Strategies for prediction under imperfect monitoring

Gabor Lugosi, Shie Mannor, Gilles Stoltz

▶ **To cite this version:**

Gabor Lugosi, Shie Mannor, Gilles Stoltz. Strategies for prediction under imperfect monitoring. 2007. hal-00124679v1

HAL Id: hal-00124679

https://hal.science/hal-00124679v1

Preprint submitted on 15 Jan 2007 (v1), last revised 7 Jan 2008 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Strategies for prediction under imperfect monitoring

Gábor Lugosi[1], Shie Mannor[2], and Gilles Stoltz[3]

[1] ICREA and Department of Economics, Universitat Pompeu Fabra Ramon Trias Fargas 25-27, 08005 Barcelona, Spain, lugosi@upf.es
[2] Department of Electrical & Computer Engineering, McGill University, 3480 University Street, Montreal, Québec, Canada H3A-2A7 shie.mannor@mcgill.ca
[3] CNRS and Département de mathématiques et applications, Ecole normale supérieure, 45 rue d'Ulm, 75005 Paris, France, gilles.stoltz@ens.fr

hal-00124679, version 1 - 15 Jan 2007

Abstract. We propose simple randomized strategies for sequential prediction under imperfect monitoring, that is, when the forecaster does not have access to the past outcomes but rather to a feedback signal. The proposed strategies are consistent in the sense that they achieve, asymptotically, the best possible average reward. It was Rustichini [11] who first proved the existence of such consistent predictors. The forecasters presented here offer the first constructive proof of consistency. Moreover, the proposed algorithms are computationally efficient. We also establish upper bounds for the rates of convergence. In the case of deterministic feedback, these rates are optimal up to logarithmic terms.

1 Introduction

Sequential prediction of arbitrary (or "individual") sequences has received a lot of attention in learning theory, game theory, and information theory; see [3] for an extensive review. In this paper we focus on the problem of prediction of sequences taking values in a finite alphabet when the forecaster has limited information about the past outcomes of the sequence.

The randomized prediction problem is described as follows. Consider a sequential decision problem where a forecaster has to predict the environment's action. At each round, the forecaster chooses an action $i \in \{1, \ldots, N\}$, and the environment chooses an action $j \in \{1, \ldots, M\}$ (which we also call an "outcome"). The forecaster's reward $r(i, j)$ is the value of a reward function $r : \{1, \ldots, N\} \times \{1, \ldots, M\} \to [0, 1]$. Now suppose that, at the t-th round, the forecaster chooses a probability distribution $\boldsymbol{p}_t = (p_{1,t}, \ldots, p_{N,t})$ over the set of actions, and plays action i with probability $p_{i,t}$. We denote the forecaster's action at time t by I_t. If the environment chooses action $J_t \in \{1, \ldots, M\}$, the reward of the forecaster is $r(I_t, J_t)$. The prediction problem is defined as follows:

RANDOMIZED PREDICTION WITH PERFECT MONITORING

Parameters: number N of actions, cardinality M of outcome space, reward function r, number n of game rounds.

For each round $t = 1, 2, \ldots, n$,

(1) the environment chooses the next outcome J_t;
(2) the forecaster chooses $\boldsymbol{p}_t$ and determines the random action I_t, distributed according to $\boldsymbol{p}_t$;
(3) the environment reveals J_t;
(4) the forecaster receives a reward $r(I_t, y_t)$.

The goal of the forecaster is to minimize the average regret

$$\max_{i=1,\ldots,N} \frac{1}{n} \sum_{t=1}^{n} r(i, J_t) - \frac{1}{n} \sum_{t=1}^{n} r(I_t, J_t),$$

that is, the realized difference between the cumulative reward of the best strategy $i \in \{1, \ldots, N\}$, in hindsight, and the reward of the forecaster. Denoting by $r(\boldsymbol{p}, j) = \sum_{i=1}^{N} p_i r(i, j)$ the linear extension of the reward function r, the Hoeffding-Azuma inequality for sums of bounded martingale differences (see [8], [1]), implies that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\frac{1}{n} \sum_{t=1}^{n} r(I_t, J_t) \geq \frac{1}{n} \sum_{t=1}^{n} r(\boldsymbol{p}_t, J_t) - \sqrt{\frac{1}{2n} \ln \frac{1}{\delta}} \ ,$$

so it suffices to study the average expected reward $(1/n) \sum_{t=1}^{n} r(\boldsymbol{p}_t, J_t)$. Hannan [7] and Blackwell [2] were the first to show the existence of a forecaster whose regret is $o(1)$ for all possible behaviors of the opponent. Here we mention one of the simplest, yet quite powerful forecasting strategies, the *exponentially weighted average* forecaster. This forecaster selects, at time t, an action I_t according to the probabilities

$$p_{i,t} = \frac{\exp\left(\eta \sum_{s=1}^{t-1} r(i, J_s)\right)}{\sum_{k=1}^{N} \exp\left(\eta \sum_{s=1}^{t-1} r(k, J_s)\right)} \quad i = 1, \ldots, N,$$

where $\eta > 0$ is a parameter of the forecaster. One of the basic well-known results in the theory of prediction of individual sequences states that the regret of the exponentially weighted average forecaster is bounded as

$$\max_{i=1,\ldots,N} \frac{1}{n} \sum_{t=1}^{n} r(i, J_t) - \frac{1}{n} \sum_{t=1}^{n} r(\boldsymbol{p}_t, J_t) \leq \frac{\ln N}{n\eta} + \frac{\eta}{8} \ . \tag{1}$$

With the choice $\eta = \sqrt{8 \ln N / n}$ the upper bound becomes $\sqrt{\ln N / (2n)}$. Different versions of this result have been proved by several authors; see [3] for a review.

In this paper we are concerned with problems in which the forecaster does not have access to the outcomes J_t. The information available to the forecaster at each round is called the *feedback*. These feedbacks may depend on the outcomes J_t only or on the action–outcome pairs (I_t, J_t) and may be deterministic or drawn at random. In the simplest case when the feedback is deterministic, the information available to the forecaster is $s_t = h(I_t, J_t)$, given by a fixed (and known) deterministic feedback function $h : \{1, \ldots, N\} \times \{1, \ldots, M\} \to \mathcal{S}$ where $\mathcal{S}$ is the finite set of signals. In the most general case, the feedback is governed by a random feedback function of the form $H : \{1, \ldots, N\} \times \{1, \ldots, M\} \to \mathcal{P}(\mathcal{S})$ where $\mathcal{P}(\mathcal{S})$ is the set of probability distributions over the signals. The received feedback s_t is then drawn at random according to the probability distribution $H(I_t, J_t)$ by using an external independent randomization.

