MATHEMATICS OF OPERATIONS RESEARCH
Vol. 33, No. 3, August 2008, pp. 513–528
ISSN 0364-765X | EISSN 1526-5471 | 08 | 3303 | 0513

informs®

DOI 10.1287/moor.1080.0312

Strategies for Prediction Under Imperfect Monitoring

Gábor Lugosi
ICREA and Department of Economics, Pompeu Fabra University, Barcelona, Spain,
lugosi@upf.es, http://www.econ.upf.es/~lugosi

Shie Mannor
Department of Electrical and Computer Engineering, McGill University, Montreal, Québec, Canada,
shie.mannor@mcgill.ca, http://www.ece.mcgill.ca/~smanno1/

Gilles Stoltz
Département de Mathématiques et Applications, Ecole Normale Supérieure, CNRS, Paris, France, and
HEC Paris School of Management, CNRS, Jouy-en-Josas, France, gilles.stoltz@ens.fr, http://www.dma.ens.fr/~stoltz

We propose simple randomized strategies for sequential decision (or prediction) under imperfect monitoring, that is, when the decision maker (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 among all fixed actions. It was Rustichini [Rustichini, A. 1999. Minimizing regret: The general case. *Games Econom. Behav.* **29** 224–243] 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 signals, these rates are optimal up to logarithmic terms.

Key words: repeated games; regret; Hannan consistency; imperfect monitoring; online learning
MSC2000 subject classification: Primary: 91A20, 62L12; secondary: 68Q32
OR/MS subject classification: Primary: computer science: artificial intelligence; decision analysis: sequential; secondary: games/group decisions: noncooperative
History: Received May 29, 2007; revised December 20, 2007.

1. Introduction. In sequential decision problems, a decision maker (or forecaster) tries to predict the outcome of a certain unknown process at each (discrete) time instance and takes an action accordingly. Depending on the outcome of the predicted event and the action taken, the decision maker receives a reward. Very often, probabilistic modeling of the underlying process is difficult. For such situations, the prediction problem can be formalized as a repeated game between the decision maker and the environment. This formulation goes back to the 1950s when Hannan [16] and Blackwell [6] showed that the decision maker has a randomized strategy that guarantees, regardless of the outcome sequence, an average asymptotic reward as high as the maximal reward one could get by knowing the empirical distribution of the outcome sequence in advance. Such strategies are called *Hannan consistent*. To prove this result, Hannan and Blackwell assumed that the decision maker has full access to the past outcomes. This case is termed the *full-information* or the *perfect monitoring case*. However, in many important applications, the decision maker has limited information about the past elements of the sequence to be predicted. Various models of limited feedback have been considered in the literature. Perhaps the best known of them is the so-called *multi-armed bandit problem* in which the forecaster is only informed of its own reward but not the actual outcome; see Baños [4], Megiddo [23], Foster and Vohra [14], Auer et al. [2], and Hart and Mas Colell [17, 18]. For example, it is shown in Auer et al. [2] that Hannan consistency is achievable in this case as well.

Sequential decision problems like the ones considered in this paper have been studied in different fields under various names such as repeated games, regret minimization, online learning, prediction of individual sequences, and sequential prediction. The vocabulary of different subcommunities differ. Ours is perhaps closest to that used by learning theorists. For a general introduction and survey of the sequential prediction problem, we refer to Cesa-Bianchi and Lugosi [9].

In this paper, we consider a general model in which the information available to the forecaster is a general given (possibly randomized) function of the outcome and the decision of the forecaster. It is well understood under what conditions Hannan consistency is achievable in this setup; see Piccolboni and Schindelhauer [25] and Cesa-Bianchi et al. [10]. Roughly speaking, this is possible whenever, after suitable transformations of the problem, the reward matrix can be expressed as a linear function of the matrix of (expected) feedback signals. However, this condition is not always satisfied and then the natural question is what the best achievable performance for the decision maker is. This question was answered by Rustichini [26], who characterized the maximal achievable average reward that can be guaranteed asymptotically for all possible outcome sequences (in an almost sure sense).

