} \ .$$

The bound proposed by Corollary 1 shows that for an effective range of M, say if the payoffs all fall in $[0, M]$, the regret is lower bounded by a quantity equal to $-2M\sqrt{n \ln N}$ (a closer look at the proof of Theorem 4 shows that the constant factor may be even equal to 1.9). The best leading constant for such bounds is, to our knowledge, $\sqrt{2}$ (see [8]). This shows that the improved dependence in the bound does not come at a significant increase in the magnitude of the leading coefficient.

Improvements for One-sided Games. The main drawback of V_n, used in Theorem 4, is that it is defined directly in terms of the forecaster's distributions $\boldsymbol{p}_t$.

We now show how this dependence could be removed. Assume $|x_{i,t}| \leq M$ for all t and i. The following corollary of Theorem 4 reveals that weighted majority suffers a small regret in one-sided games whenever $|X_n^*|$ or $Mn - |X_n^*|$ is small (where $|x_{i,t}| \leq M$ for all t and i); that is, whenever $|X_n^*|$ is very small or very large. Improvements of the same flavour were obtained by Auer, Cesa-Bianchi, and Gentile [4] for loss games; however, their result cannot be converted in a straightforward manner to a corresponding useful result for gain games. Allenberg and Neeman [2] proved, in a gain game and for a related algorithm, a bound of the order of $11.4\sqrt{M}\min\bigl\{\sqrt{X_n^*},\, \sqrt{Mn - X_n^*}\bigr\}$. That algorithm was specifically designed to ensure a regret bound of this form, and is different from the algorithm whose performance we discussed before the statement of Corollary 1. Our weighted majority forecaster achieves a better bound, even though it was not directly constructed to do so.

Corollary 2. *Consider the weighted majority forecaster using the time-varying learning rate (6). Then, for all sequences of payoffs in a one-sided game (i.e., payoffs are all non-positive or all nonnegative),*

$$\widehat{X}_n - X_n^* \geq -4\sqrt{|X_n^*|\left(M - \frac{|X_n^*|}{n}\right)\ln N} - 65\,\max\{1\,,M\}\max\{1\,,\ln N\} - 5\,M^2$$

where $M = \max_{t=1,\ldots,n}\max_{i=1,\ldots,N}|x_{i,t}|$.

Proof. We give the proof for a gain game. Since the payoffs are in $[0, M]$, we can write

$$\begin{aligned} V_n &\leq \sum_{t=1}^{n}\left(M\sum_{i=1}^{N}p_{i,t}x_{i,t} - \left(\sum_{i=1}^{N}p_{i,t}x_{i,t}\right)^2\right) = \sum_{t=1}^{n}(M - \widehat{x}_t)\widehat{x}_t \\ &\leq n\left(\frac{M\widehat{X}_n}{n} - \left(\frac{\widehat{X}_n}{n}\right)^2\right) = \widehat{X}_n\left(M - \frac{\widehat{X}_n}{n}\right) \end{aligned}$$

where we used the concavity of $x \mapsto Mx - x^2$. Assume that $\widehat{X}_n \leq X_n^*$ (otherwise the result is trivial). Then, Theorem 4 ensures that

$$\widehat{X}_n - X_n^* \geq -4\sqrt{X_n^*\left(M - \frac{\widehat{X}_n}{n}\right)\ln N} - \kappa$$

where $\kappa = 16\,\max\{M,\,1\}\ln N + 8\,\max\{M,\,1\} + M^2$. We solve for $\widehat{X}_n$ obtaining

$$\widehat{X}_n - X_n^* \geq -4\sqrt{X_n^*\left(M - \frac{X_n^*}{n} + \frac{\kappa}{n}\right)\ln N} - \kappa - 16\frac{X_n^*}{n}\ln N\ .$$

Using the crude upper bound $X_n^*/n \leq M$ and performing some simple algebra, we get the desired result. □

Quite surprisingly, a bound of the same form as the one shown in Corollary 2 can be derived as a Corollary of Theorem 3. The derivation uses a payoff translation technique similar to the one we discussed in the previous paragraph. However, unlike the approach presented there for the weighted majority based forecaster, here the payoffs have to be explicitely translated by the forecaster. (And each translation rule corresponds to a different forecaster.)