A motivating example for such a prediction problem arises naturally in multi-access channels that are prevalent in both wired and wireless networks. In such networks, the communication medium is shared between multiple decision makers. It is often technically difficult to synchronize between the decision makers. Channel sharing protocols such as ALOHA and several variants of spread spectrum allow multiple agents to use the same channel (or channels that may interfere with each other) simultaneously. More specifically, consider a wireless system where multiple agents can choose in which channel to transmit data at any given time. The quality of each channel may be different and interference from other users using this channel (or other "close" channels) may affect the base-station reception. The transmitting agent may choose which channel to use and how much power to spend on every transmission. The agent has a tradeoff between the amount of power wasted on a transmission and the cost of having its message only partially received. The transmitting agent does not receive immediate feedback on how much data were received in the base station (even if feedback is received, it often happens on a much higher layer of the communication protocol). Instead, the transmitting agent can monitor the transmissions of the other agents. However, since the transmitting agent is physically far from the base-station and the other agents, the information about the channels chosen by other agents and the amount of power they used is imperfect. This naturally abstracts to an online learning problem with imperfect monitoring.

To make notation uniform throughout the paper, we identify a deterministic feedback function $h : \{1, \ldots, N\} \times \{1, \ldots, M\} \to \mathcal{S}$ with the random feedback function $H : \{1, \ldots, N\} \times \{1, \ldots, M\} \to \mathcal{P}(\mathcal{S})$ which, to each pair (i, j), assigns $\delta_{h(i,j)}$ where δ_s is the probability distribution over the set of signals $\mathcal{S}$ concentrated on the single element $s \in \mathcal{S}$.

We will see that the prediction problem becomes significantly simpler in the special case when the feedback distribution depends only on the outcome, that is, when for all $i = 1, \ldots, N$, $H(i, \cdot)$ is constant. In other words, H depends on the outcome J_t but not on the forecaster's action I_t. To simplify notation in this case, we write $H(J_t) = H(I_t, J_t)$ for the feedback at time t ($h(J_t) = h(I_t, J_t)$ in case of deterministic feedback).

The sequential prediction problem under imperfect monitoring is formalized in Figure 1.

RANDOMIZED PREDICTION UNDER IMPERFECT MONITORING

Parameters: number $N$of actions, number M of outcomes, reward function r, random feedback function H, number n of rounds.

For each round $t = 1, 2 \ldots, n$,

1. the environment chooses the next outcome $J_t \in \{1, \ldots, M\}$ without revealing it;
2. the forecaster chooses a probability distribution $\boldsymbol{p}_t$ over the set of N actions and draws an action $I_t \in \{1, \ldots, N\}$ according to this distribution;
3. the forecaster receives reward $r(I_t, J_t)$ and each action i gets reward $r(i, J_t)$, where none of these values is revealed to the forecaster;
4. a feedback s_t drawn at random according to $H(I_t, J_t)$ is revealed to the forecaster.

Fig. 1. The game of randomized prediction under imperfect monitoring

Next we describe a reasonable goal for the forecaster and define the appropriate notion of consistency. To this end, we introduce some notation. If $\boldsymbol{p} = (p_1, \ldots, p_N)$ and $\boldsymbol{q} = (q_1, \ldots, q_M)$ are probability distributions over $\{1, \ldots, N\}$ and $\{1, \ldots, M\}$, respectively, then, with a slight abuse of notation, we write

$$r(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{N} \sum_{j=1}^{M} p_i q_j r(i, j)$$

for the linear extension of the reward function r. We also extend linearly the random feedback function in its second argument: for a probability distribution $\boldsymbol{q} = (q_1, \ldots, q_M)$ over $\{1, \ldots, M\}$, define the vector in $\mathbb{R}^{|\mathcal{S}|}$

$$H(i, \boldsymbol{q}) = \sum_{j=1}^{M} q_j H(i, j) \ , \qquad i = 1, \ldots, N.$$

Denote by $\mathcal{F}$ the convex set of all the N-vectors $H(\cdot, \boldsymbol{q}) = (H(1, \boldsymbol{q}), \ldots, H(N, \boldsymbol{q}))$ of probability distributions obtained this way when $\boldsymbol{q}$ varies. ($\mathcal{F} \subset \mathbb{R}^{|\mathcal{S}|N}$ is the set of feasible distributions over the signals). In the case where the feedback only depends on the outcome, all components of this vector are equal and we denote their common value by $H(\boldsymbol{q})$. We note that in the general case, the set $\mathcal{F}$ is the convex hull of the M vectors $H(\cdot, j)$. Therefore, performing a Euclidean projection on $\mathcal{F}$ can be done efficiently using quadratic programming.

To each probability distribution $\boldsymbol{p}$ over $\{1,\dots,N\}$ and probability distribution $\Delta \in \mathcal{F}$, we may assign the quantity

$$\rho(\boldsymbol{p},\Delta) = \min_{\boldsymbol{q}:H(\cdot,\boldsymbol{q})=\Delta} r(\boldsymbol{p},\boldsymbol{q}) \ .$$

Note that $\rho \in [0,1]$, and ρ is concave in $\boldsymbol{p}$ and convex in Δ.

To define the goal of the forecaster, let $\overline{\boldsymbol{q}}_n$ denote the empirical distribution of the outcomes $J_1,\dots,J_n$ up to round n. This distribution may be unknown to the forecaster since the forecaster observes the signals rather than the outcomes. The best the forecaster can hope for is an average reward close to $\max_{\boldsymbol{p}} \rho(\boldsymbol{p}, H(\cdot,\overline{\boldsymbol{q}}_n))$. Indeed, even if $H(\cdot,\overline{\boldsymbol{q}}_n)$ was known beforehand, the maximal expected reward for the forecaster would be $\max_{\boldsymbol{p}} \rho(\boldsymbol{p}, H(\cdot,\overline{\boldsymbol{q}}_n))$, simply because without any additional information the forecaster cannot hope to do better than against the worst element which is equivalent to $\boldsymbol{q}$ as far as the signals are concerned.

Based on this argument, the (per-round) regret R_n is defined as the averaged difference between the obtained cumulative reward and the target quantity described above, that is,

$$R_n = \max_{\boldsymbol{p}} \rho(\boldsymbol{p}, H(\cdot,\overline{\boldsymbol{q}}_n)) - \frac{1}{n}\sum_{t=1}^{n} r(I_t, J_t) \ .$$

Rustichini [11] proves the existence of a forecasting strategy whose per-round regret is guaranteed to satisfy $\limsup_{n\to\infty} R_n \le 0$ with probability one, for all possible imperfect monitoring problems. However, Rustichini's proof is not constructive and it seems unlikely that his proof method can give rise to computationally efficient prediction algorithms.