However, Rustichini's proof of achievability is not constructive. It uses abstract approachability theorems due to Mertens et al. [24] and it seems unlikely that his proof method can give rise to computationally efficient prediction algorithms, as noted in the conclusion of Rustichini [26]. A simplified efficient approachability-based strategy in the special case where the feedback is a function of the action of nature alone was shown in Mannor and Shimkin [22]. In the general case, the simplified approachability-based strategy of Mannor and Shimkin [22] falls short of the maximal achievable average reward characterized by Rustuchini [26]. The goal of this paper is to develop computationally efficient forecasters in the general prediction problem under imperfect monitoring that achieve the best-possible asymptotic performance.

We introduce several forecasting strategies that exploit some specific properties of the problem at hand. We separate four cases, according to whether the feedback signal only depends on the outcome or both on the outcome and the forecaster's action and whether the feedback signal is deterministic or not. We design different prediction algorithms for all four cases.

As a by-product, we also obtain finite-horizon performance bounds with explicit guaranteed rates of convergence in terms of the number n of rounds the prediction game is played. In the case of deterministic feedback signals, these rates are optimal up to logarithmic factors. In the random feedback signal case, we do not know if it is possible to construct forecasters with a significantly smaller regret.

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, and, in particular, 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 trade-off between the amount of power wasted on transmission and the cost of having its message only partially received. The transmitting agent may 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, because 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.

This paper is organized as follows. In the next section, we formalize the prediction problem we investigate, introduce the target quantity, that is, the best achievable reward, and the notion of regret. In §3, we describe some analytical properties of a key function ρ, defined in §2. This function represents the worst-possible average reward for a given vector of observations and is needed in our analysis. In §4, we consider the simplest special case when the actions of the forecaster do not influence the feedback signal, 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), where n is the number of rounds of the prediction game. In §5, we study random feedback signals 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 §6. The forecaster introduced there has a regret of the order of $n^{-1/5}\sqrt{\log n}$. Finally, in §7 we show that this may be improved to $O(n^{-1/3})$ in the case of deterministic feedback signals, which is known to be optimal (see Cesa-Bianchi et al. [10]).

2. Problem setup and notation. The randomized prediction problem is described as follows. Consider a sequential decision problem in which a forecaster has to predict an outcome that may be thought of as an action taken by the environment.

At each round, $t = 1, 2, \ldots, n$, 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 tth round, the forecaster chooses a probability distribution $\mathbf{p}_t = (p_{1,t}, \ldots, p_{N,t})$ over the set of actions, and plays action i with probability $p_{i,t}$. We denote the forecaster's (random) action at time t by I_t. If the environment chooses action $J_t \in \{1, \ldots, M\}$, then the reward of the forecaster is $r(I_t, J_t)$. The prediction problem is defined as follows:

Randomized prediction under perfect monitoring. Parameters. Number N of actions, cardinality M of outcome space, reward function r, and 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 $\mathbf{p}_t$ and determines the random action I_t, distributed according to $\mathbf{p}_t$;

(3) the environment reveals J_t;
(4) the forecaster receives a reward $r(I_t, J_t)$.

Note in particular that the environment may react to the forecaster's strategy by using a possibly randomized strategy. Below, the probabilities of the considered events are taken with respect to the forecaster's and the environment's randomized strategies. The goal of the forecaster is to minimize the average regret and to enforce that

$$\limsup_{n\to\infty}\left(\max_{i=1,\ldots,N}\frac{1}{n}\sum_{t=1}^{n} r(i, J_t) - \frac{1}{n}\sum_{t=1}^{n} r(I_t, J_t)\right) \le 0 \quad \text{a.s.},$$

that is, the per-round realized differences between the cumulative reward of the best fixed strategy $i \in \{1, \ldots, N\}$, in hindsight, and the reward of the forecaster, are asymptotically nonpositive. Denoting by $r(\mathbf{p}, j) = \sum_{i=1}^{N} p_i r(i, j)$ the linear extension of the reward function r, the Hoeffding-Azuma inequality for sums of bounded martingale differences (see Hoeffding [19], Azuma [3]) implies that for any $\delta \in (0, 1)$, with probability at least $1-\delta$,

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

so it suffices to study the average expected reward $(1/n)\sum_{t=1}^{n} r(\mathbf{p}_t, J_t)$. Hannan [16] and Blackwell [6] 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 a simple yet powerful forecasting strategy known as the *exponentially weighted average forecaster*. This forecaster selects, at time t, an action I_t according to the probabilities

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

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

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

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 Littlestone and Warmuth [21], Vovk [27, 28], Cesa-Bianchi et al. [12], and Cesa-Bianchi [7]; see also Cesa-Bianchi and Lugosi [8].