A simplified Algorithm for Bandit Loss Games. We close this section with a result that is not a direct consequence of Theorems 3 or 4. Rather, we derive it via an extension of Lemma 4, one of our key results at the core of the second-order analysis in Section 3.

Recall that payoffs $x_{i,t}$ in loss game are all non-positive. We use $\ell_{i,t} = -x_{i,t}$ to denote the loss of action i at time t. Similarly, $\widehat{\ell}_t = \ell_{1,t}p_{1,t} + \ldots + \ell_{N,t}p_{N,t}$ is the loss of the forecaster using $\boldsymbol{p}_t$ as probability assignment at time t. We make the simplifying assumption $\ell_{i,t} \in [0,1]$ for all i, t.

The bandit loss game (see [3] and references therein) is a loss game with the only difference that, at each time step t, the forecaster has no access to the loss vector $\boldsymbol{\ell}_t = (\ell_{1,t}, \ldots, \ell_{N,t})$. Therefore, the loss $\widehat{\ell}_t$ cannot be computed and the individual losses $\ell_{i,t}$ can not be used to adjust the probability assignment $\boldsymbol{p}_t$. The only information the forecaster receives at the end of each round t is the loss $\ell_{I_t,t}$, where I_t takes value i with probability $p_{i,t}$ for $i = 1, \ldots, N$.

In bandit problems and, more generally, in all incomplete information problems like label-efficient prediction or prediction with partial monitoring, a crucial point is to estimate the unobserved losses. In bandit algorithms based on weighted majority, this is usually done by shifting the probability distribution $\boldsymbol{p}_t$ so that all components are larger than a given threshold. Allenberg and Auer [1] apply the shifting technique to weighted majority obtaining, in bandit loss games, a regret bound of order $\sqrt{NL_n^* \ln N} + N \ln(n\,N) \ln n$ where L_n^* is the cumulative loss of the best action after n rounds (note that using the results of [3], derived for gain games, one would only obtain $\sqrt{Nn \ln(n\,N)}$). We show that *without any shifting*, a slight modification of weighted majority achieves a regret of order $N\sqrt{L_n^* \ln n} + N \ln n$. The new bound becomes better than the one by Allenberg and Auer when L_n^* is so small that $L_n^* = o((\ln n)^3)$.

The bandit algorithm, which we call EXP3LIGHT, performs the weight update $w_{i,t+1} = w_{i,t}\, e^{-\eta \widetilde{\ell}_{i,t}}$. The pseudo-losses $\widetilde{\ell}_{i,t}$ are defined by $\widetilde{\ell}_{i,t} = (\ell_{i,t}/p_{i,t})Z_{i,t}$ for $i = 1, \ldots, N$. The Bernoulli random variable $Z_{i,t}$ takes value 1 if the forecaster has drawn action i at time t; i.e., $I_t = i$.

We start with a variant of Lemma 4 for loss games (proof omitted from this extended abstract).

Lemma 5. *For all $\eta > 0$, all losses $\ell_{i,t} \geq 0$, and all sets $S_t \subseteq \{1, \ldots, N\}$,*

$$\Phi(\boldsymbol{p}_t,\, \eta,\, -\boldsymbol{\ell}_t) \leq \frac{\eta}{2} \sum_{i \in S_t} p_{i,t}\, \ell_{i,t}^2 + \sum_{i \in S_t} p_{i,t}\, \ell_{i,t}\ .$$

Lemma 5 is applied as follows (the proofs of Proposition 1 and Theorem 5 are omitted from this extended abstract).

Proposition 1. *Assume the forecaster* Exp3Light *plays a bandit loss game, with losses bounded between 0 and 1. For all $\eta > 0$, the cumulative pseudo-loss of* Exp3Light *satisfies*

$$\widetilde{L}_n \leq \frac{(\ln N) + N(\ln n)}{\eta} + \frac{\eta}{2} N \widetilde{L}^* + \Delta_n$$

where $\widetilde{L}_n = \sum_{t=1}^{n} \sum_{i=1}^{N} p_{i,t} \widetilde{\ell}_{i,t}$, $\widetilde{L}_{k,n} = \sum_{t=1}^{n} \widetilde{\ell}_{k,t}$, $\widetilde{L}^* = \min_{k=1,\ldots,N} \widetilde{L}_{k,n}$,
and Δ_n is a random variable with expectation less than $2\,N$.