Several partial solutions had been proposed so far. Piccolboni and Schindelhauer [10] and Cesa-Bianchi, Lugosi, and Stoltz [4] study the case when $\max_{\boldsymbol{p}} \rho(\boldsymbol{p}, H(\cdot,\overline{\boldsymbol{q}}_n)) = \max_{i=1,\dots,N} r(i,\overline{\boldsymbol{q}}_n) = \max_{i=1,\dots,N} (1/n)\sum_{t=1}^{n} r(i,J_t)$. In this case strategies with a vanishing per-round regret are called *Hannan consistent*. This case turns out to be considerably simpler than the general case and computationally tractable explicit algorithms have been derived. Also, it is shown in [4] that in this case it is possible to construct strategies whose regret decreases as $O_p(n^{-1/3})$. The general case was considered by Mannor and Shimkin [9] who construct an algorithm with vanishing regret in the case when the feedback depends only on the outcome.

In this paper we construct simple and computationally efficient strategies whose regret vanishes with probability one. In Section 2 we consider the simplest special case when the actions of the forecaster do not influence the feedback which is, moreover, deterministic. This case is basically as easy as the full information case and we obtain a regret bound of the order of $n^{-1/2}$ (with high probability). In Section 3 we study random feedback but still with the restriction that it is only determined by the outcome. Here we are able to obtain a regret of the order of $n^{-1/4}\sqrt{\log n}$. The most general case is dealt with in Section 4. The forecaster introduced there has a regret of the order of $n^{-1/5}\sqrt{\log n}$. Finally, in Section 5 we show that this may be improved to $n^{-1/3}$ in the case of deterministic feedback, which is known to be optimal (see [4]).

2 Deterministic feedback only depends on outcome

We start with the simplest case when the feedback signal is deterministic and it does not depend on the action I_t of the forecaster. In other words, after making the prediction at time t, the forecaster observes $h(J_t)$.

In this case, we group the outcomes according to the deterministic feedback they are associated to. Each signal s is uniquely associated to a group of outcomes. This situation is very similar to the case of full monitoring except that rewards are measured by ρ and not by r. This does not pose a problem since r is lower bounded by ρ in the sense that for all $\boldsymbol{p}$ and j,

$$r(\boldsymbol{p}, j) \geq \rho(\boldsymbol{p}, \delta_{h(j)}) \ .$$

We introduce a forecaster that resembles the gradient-based strategies described, for example, in Cesa-Bianchi and Lugosi [3, Section 2.5]. The forecaster uses any sub-gradient of $\rho(\cdot, \delta_{h(J_t)})$ at time t. (Recall that if f is a concave function defined over a convex subset of $\mathbb{R}^d$, any vector $\boldsymbol{b}(\boldsymbol{x}) \in \mathbb{R}^d$ is a sub-gradient of f at $\boldsymbol{x}$ if $f(\boldsymbol{y}) - f(\boldsymbol{x}) \leq \boldsymbol{b}(\boldsymbol{x}) \cdot (\boldsymbol{y} - \boldsymbol{x})$ for all $\boldsymbol{y}$ in the domain of f. Sub-gradients always exist in the interior of the domain of a concave function. Here, in view of the exponentially weighted update rules, we only evaluate them in the interior of the simplex.) The forecaster requires a tuning parameter $\eta > 0$. The i-th component of $\boldsymbol{p}_t$ is

$$p_{i,t} = \frac{e^{\eta \sum_{s=1}^{t-1} \left(\widetilde{r}(\boldsymbol{p}_s, \delta_{h(J_s)})\right)_i}}{\sum_{j=1}^{N} e^{\eta \sum_{s=1}^{t-1} \left(\widetilde{r}(\boldsymbol{p}_s, \delta_{h(J_s)})\right)_j}},$$

where $\left(\widetilde{r}(\boldsymbol{p}_s, \delta_{h(J_s)})\right)_i$ is the i-th component of any sub-gradient $\widetilde{r}(\boldsymbol{p}_s, \delta_{h(J_s)}) \in \nabla\rho(\boldsymbol{p}_s, \delta_{h(J_s)})$ of the concave function $f(\cdot) = \rho(\cdot, \delta_{h(J_s)})$.

The computation of a sub-gradient is trivial whenever $\rho(\boldsymbol{p}_s, \delta_{h(J_s)})$ is differentiable because it is then locally linear and the gradient equals the column of the reward matrix corresponding to the outcome y_s for which $r(\boldsymbol{p}_s, y_s) = \rho(\boldsymbol{p}_s, \delta_{h(J_s)})$. Note that $\rho(\cdot, \delta_{h(J_s)})$ is differentiable exactly at those points at which it is locally linear. Since it is concave, the Lebesgue measure of the set where it is non-differentiable equals zero. To avoid such values, one may add a small random perturbation to $\boldsymbol{p}_t$ or just calculate a sub-gradient using the simplex method. Note that the components of the sub-gradients are always bounded by a constant that depends on the game parameters. This is the case since $\rho(\cdot, \delta_{h(J_s)})$ is concave and continuous on a compact set and is therefore Lipschitz leading to a bounded sub-gradient. Let K denote a constant such that $\sup_{\boldsymbol{p}} \max_j \|\widetilde{r}(\boldsymbol{p}, \delta_{h(j)})\|_\infty \leq K$. This constant depends on the specific parameters of the game. The regret is bounded as follows. Note that the following bound (and the considered forecaster) coincide with those of (1) in case of perfect monitoring. (In that case, $\rho(\cdot, \delta_{h(j)}) = r(\cdot, j)$, the subgradients are given by r, and therefore, are bounded between 0 and 1.)

Proposition 1. *For all $\eta > 0$, for all strategies of the environment, for all $\delta > 0$, the above strategy of the forecaster ensures that, with probability at least*

$1-\delta$,

$$R_n \le \frac{\ln N}{\eta n} + \frac{K^2\eta}{2} + \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}\ .$$

In particular, choosing $\eta \sim \sqrt{(\ln N)/n}$ *yields* $R_n = O(n^{-1/2}\sqrt{\ln(N/\delta)})$.

Proof. Note that since the feedback is deterministic, $H(\overline{\boldsymbol{q}}_n)$ takes the simple form $H(\overline{\boldsymbol{q}}_n) = \frac{1}{n}\sum_{t=1}^n \delta_{h(J_t)}$. Now, for any $\boldsymbol{p}$,

$$\begin{aligned}
&n\rho(\boldsymbol{p}, H(\overline{\boldsymbol{q}}_n)) - \sum_{t=1}^n r(\boldsymbol{p}_t, J_t) \\
&\le n\rho(\boldsymbol{p}, H(\overline{\boldsymbol{q}}_n)) - \sum_{t=1}^n \rho(\boldsymbol{p}_t, \delta_{h(J_t)}) \quad \text{(by the lower bound on } r \text{ in terms of } \rho) \\
&\le \sum_{t=1}^n \big(\rho(\boldsymbol{p}, \delta_{h(J_t)}) - \rho(\boldsymbol{p}_t, \delta_{h(J_t)})\big) \quad \text{(by convexity of } \rho \text{ in the second argument)} \\
&\le \sum_{t=1}^n \widetilde{r}(\boldsymbol{p}_t, \delta_{h(J_t)}) \cdot (\boldsymbol{p} - \boldsymbol{p}_t) \quad \text{(by concavity of } \rho \text{ in the first argument)} \\
&\le \frac{\ln N}{\eta} + \frac{nK^2\eta}{2} \quad \text{(by (1), after proper rescaling),}
\end{aligned}$$

where at the last step we used the fact that the forecaster is just the exponentially weighted average predictor based on the rewards $(\widetilde{r}(\boldsymbol{p}_s, \delta_{h(J_s)}))_i$ and that all these reward vectors have components between $-K$ and K. The proof is concluded by the Hoeffding-Azuma inequality, which ensures that, with probability at least $1-\delta$,

$$\sum_{t=1}^n r(I_t, J_t) \ge \sum_{t=1}^n r(\boldsymbol{p}_t, J_t) - \sqrt{\frac{n}{2}\ln\frac{1}{\delta}}\ . \tag{2}$$