In this paper, we are concerned with problems in which the forecaster does not have access to the outcomes J_t or to the rewards $r(i, J_t)$. The information available to the forecaster at each round is called the *feedback signal*. These feedback signals 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 signal 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 signal 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 signal s_t is then drawn at random according to the probability distribution $H(I_t, J_t)$ by using an external independent randomization.

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 concentrated on the single element $s \in \mathcal{S}$.

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

In many interesting situations, the feedback signal the forecaster receives is independent of the forecaster's action and only depends on the outcome, that is, for all $j = 1, \ldots, M$, $H(\cdot, j)$ is constant. In other words, H depends on the outcome J_t but not on the forecaster's action I_t. We will see that the prediction problem becomes significantly simpler in this special case. To simplify notation in this case, we write $H(J_t) = H(I_t, J_t)$ for the feedback signal at time t ($h(J_t) = h(I_t, J_t)$ in case of deterministic feedback signals). This setting includes the full-information case (when the outcomes J_t are revealed) but also the case of noisy observations (when a random variable with distribution depending only on J_t is observed); see Weissman and Merhav [29] and Weissman et al. [30].

Parameters: Number N of actions, number M of outcomes, reward function r, random feedback function H, number n of rounds.
For each round, $t = 1, 2, \ldots, n$,
(i) the environment chooses the next outcome $J_t \in \{1, \ldots, M\}$ without revealing it;
(ii) the forecaster chooses a probability distribution $\mathbf{p}_t$ over the set of N actions and draws an action $I_t \in \{1, \ldots, N\}$ according to this distribution;
(iii) the forecaster receives reward $r(I_t, J_t)$ and each action i gets reward $r(i, J_t)$, but none of these values is revealed to the forecaster;
(iv) a feedback signal s_t drawn at random according to $H(I_t, J_t)$ is revealed to the forecaster.

FIGURE 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 $\mathbf{p} = (p_1, \ldots, p_N)$ and $\mathbf{q} = (q_1, \ldots, q_M)$ are probability distributions over $\{1, \ldots, N\}$ and $\{1, \ldots, M\}$, respectively, then, with a slight abuse of notation, we write

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

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 $\mathbf{q} = (q_1, \ldots, q_M)$ over $\{1, \ldots, M\}$, define the vector in $\mathcal{P}(\mathcal{S})$,

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

Denote by $\mathcal{F}$ the convex set of all N-vectors $H(\cdot, \mathbf{q}) = (H(1, \mathbf{q}), \ldots, H(N, \mathbf{q}))$ of probability distributions obtained this way when $\mathbf{q}$ varies. ($\mathcal{F} \subset \mathcal{P}(\mathcal{S})^N$ is the set of feasible distributions over the signals.) In the case when the feedback signals only depend on the outcome, all components of this vector are equal and we denote their common value by $H(\mathbf{q})$. 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 $\mathbf{p}$ over $\{1, \ldots, N\}$ and probability distribution $\Delta \in \mathcal{F}$, we may assign the quantity

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

which is the reward guaranteed by the mixed action $\mathbf{p}$ of the forecaster against any distribution of the outcomes that induces the given distribution of feedback signals Δ. Note that $\rho \in [0, 1]$ and that ρ is concave in $\mathbf{p}$ (because it is an infimum of linear functions; because this infimum is taken on a convex set, the infimum is indeed a minimum). Finally, ρ is also convex in Δ because the condition defining the minimum is linear in Δ.

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

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

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

Rustichini [26] 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.