Theorem 5. *Consider the forecaster that runs algorithm* Exp3Light *in epochs as follows. In each epoch $r = 0, 1 \ldots$ the algorithm uses*

$$\eta_r = \sqrt{\frac{2\left((\ln N) + N \ln n\right)}{N 4^r}}$$

and epoch r stops whenever the pseudo-loss $\widetilde{L}^*$ in this epoch is larger than 4^r. For any bandit loss game with $\ell_{i,t} \in [0,1]$ for all i and t, the expected cumulative loss of this forecaster satisfies

$$\mathbb{E}\left[\sum_{t=1}^{n} \ell_{I_t,t}\right] - L_n^* \leq 2\sqrt{2\left((\ln N) + N \ln n\right) N \left(1 + 3L_n^*\right)} + (2N+1)\left(1 + \log_4(3n+1)\right) \ .$$

5 Discussion and Open Problems

Though the results of Sections 2 and 3 cannot be easily compared, the two underlying algorithms apply to loss games, gain games, as well as to signed games. In addition, note that the bounds proposed by Theorem 3 and by Theorem 4 (or, more precisely, the variant of this bound using payoffs translated by $\widehat{x}_t$) are both stable under many transformations, such as translations or changes of signs. Consequently, and most importantly, they are invariant under the change $\ell_{i,t} = M - x_{i,t}$, that converts bounded nonnegative payoffs into bounded losses, and vice versa. However, the occurrence of terms like $\max\{M, 1\}$ and M^2 makes these bounds not stable under rescaling of the payoffs. This means that if the payoffs are all multiplied by a positive number α (which may be more or less than 1), then the bounds on the regret are not necessarily multiplied by the same quantity α.

Modifying the proof of Theorem 4 we also obtained a regret bound equal to $-4\sqrt{V_n \ln N} - 16\,M \ln N - 8\,M - 2M \log M^2/V_1$. This bound is indeed stable under rescalings and improves on Theorem 4 for instance when M much smaller than 1, or even when M is large and V_1 is not too small. We hope that the unconvenient factor $1/V_1$ could be removed soon.

A practical advantage of the weighted majority forecaster is that its update rule is completely incremental and never needs to reset the weights. This in contrast to the forecaster `prod-MQ` of Theorem 3 that uses a nested doubling trick. On the other hand, the bound proposed in Theorem 4 is not in closed form, as it still explicitely depends through V_n on the forecaster's rewards $\widehat{x}_t$.

Several issues are left open. The following list mentions some of them.

- Design and analyze incremental updates for the forecaster `prod`(η) of Section 2.
- Obtain second order bounds with updates that are not multiplicative; for instance, updates based on the polynomial potentials (see [8]).
- Extend the analysis of `prod-MQ` to obtain an oracle inequality of the form

$$\widehat{X}_n \geq \max_{k=1,\ldots,N} \left(X_{k,n} - \gamma_1 \sqrt{Q_{k,n} \ln N} \right) - \gamma_2 M \ln N$$

  where γ_1 and γ_2 are absolute constants. Inequalities of this form can be viewed as game-theoretic versions of the model selection bounds in statistical learning theory.

References

1. C. Allenberg-Neeman and P. Auer. Personal communication.
2. C. Allenberg-Neeman and B. Neeman. Full information game with gains and losses. Algorithmic Learning Theory, 15th International Conference, ALT 2004, Padova, Italy, October 2004, Proceedings, volume 3244 of Lecture Notes in Artificial Intelligence, pages 264-278. Springer, 2004.
3. P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32:48–77, 2002.
4. P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64:48–75, 2002.
5. N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, D. Haussler, R. Schapire, and M.K. Warmuth. How to use expert advice. *Journal of the ACM*, 3:427–485, 1997.
6. N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, to appear.
7. N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. Submitted for journal publication, 2004.
8. N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, to appear.
9. Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
10. S. Hart and A. Mas-Colell. A Reinforcement Procedure Leading to Correlated Equilibrium. Economic Essays, Gerard Debreu, Wilhelm Neuefeind and Walter Trockel (editors), Springer (2001), 181-200
11. N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

12. A. Piccolboni and C. Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Proceedings of the 14th Annual Conference on Computational Learning Theory*, pages 208–223, 2001.
13. V.G. Vovk. A Game of Prediction with Expert Advice. *Journal of Computer and System Sciences*, 56(2):153–73, 1998.