3 Random feedback only depends on outcome

Next we consider the case when the feedback does not depend on the forecaster's actions, but, at time t, the signal s_t is drawn at random according to the distribution $H(J_t)$. In this case the forecaster does not have a direct access to

$$H(\overline{\boldsymbol{q}}_n) = \frac{1}{n}\sum_{t=1}^n H(J_t)$$

anymore, but only observes the realizations s_t drawn at random according to $H(J_t)$. In order to overcome this problem, we group together several consecutive time rounds (m of them) and estimate the probability distributions according to which the signals have been drawn.

To this end, denote by Π the Euclidean projection onto $\mathcal{F}$ (since the feedback depends only on the outcome we may now view the set $\mathcal{F}$ of feasible distributions

Parameters: Integer $m \geq 1$, real number $\eta > 0$.
Initialization: $\boldsymbol{w}^0 = (1, \ldots, 1)$.

For each round $t = 1, 2, \ldots$

1. If $bm + 1 \leq t < (b+1)m$ for some integer b, choose the distribution $\boldsymbol{p}_t = \boldsymbol{p}^b$ given by
$$p_{k,t} = p_k^b = \frac{w_k^b}{\sum_{j=1}^N w_j^b}$$
and draw an action I_t from $\{1, \ldots, N\}$ according to it;
2. if $t = (b+1)m$ for some integer b, perform the update
$$w_k^{b+1} = w_k^b \, e^{\eta \left(\widetilde{r}(\boldsymbol{p}^b, \widehat{\Delta}^b)\right)_k} \qquad \text{for each } k = 1, \ldots, N,$$
where for all Δ, $\widetilde{r}(\cdot, \Delta)$ is a sub-gradient of $\rho(\cdot, \Delta)$ and $\widehat{\Delta}^b$ is defined in (3).

Fig. 2. The forecaster for random feedback depending only on outcome.

over the signals as a subset of $\mathcal{P}(\mathcal{S})$, the latter being identified with a subset of $\mathbb{R}^{|\mathcal{S}|}$ in a natural way). Let m, $1 \leq m \leq n$, be a parameter of the algorithm. For $b = 0, 1, \ldots$, we denote

$$\widehat{\Delta}^b = \Pi\left(\frac{1}{m} \sum_{t=bm+1}^{(b+1)m} \delta_{s_t}\right) . \tag{3}$$

For the sake of the analysis, we also introduce

$$\Delta^b = \frac{1}{m} \sum_{t=bm+1}^{(b+1)m} H(J_t) .$$

The proposed strategy is described in Figure 2. Observe that the practical implementation of the forecaster only requires the computation of (sub)gradients and of ℓ^2 projections, which can be done in polytime. The next theorem bounds the regret of the strategy which is of the order of $n^{-1/4}\sqrt{\log n}$. The price we pay for having to estimate the distribution is thus a deteriorated rate of convergence (from the $O(n^{-1/2})$ obtained in the case of deterministic feedback). We do not know whether this rate can be improved significantly as we do not know of any nontrivial lower bound in this case.

Theorem 1. *For all integers $m \geq 1$, for all $\eta > 0$, and for all $\delta > 0$, the regret against any strategy of the environment is bounded, with probability at least $1 - (n/m + 1)\delta$, by*

$$R_n \leq 2\sqrt{2}\, L \, \frac{1}{\sqrt{m}} \sqrt{\ln \frac{2}{\delta}} + \frac{m \ln N}{n\eta} + \frac{K^2 \eta}{2} + \frac{m}{n} + \sqrt{\frac{1}{2n} \ln \frac{1}{\delta}} \, ,$$

where K, L are constants which depend only on the parameters of the game. The choices $m = \lceil \sqrt{n} \rceil$ and $\eta \sim \sqrt{(m \ln N)/n}$ imply $R_n = O(n^{-1/4}\sqrt{\ln(nN/\delta)})$ with probability of at least $1 - \delta$.

Proof. We start by grouping time rounds m by m. For simplicity, we assume that $n = (B+1)m$ for some integer B (this accounts for the m/n term in the bound). For all $\boldsymbol{p}$,

$$n\,\rho(\boldsymbol{p}, H(\overline{\boldsymbol{q}}_n)) - \sum_{t=1}^{n} r(\boldsymbol{p}_t, J_t) \le \sum_{b=0}^{B} \left(m\,\rho\left(\boldsymbol{p}, \Delta^b\right) - m\,r\left(\boldsymbol{p}^b, \frac{1}{m}\sum_{t=bm+1}^{(b+1)m} \delta_{J_t}\right)\right)$$
$$\le m \sum_{b=0}^{B} \left(\rho\left(\boldsymbol{p}, \Delta^b\right) - \rho\left(\boldsymbol{p}^b, \Delta^b\right)\right) \ ,$$

where we used the definition of the algorithm, convexity of ρ in its second argument, and finally, the definition of ρ as a minimum. We proceed by estimating Δ^b by $\widehat{\Delta}^b$. By a version of the Hoeffding-Azuma inequality in Hilbert spaces proved by Chen and White [5, Lemma 3.2], and since the ℓ^2 projection can only help, for all b, with probability at least $1 - \delta$,

$$\left\| \Delta^b - \widehat{\Delta}^b \right\|_2 \le \sqrt{\frac{2\ln\frac{2}{\delta}}{m}} \ .$$

By Proposition 2, ρ is uniformly Lipschitz in its second argument (with constant L), and therefore we may further bound as follows. With probability $1-(B+1)\delta$,

$$m \sum_{b=0}^{B} \left(\rho\left(\boldsymbol{p}, \Delta^b\right) - \rho\left(\boldsymbol{p}^b, \Delta^b\right)\right) \le m \sum_{b=0}^{B} \left(\rho\left(\boldsymbol{p}, \widehat{\Delta}^b\right) - \rho\left(\boldsymbol{p}^b, \widehat{\Delta}^b\right)\right) + 2\,L(B+1)\sqrt{2m\ln\frac{2}{\delta}} \ .$$