Rustichini's proof is not constructive but in several special cases constructive and computationally efficient prediction algorithms have been proposed. Among the partial solutions proposed so far, we mention Piccolboni and Schindelhauer [25] and Cesa-Bianchi et al. [10], who study the case when

$$\max_{\mathbf{p}} \rho(\mathbf{p}, H(\cdot, \bar{\mathbf{q}}_n)) = \max_{i=1,\ldots,N} r(i, \bar{\mathbf{q}}_n) = \max_{i=1,\ldots,N} \frac{1}{n} \sum_{t=1}^{n} r(i, J_t).$$

In this case, strategies with a vanishing per-round regret are called *Hannan consistent*. In such cases, the feedback is sufficiently rich so that one may achieve the same asymptotic reward as in the full-information case, although the rate of convergence may be slower. This case turns out to be considerably simpler to handle than the general problem and computationally tractable explicit algorithms have been derived. Also, it is shown in Cesa-Bianchi et al. [10] that in this case, it is possible to construct strategies whose regret decreases at a rate of $n^{-1/3}$ (with high probability) and that this rate of convergence cannot be improved in general. (Note that Hannan consistency is achievable, for example, in the adversarial multiarmed bandit problem; see Remark B.1 in the appendix.) Mannor and Shimkin [22] construct an approachability-based algorithm with vanishing regret for the special case where the feedback signals depend only on the outcome. In addition, Mannor and Shimkin discuss the more general case of feedback signals that depend on both the action and the outcome and provide an algorithm that attains a relaxed goal comparing to the one attained in this work.

The following example demonstrates the structure of the model.

Example 2.1. Consider the simple game where $N = 2$, $M = 3$, $\mathcal{S} = \{a, b\}$, and the reward and feedback functions are as follows. The reward function is described by the matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

To identify the possible distributions of the feedback signals, we need to specify some elements of $\mathcal{P}(\mathcal{S})$. We describe such a member of $\mathcal{P}(\mathcal{S})$ by the probability of observing a. The feedback function is parameterized by some $\varepsilon > 0$ and is then given by

$$\begin{bmatrix} 1 & 1-\varepsilon & 0 \\ 1 & 1-\varepsilon & 0 \end{bmatrix}.$$

In words, outcome 1 leads to a deterministic feedback signal of a, outcome 3 leads to a deterministic feedback signal of b, and outcome 2 leads to a feedback signal of a with probability $1-\varepsilon$ and b with probability ε. Note that the feedback signals depend only on the outcome and not on the action taken. We recall that Δ, as a member of $\mathcal{P}(\mathcal{S})$, is identified with the probability of observing the feedback signal a and it follows that $\mathcal{F}$ is the interval $[0, 1]$. We now compute the function ρ. Letting p denote the probability of selecting the first action (i.e., $\mathbf{p} = (p, 1-p)$), we have

$$\begin{aligned} \rho(\mathbf{p}, \Delta) &= \min_{\mathbf{q}:\, q_1+(1-\varepsilon)q_2=\Delta} \left(p q_1 + (1-p)\frac{q_1+q_2+q_3}{2} \right) = \min_{\mathbf{q}:\, \varepsilon q_1-(1-\varepsilon)q_3=\Delta-(1-\varepsilon)} p\, q_1 + \frac{1-p}{2} \\ &= \frac{1-p}{2} + \begin{cases} 0 & \text{for } \Delta \le 1-\varepsilon, \\ p\dfrac{\Delta-(1-\varepsilon)}{\varepsilon} & \text{for } 1-\varepsilon \le \Delta \le 1. \end{cases} \end{aligned}$$

Optimizing over p, we obtain

$$\max_{\mathbf{p}} \rho(\mathbf{p}, \Delta) = \begin{cases} \dfrac{1}{2} & \text{for } \Delta \le 1-\varepsilon/2, \\ \dfrac{\Delta-(1-\varepsilon)}{\varepsilon} & \text{for } 1-\varepsilon/2 \le \Delta \le 1. \end{cases}$$

The intuition here is that for $\Delta = 1$, there is certainty that the outcome is 1 so that an action of $p = 1$ is optimal. For $\Delta \le 1-\varepsilon$, the forecaster does not know if the outcome was consistently 2 or some mixture of outcomes 1 and 3. By playing the second action, the forecaster can guarantee a reward of 1/2. The function $\Delta \mapsto \max_{\mathbf{p}} \rho(\mathbf{p}, \Delta)$ is depicted in Figure 2.

In this paper, we construct simple and computationally efficient strategies whose regret vanishes with probability one. The main idea behind the forecasters we introduce in the next sections is based on the gradient-based strategies described, for example, in Cesa-Bianchi and Lugosi [9, §2.5]. Our forecasters use subgradients of concave functions. In the next section, we briefly recall some basic facts on the existence, computation, and boundedness of these subgradients.

3. Some analytical properties of ρ.

For a concave function f defined over a convex subset of $\mathbb{R}^d$, a vector $\mathbf{b}(\mathbf{x}) \in \mathbb{R}^d$ is a subgradient of f at $\mathbf{x}$ if $f(\mathbf{y}) - f(\mathbf{x}) \le \mathbf{b}(\mathbf{x}) \cdot (\mathbf{y}-\mathbf{x})$ for all $\mathbf{y}$ in the domain of f. We denote by

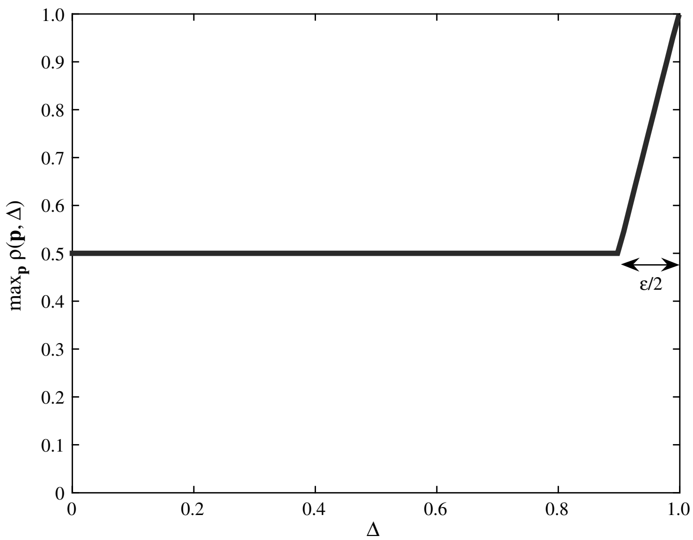

FIGURE 2. The function $\Delta \mapsto \max_{\mathbf{p}} \rho(\mathbf{p}, \Delta)$ for Example 2.1.

$\partial f(\mathbf{x})$ the set of subgradients of f at $\mathbf{x}$, which is also known as the subdifferential. Subgradients always exist, that is, $\partial f(\mathbf{x})$ is nonempty in the interior of the domain of a concave function. In this paper, we are interested in subgradients of concave functions of the form $f(\cdot) = \rho(\cdot, \hat{\Delta}_t)$, where $\hat{\Delta}_t$ is an observed or estimated distribution of the feedback signal at round t. (For example, in §4, $\hat{\Delta}_t = \delta_{h(J_t)}$ is observed; in the other sections, it will be estimated.) In view of the exponentially weighted update rules that are used below, we only evaluate these functions in the interior of the definition domain (the simplex). Thus, the existence of subgradients is ensured throughout.

In the general case, subgradients may be computed efficiently by the simplex method. However, their computation is often even simpler, as in the case described in §4, that is, when one faces deterministic feedback signals not depending on the actions of the forecaster. Indeed, at round t, it is trivial whenever $\mathbf{p} \mapsto \rho(\mathbf{p}, \delta_{h(J_t)})$ is differentiable at the considered point $\mathbf{p}_t$ because it is differentiable exactly at those points at which it is locally linear, and thus the gradient equals the column of the reward matrix corresponding to the outcome y_t for which $r(\mathbf{p}_t, y_t) = \rho(\mathbf{p}_t, \delta_{h(J_t)})$. But because $\rho(\cdot, \delta_{h(J_t)})$ is concave, the Lebesgue measure of the set where it is nondifferentiable equals zero. It thus suffices to resort to the simplex method only at these points to compute the subgradients.

Note that the components of the subgradients are always bounded by a constant that depends on the game parameters. This is the case because the $\rho(\cdot, \hat{\Delta}_t)$ are concave and continuous on a compact set and are therefore Lipschitz, leading to a bounded subgradient. In the sequel, we denote by K the value $\sup_{\mathbf{p}} \sup_{\Delta} \sup_{\mathbf{b} \in \partial\rho(\mathbf{p}, \Delta)} \|\mathbf{b}\|_\infty$, where $\partial\rho(\mathbf{p}, \Delta)$ denotes the subgradient at $\mathbf{p}$ of the concave function $\rho(\cdot, \Delta)$ with Δ fixed. This constant depends on the specific parameters of the game. Because the parameters of the game are supposed to be known to the forecaster, in principle, the forecaster can compute the value of K. In any case, the value of K can be bounded by the supremum norm of the payoff function as the following lemma asserts.