The term containing $(B+1)\sqrt{m} = n/\sqrt{m}$ is the first term in the upper bound. The remaining part is bounded by using the same slope inequality argument as in the previous section (recall that $\widetilde{r}$ denotes a sub-gradient),

$$m \sum_{b=0}^{B} \left(\rho\left(\boldsymbol{p}, \widehat{\Delta}^b\right) - \rho\left(\boldsymbol{p}^b, \widehat{\Delta}^b\right)\right) \le m \sum_{b=0}^{B} \widetilde{r}\left(\boldsymbol{p}^b, \widehat{\Delta}^b\right) \cdot \left(\boldsymbol{p} - \boldsymbol{p}^b\right)$$
$$\le m\left(\frac{\ln N}{\eta} + \frac{(B+1)K^2\eta}{2}\right) = \frac{m\ln N}{\eta} + \frac{nK^2\eta}{2}$$

where we used Theorem 1 and the boundedness of the function $\widetilde{r}$ between $-K$ and K. The proof is concluded by the Hoeffding-Azuma inequality which, as in (2), gives the final term in the bound. The union bound indicates that the obtained bound holds with probability at least $1 - (B+2)\delta \ge 1 - (n/m+1)\delta$.

4 Random feedback depends on action–outcome pair

We now turn to the most general case, where the feedback is random and depends on the action–outcome pairs (I_t, J_t). The key is, again, to exhibit efficient estimators of the (unobserved) $H(\cdot, \overline{\boldsymbol{q}}_n)$.

Denote by $\mathit{\Pi}$ the projection, in the Euclidian distance, onto $\mathcal{F}$ (where $\mathcal{F}$, as a subset of $(\mathcal{P}(\mathcal{S}))^N$, is identified with a subset of $\mathbb{R}^{|\mathcal{S}|\,N}$). For $b = 0, 1, \ldots$, denote

$$\widehat{\mathit{\Delta}}^b = \mathit{\Pi}\left(\frac{1}{m}\sum_{t=bm+1}^{(b+1)m}\left[\widehat{h}_{i,t}\right]_{i=1,\ldots,N}\right) \tag{4}$$

where the distribution $H(i, J_t)$ of the random signal s_t received by action i at round t is estimated by

$$\widehat{h}_{i,t} = \frac{\delta_{s_t}}{p_{i,t}}\mathbb{1}_{I_t=i}\ .$$

We prove that the $\widehat{h}_{i,t}$ are conditionally unbiased estimators. Denote by $\mathbb{E}_t$ the conditional expectation with respect to the information available to the forecaster at the beginning of round t. This conditioning fixes the values of $\boldsymbol{p}_t$ and J_t. Thus,

$$\mathbb{E}_t\left[\widehat{h}_{i,t}\right] = \frac{1}{p_{i,t}}\,\mathbb{E}_t\left[\delta_{s_t}\,\mathbb{1}_{I_t=i}\right] = \frac{1}{p_{i,t}}\,\mathbb{E}_t\left[H(I_t, J_t)\mathbb{1}_{I_t=i}\right] = \frac{1}{p_{i,t}}H(i, J_t)p_{i,t} = H(i, J_t)\,.$$

For the sake of the analysis, introduce

$$\mathit{\Delta}^b = \frac{1}{m}\sum_{t=bm+1}^{(b+1)m} H(\cdot, J_t)\ .$$

The proposed forecasting strategy is sketched in Figure 3. Here again, the practical implementation of the forecaster only requires the computation of (sub)gradients and of ℓ^2 projections, which can be done efficiently. The next theorem states that the regret in this most general case is at most of the order of $n^{-1/5}\sqrt{\log n}$. Again, we don't know whether this bound can be improved significantly.

Theorem 2. *For all integers $m \geq 1$, for all $\eta > 0$, $\gamma \in (0,1)$, and $\delta > 0$, the regret against any strategy of the environment is bounded, with probability at least $1 - (n/m + 1)\delta$, as*

$$R_n \leq L\,N\sqrt{\frac{2\,|S|}{\gamma m}\ln\frac{2N\,|S|}{\delta}} + L\frac{N^{3/2}\sqrt{|S|}}{3\gamma m}\ln\frac{2N\,|S|}{\delta} + \frac{m\ln N}{n\eta} + \frac{K^2\eta}{2} + \gamma + \frac{m}{n} + \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}\ ,$$

where L and K are constants which depend on the parameters of the game. The choices $m = \lceil n^{3/5}\rceil$, $\eta \sim \sqrt{(m\ln N)/n}$, and $\gamma \sim n^{-1/5}$ ensure that, with probability at least $1-\delta$, $R_n = O\left(n^{-1/5}N\sqrt{\ln\frac{Nn}{\delta}} + n^{-2/5}N^{3/2}\ln\frac{Nn}{\delta}\right)$

Parameters: Integer $m \geq 1$, real numbers $\eta, \gamma > 0$.
Initialization: $\boldsymbol{w}^0 = (1, \ldots, 1)$.

For each round $t = 1, 2, \ldots$

1. if $bm + 1 \leq t < (b+1)m$ for some integer b, choose the distribution $\boldsymbol{p}_t = \boldsymbol{p}^b = (1-\gamma)\widetilde{\boldsymbol{p}}^b + \gamma \boldsymbol{u}$, where $\widetilde{\boldsymbol{p}}^b$ is defined component-wise as

$$\widetilde{p}_k^b = \frac{w_k^b}{\sum_{j=1}^N w_j^b}$$

   and $\boldsymbol{u}$ denotes the uniform distribution, $\boldsymbol{u} = (1/N, \ldots, 1/N)$;
2. draw an action I_t from $\{1, \ldots, N\}$ according to it;
3. if $t = (b+1)m$ for some integer b, perform the update

$$w_k^{b+1} = w_k^b \, e^{\eta \left(\widetilde{r}\left(\boldsymbol{p}^b, \widehat{\Delta}^b\right)\right)_k} \qquad \text{for each } k = 1, \ldots, N,$$

   where for all $\Delta \in \mathcal{F}$, $\widetilde{r}(\cdot, \Delta)$ is a sub-gradient of $\rho(\cdot, \Delta)$ and $\widehat{\Delta}^b$ is defined in (4).

Fig. 3. The forecaster for random feedback depending on action–outcome pair.

Proof. The proof is similar to the one of Theorem 1. A difference is that we bound the accuracy of the estimation of the Δ^b via a martingale analog of Bernstein's inequality due to Freedman [6] rather than the Hoeffding-Azuma inequality. Also, the mixing with the uniform distribution of Step 1 needs to be handled.

We start by grouping time rounds m by m. Assume, for simplicity, that $n = (B+1)m$ for some integer B (this accounts for the m/n term in the bound). As before, we get that, for all $\boldsymbol{p}$,

$$n\,\rho(\boldsymbol{p}, H(\cdot, \overline{\boldsymbol{q}}_n)) - \sum_{t=1}^{n} r(\boldsymbol{p}_t, J_t) \leq m \sum_{b=0}^{B} \left(\rho\left(\boldsymbol{p}, \Delta^b\right) - \rho\left(\boldsymbol{p}^b, \Delta^b\right)\right) \tag{5}$$

and proceed by estimating Δ^b by $\widehat{\Delta}^b$. Freedman's inequality [6] (see, also, [4, Lemma A.1]) implies that for all $b = 0, 1, \ldots, B$, $i = 1, \ldots, N$, $s \in \mathcal{S}$, and $\delta > 0$,

$$\left| \Delta_i^b(s) - \frac{1}{m} \sum_{t=bm+1}^{(b+1)m} \widehat{h}_{i,t}(s) \right| \leq \sqrt{2 \frac{N}{\gamma m} \ln \frac{2}{\delta}} + \frac{1}{3} \frac{N}{\gamma m} \ln \frac{2}{\delta}$$

where $\widehat{h}_{i,t}(s)$ is the probability mass put on s by $\widehat{h}_{i,t}$ and $\Delta_i^b(s)$ is the i-th component of Δ^b. This is because the sums of the conditional variances are bounded as

$$\sum_{t=bm+1}^{(b+1)m} \mathrm{Var}_t \left(\frac{\mathbb{1}_{I_t = i, s_t = s}}{p_{i,t}} \right) \leq \sum_{t=bm+1}^{(b+1)m} \frac{1}{p_{i,t}} \leq \frac{mN}{\gamma} \,.$$

Summing (since the ℓ^2 projection can only help), the union bound shows that for all b, with probability at least $1-\delta$,

$$\left\| \Delta^b - \widehat{\Delta}^b \right\|_2 \leq d \stackrel{\text{def}}{=} \sqrt{N\,|S|} \left(\sqrt{2\frac{N}{\gamma m} \ln \frac{2N\,|S|}{\delta}} + \frac{1}{3}\frac{N}{\gamma m} \ln \frac{2N\,|S|}{\delta} \right) .$$

By using uniform Lipschitzness of ρ in its second argument (with constant L; see Proposition 2), we may further bound (5) with probability $1-(B+1)\delta$ by

$$\begin{aligned} m \sum_{b=0}^{B} \left(\rho\left(\boldsymbol{p}, \Delta^b\right) - \rho\left(\boldsymbol{p}^b, \Delta^b\right) \right) &\leq m \sum_{b=0}^{B} \left(\rho\left(\boldsymbol{p}, \widehat{\Delta}^b\right) - \rho\left(\boldsymbol{p}^b, \widehat{\Delta}^b\right) + L\,d \right) \\ &= m \sum_{b=0}^{B} \left(\rho\left(\boldsymbol{p}, \widehat{\Delta}^b\right) - \rho\left(\boldsymbol{p}^b, \widehat{\Delta}^b\right) \right) + m(B+1) L\,d\ . \end{aligned}$$

The terms $m(B+1)L\,d = nL\,d$ are the first two terms in the upper bound of the theorem. The remaining part is bounded by using the same slope inequality argument as in the previous section (recall that $\widetilde{r}$ denotes a sub-gradient bounded between $-K$ and K):

$$m \sum_{b=0}^{B} \left(\rho\left(\boldsymbol{p}, \widehat{\Delta}^b\right) - \rho\left(\boldsymbol{p}^b, \widehat{\Delta}^b\right) \right) \leq m \sum_{b=0}^{B} \widetilde{r}\left(\boldsymbol{p}^b, \widehat{\Delta}^b\right) \cdot \left(\boldsymbol{p} - \boldsymbol{p}^b\right)\ .$$

Finally, we deal with the mixing with the uniform distribution:

$$\begin{aligned} m \sum_{b=0}^{B} \widetilde{r}\left(\boldsymbol{p}^b, \widehat{\Delta}^b\right) \cdot \left(\boldsymbol{p} - \boldsymbol{p}^b\right) &\leq (1-\gamma) m \sum_{b=0}^{B} \widetilde{r}\left(\boldsymbol{p}^b, \widehat{\Delta}^b\right) \cdot \left(\boldsymbol{p} - \widetilde{\boldsymbol{p}}^b\right) + \gamma m (B+1) \\ &\qquad \text{(since, by definition, } \boldsymbol{p}^b = (1-\gamma)\widetilde{\boldsymbol{p}}^b + \gamma \boldsymbol{u}) \\ &\leq (1-\gamma) m \left(\frac{\ln N}{\eta} + \frac{(B+1)K^2\eta}{2} \right) + \gamma m (B+1) \\ &\qquad \text{(by (1))} \\ &\leq \frac{m \ln N}{\eta} + \frac{nK^2\eta}{2} + \gamma n\ . \end{aligned}$$

The proof is concluded by the Hoeffding-Azuma inequality which, as in (2), gives the final term in the bound. The union bound indicates that the obtained bound hold with probability at least $1-(B+2)\delta \geq 1-(n/m+1)\delta$.

5 Deterministic feedback depends on action–outcome pair

In this last section we explain how in the case of deterministic feedback the forecaster of the previous section can be modified so that the order of magnitude of

the per-round regret improves to $n^{-1/3}$. This relies on the linearity of ρ in its second argument. In the case of random feedback, ρ may not be linear which required grouping rounds of size m. If the feedback is deterministic, such grouping is not needed and the $n^{-1/3}$ rate is obtained as a trade-off between an exploration term (γ) and the cost payed for estimating the feedbacks ($\sqrt{1/(\gamma n)}$). This rate of convergence has been shown to be optimal in [4] even in the Hannan consistent case. The key property is summarized in the next technical lemma whose proof is omitted for the lack of space.

Lemma 1. *For every fixed $\boldsymbol{p}$, the function $\rho(\boldsymbol{p}, \cdot)$ is linear on $\mathcal{F}$.*

Next we describe the modified forecaster. Denote by $\mathcal{H}$ the vector space generated by $\mathcal{F} \subset \mathbb{R}^{|\mathcal{S}|N}$ and Π the linear operator which projects any element of $\mathbb{R}^{|\mathcal{S}|N}$ onto $\mathcal{H}$. Since the $\rho(\boldsymbol{p}, \cdot)$ are linear on $\mathcal{F}$, we may extend them linearly to $\mathcal{H}$ (and with a slight abuse of notation we write ρ for the extension). As a consequence, the functions $\rho(\boldsymbol{p}, \Pi(\cdot))$ are linear defined on $\mathbb{R}^{|\mathcal{S}|N}$ and coincide with the original definition on $\mathcal{F}$. We denote by $\widetilde{r}$ a sub-gradient (i.e., for all $\Delta \in \mathbb{R}^{|S|N}$, $\widetilde{r}(\cdot, \Delta)$ is a sub-gradient of $\rho(\cdot, \Pi(\Delta))$).

The sub-gradients are evaluated at the following points. (Recall that since the feedback is deterministic, $s_t = h(I_t, J_t)$.) For $t = 1, 2, \ldots$, let

$$\widehat{h}_t = \left[\widehat{h}_{i,t}\right]_{i=1,\ldots,N} = \left[\frac{\delta_{s_t}}{p_{i,t}} \mathbb{1}_{I_t=i}\right]_{i=1,\ldots,N} . \tag{6}$$

The $\widehat{h}_{i,t}$ estimate the feedbacks $H(i, J_t) = \delta_{h(i,J_t)}$ received by action i at round t. They are still conditionally unbiased estimators of the $h(i, J_t)$, and so is $\widehat{h}_t$ for $H(\cdot, J_t)$. The proposed forecaster is defined in Figure 4 and the regret bound is established in Theorem 3.

Theorem 3. *There exists a constant C only depending on r and h such that for all $\delta > 0$, $\gamma \in (0, 1)$, and $\eta > 0$, the regret against any strategy of the environment is bounded, with probability at least $1 - \delta$, as*

$$R_n \leq 2NC\sqrt{\frac{2}{n\gamma} \ln \frac{2}{\delta}} + \frac{NC}{3\gamma n} \ln \frac{2}{\delta} + \frac{\ln N}{\eta n} + \frac{\eta K^2}{2} + \gamma + \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}} .$$

The choice $\gamma \sim n^{-1/3} N^{2/3}$ and $\eta \sim \sqrt{(\ln N)/n}$ ensures that, with probability at least $1 - \delta$, $R_n = O\left(n^{-1/3} N^{2/3} \sqrt{\ln(1/\delta)}\right)$.

Proof. The proof is similar to the one of Theorem 2, except that we do not have to consider the grouping steps and that we do not apply the Hoeffding-Azuma inequality to the estimated feedbacks but to the estimated rewards. By the bound on r in terms of ρ and convexity (linearity) of ρ in its second argument,

$$n\, \rho(\boldsymbol{p}, H(\cdot, \overline{\boldsymbol{q}}_n)) - \sum_{t=1}^{n} r(\boldsymbol{p}_t, J_t) \leq \sum_{t=1}^{n} \left(\rho\left(\boldsymbol{p}, H(\cdot, J_t)\right) - \rho\left(\boldsymbol{p}_t, H(\cdot, J_t)\right)\right) .$$

Parameters: Real numbers $\eta, \gamma > 0$.
Initialization: $\boldsymbol{w}_1 = (1, \ldots, 1)$.

For each round $t = 1, 2, \ldots$

1. choose the distribution $\boldsymbol{p}_t = (1-\gamma)\widetilde{\boldsymbol{p}}_t + \gamma \boldsymbol{u}$, where $\widetilde{\boldsymbol{p}}_t$ is defined component-wise as
$$\widetilde{p}_{k,t} = \frac{w_{k,t}}{\sum_{j=1}^{N} w_{j,t}}$$
and $\boldsymbol{u}$ denotes the uniform distribution, $\boldsymbol{u} = (1/N, \ldots, 1/N)$; then draw an action I_t from $\{1, \ldots, N\}$ according to $\boldsymbol{p}_t$;
2. perform the update
$$w_{k,t+1} = w_{k,t}\, e^{\eta \left(\widetilde{r}(\boldsymbol{p}_t, \widehat{h}_t)\right)_k} \qquad \text{for each } k = 1, \ldots, N,$$
where $\mathit{\Pi}$ is the projection operator defined after the statement of Lemma 1, for all $\mathit{\Delta} \in \mathbb{R}^{|S|N}$, $\widetilde{r}(\cdot, \mathit{\Delta})$ is a sub-gradient of $\rho(\cdot, \mathit{\Pi}(\mathit{\Delta}))$, and $\widehat{h}_t$ is defined in (6).

Fig. 4. The forecaster for deterministic feedback depending on action–outcome pair.

Next we estimate

$$\rho\left(\boldsymbol{p}, H(\cdot, J_t)\right) - \rho\left(\boldsymbol{p}_t, H(\cdot, J_t)\right) \qquad \text{by} \qquad \rho\left(\boldsymbol{p}, \mathit{\Pi}\left(\widehat{h}_t\right)\right) - \rho\left(\boldsymbol{p}_t, \mathit{\Pi}\left(\widehat{h}_t\right)\right) .$$

By Freedman's inequality (see, again, [4, Lemma A.1]), since $\widehat{h}_t$ is a conditionally unbiased estimator of $H(\cdot, J_t)$ and all functions at hand are linear in their second argument, we get that, with probability at least $1 - \delta/2$,

$$\begin{aligned}
&\sum_{t=1}^{n} \left(\rho\left(\boldsymbol{p}, H(\cdot, J_t)\right) - \rho\left(\boldsymbol{p}_t, H(\cdot, J_t)\right)\right) \\
&= \sum_{t=1}^{n} \left(\rho\left(\boldsymbol{p}, \mathit{\Pi}\left(H(\cdot, J_t)\right)\right) - \rho\left(\boldsymbol{p}_t, \mathit{\Pi}\left(H(\cdot, J_t)\right)\right)\right) \\
&\le \sum_{t=1}^{n} \left(\rho\left(\boldsymbol{p}, \mathit{\Pi}\left(\widehat{h}_t\right)\right) - \rho\left(\boldsymbol{p}_t, \mathit{\Pi}\left(\widehat{h}_t\right)\right)\right) + 2NC\sqrt{2\frac{n}{\gamma}\ln\frac{2}{\delta}} + \frac{NC}{3\gamma}\ln\frac{2}{\delta}
\end{aligned}$$

where, denoting by $\boldsymbol{e}_i(\delta_{h(i,j)})$ the column vector whose i-th component is $\delta_{h(i,j)}$ and all other components equal 0,

$$C = \max_{i,j} \max_{\boldsymbol{p}} \rho\left(\boldsymbol{p}, \mathit{\Pi}\left[\boldsymbol{e}_i(\delta_{h(i,j)})\right]\right) < +\infty .$$

This is because for all t, the conditional variances are bounded as follows. For all $\boldsymbol{p}'$,

$$\mathbb{E}_t\left[\rho\left(\boldsymbol{p}', \mathit{\Pi}\left(\widehat{h}_t\right)\right)^2\right] = \sum_{i=1}^{N} p_{i,t}\,\rho\left(\boldsymbol{p}', \mathit{\Pi}\left[\boldsymbol{e}_i(\delta_{h(i,j)}/p_{i,t})\right]\right)^2$$
$$= \sum_{i=1}^{N} \frac{1}{p_{i,t}}\,\rho\left(\boldsymbol{p}', \mathit{\Pi}\left[\boldsymbol{e}_i(\delta_{h(i,j)}/p_{i,t})\right]\right)^2 \leq \sum_{i=1}^{N} \frac{C^2}{p_{i,t}} \leq \frac{C^2 N^2}{\gamma}\ .$$

The remaining part is bounded by using the same slope inequality argument as in the previous sections (recall that $\widetilde{r}$ denotes a sub-gradient in the first argument of $\rho(\cdot, \mathit{\Pi}(\cdot))$, bounded between $-K$ and K),

$$\sum_{t=1}^{n}\left(\rho\left(\boldsymbol{p}, \mathit{\Pi}\left(\widehat{h}_t\right)\right) - \rho\left(\boldsymbol{p}_t, \mathit{\Pi}\left(\widehat{h}_t\right)\right)\right) \leq \sum_{t=1}^{n} \widetilde{r}\left(\boldsymbol{p}_t, \widehat{h}_t\right) \cdot (\boldsymbol{p} - \boldsymbol{p}_t)\ .$$

Finally, we deal with the mixing with the uniform distribution:

$$\sum_{t=1}^{n} \widetilde{r}\left(\boldsymbol{p}, \widehat{h}_t\right) \cdot (\boldsymbol{p} - \boldsymbol{p}) \leq (1-\gamma) \sum_{t=1}^{n} \widetilde{r}\left(\boldsymbol{p}_t, \widehat{h}_t\right) \cdot (\boldsymbol{p} - \widetilde{\boldsymbol{p}}_t) + \gamma n$$
$$\text{(since by definition } \boldsymbol{p}_t = (1-\gamma)\widetilde{\boldsymbol{p}}_t + \gamma \boldsymbol{u})$$
$$\leq (1-\gamma)\left(\frac{\ln N}{\eta} + \frac{n\eta K^2}{2}\right) + \gamma n \quad \text{(by (1))}.$$

As before, the proof is concluded by the Hoeffding-Azuma inequality (2) and the union bound.

References

1. K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.
2. D. Blackwell. Controlled random walks. In *Proceedings of the International Congress of Mathematicians, 1954*, volume III, pages 336–338. North-Holland, 1956.
3. N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.
4. N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31, 562–580, 2006.
5. X. Chen and H. White. Laws of large numbers for Hilbert space-valued mixingales with applications. *Econometric Theory*, 12(2):284–304, 1996.
6. D.A. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3:100–118, 1975.
7. J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the theory of games*, 3:97–139, 1957.
8. W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

9. S. Mannor and N. Shimkin. On-line learning with imperfect monitoring. In *Proceedings of the 16th Annual Conference on Learning Theory*, pages 552–567. Springer, 2003.
10. 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.
11. A. Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29:224–243, 1999.

A Appendix

Proposition 2. *The function* $(\boldsymbol{p}, \Delta) \mapsto \rho(\boldsymbol{p}, \Delta)$ *is uniformly Lipschitz in its second argument.*

Proof. We consider the general case where the signal distribution depends on both the actions and outcomes. Accordingly, we can write $\rho(\boldsymbol{p}, \Delta)$ as the solution of the following linear program (we denote $\Delta = (\Delta_1, \ldots, \Delta_N) \in \mathcal{F} \subset \mathcal{P}(S)^N$):

$$\begin{aligned} \rho(\boldsymbol{p}, \Delta) = \min_{\boldsymbol{q}} \quad & \boldsymbol{q}^\top \boldsymbol{r} \\ \text{s.t.} \quad & H^k \boldsymbol{q} = \Delta_k \ , \quad k = 1, 2, \ldots, N \ , \\ & \boldsymbol{q}^\top \boldsymbol{e}_M = 1 \ , \\ & \boldsymbol{q} \geq 0 \ , \end{aligned}$$

where $\boldsymbol{r} = (r_j)_j = (\sum_{i=1}^N p_i r(i,j))_j$ is an M-dimensional vector, $\boldsymbol{e}_M$ is an M-dimensional vector of ones, and $H^k = H(k, \cdot)$ is the $S \times M$ matrix, whose entry (s, j) is the probability of observing signal s when action k is chosen and the outcome is j.

The program is feasible for every $\Delta \in \mathcal{F}$ so by the duality theorem:

$$\begin{aligned} \rho(\boldsymbol{p}, \Delta) = \max_{\boldsymbol{y} \in \mathbb{R}^{NS+1}} \quad & \boldsymbol{y}^\top [\Delta_1 \, \Delta_2 \, \ldots \, \Delta_N \, 1] \\ \text{s.t.} \quad & [H_1^1 \, H_1^2 \, \ldots \, H_1^N \, 1] \, \boldsymbol{y} = r_1 \ , \\ & \qquad \vdots \\ & [H_M^1 \, H_M^2 \, \ldots \, H_M^N \, 1] \, \boldsymbol{y} = r_M \ , \\ & \boldsymbol{y} \geq 0 \end{aligned} \tag{7}$$

where $H_j^k = H(k, j)$ is the $|\mathcal{S}|$-dimensional vector whose ℓ-th entry is the probability of observing signal ℓ if the action is k and the outcome is j. We first claim that $\Delta \mapsto \rho(\boldsymbol{p}, \Delta)$ is Lipschitz for every fixed $\boldsymbol{p}$. Indeed, for every fixed $\boldsymbol{p}$ the optimization problem involves Δ only through the objective function. We thus have that the solution to the optimization problem is obtained at one of finitely many values of $\boldsymbol{y}$ (the vertices of the feasible cone defined by the constraints of program (7)). (More precisely, the obtained cone may be unbounded if there are some unconstrained components of $\boldsymbol{y}$. This happens when $H_j^k(\ell) = 0$ for all j. But then $\Delta_k(\ell) = 0$ as well and we do not care about the component $(k-1)N+\ell$

of $\boldsymbol{y}$.) Since $\rho(\boldsymbol{p}, \cdot)$ is a maximum of finitely many linear functions we obtain that it is Lipschitz. We now prove that the Lipschitz constant is uniform with respect to $\boldsymbol{p}$. Since the objective function is non-negative ($\Delta \geq 0$) we can replace the equality signs with inequality signs in the constraints of program (7) while still having the same solutions. Consider the polytope defined by:

$$\left\{ \boldsymbol{y} \in \mathbb{R}^{SN+1} : \boldsymbol{y} \geq 0; [H_m^1 \, H_m^2 \, \ldots \, H_m^N \, 1] \, \boldsymbol{y} \leq 1, \; m = 1, 2, \ldots, M \right\}.$$

This is a cone, and the $\boldsymbol{y}$ with the maximum ℓ_1 norm upper bounds the Lipschitz constant of the $\rho(\boldsymbol{p}, \cdot)$, for all $\boldsymbol{p}$. (As before, any unbounded components of $\boldsymbol{y}$ do not matter to the optimization problem.)