LEMMA 3.1. *The constant K satisfies $K \le 1$.*

PROOF. Fix Δ and consider $Z^\Delta = \{\mathbf{q}\colon H(\cdot, \mathbf{q}) = \Delta\}$. Define $\varphi\colon (\mathbf{p}, \mathbf{q}) \in \mathbb{R}^n \times Z^\Delta \mapsto \varphi(\mathbf{p}, \mathbf{q}) \in \mathbb{R}$ as the linear extension restriction of r to $\mathbb{R}^n \times Z^\Delta$, that is, $\varphi(\mathbf{p}, \mathbf{q}) = \sum_{i,j} p_i q_j r(i, j)$. Further, let $Z_0^\Delta(\mathbf{p}) = \{\bar{\mathbf{q}}\colon \varphi(\mathbf{p}, \bar{\mathbf{q}}) = \min_{\mathbf{q} \in Z^\Delta} \varphi(\mathbf{p}, \mathbf{q})\}$. It follows that under our notation, for any probability distribution $\mathbf{p}$, one has $\rho(\mathbf{p}, \Delta) = \min_{\mathbf{q} \in Z^\Delta} \varphi(\mathbf{p}, \mathbf{q})$. Now, from Danskin's theorem (see, e.g., Bertsekas [5]), we have that the subdifferential satisfies

$$\partial\rho(\mathbf{p}, \Delta) = \operatorname{conv}\left(\frac{\partial\varphi(\mathbf{p}, \mathbf{z})}{\partial\mathbf{p}}\colon \mathbf{z} \in Z_0^\Delta(\mathbf{p})\right),$$

where $\operatorname{conv}(A)$ denotes the convex hull of a set A. Because $r(i, j) \in [0, 1]$, it follows that $\|\partial\rho(\mathbf{p}, \mathbf{z})/\partial\mathbf{p}\|_\infty \le 1$ for all $\mathbf{z} \in Z^\Delta$. Because the convex hull does not increase the infinity norm, the result follows. □

REMARK 3.1. The constant K for the game described in Example 2.1 is $1/2$. However, the gradient of the function $\max_{\mathbf{p}} \rho(\mathbf{p}, \Delta)$ as a function of Δ is $1/\varepsilon$. This happens because the $\mathbf{p}$ that attains the maximum changes rapidly in the interval $[1 - \varepsilon/2, 1]$. We further note that K may be much smaller than 1. Because our regret bounds below depend on K linearly, having a tighter bound on K can lead to considerable convergence rate speedup; see Remark 4.1.

4. Deterministic feedback signals only depending on outcome. We start with the simplest case when the feedback signal is deterministic and does not depend on the action I_t of the forecaster. In other words, after making the prediction at time t, the forecaster observes $h(J_t)$. This simplifying assumption may be naturally satisfied in applications in which the forecaster's decisions do not effect the environment.

In this case, we group the outcomes according to the deterministic feedback signal 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 because r is lower bounded by ρ in the sense that for all $\mathbf{p}$ and j,

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

As mentioned in the previous section, we introduce a forecaster based on the subgradients of $\rho(\cdot, \delta_{h(J_t)})$, $t = 1, 2, \ldots$. The forecaster requires a tuning parameter $\eta > 0$. The ith component of $\mathbf{p}_t$ is

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

where $(\tilde{r}(\mathbf{p}_s, \delta_{h(J_s)}))_i$ is the ith component of any subgradient $\tilde{r}(\mathbf{p}_s, \delta_{h(J_s)}) \in \partial\rho(\mathbf{p}_s, \delta_{h(J_s)})$ of the concave function $\rho(\cdot, \delta_{h(J_s)})$. This forecaster is inspired by a gradient-based predictor introduced by Kivinen and Warmuth [20].

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.)

Proposition 4.1. *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 $1 - \delta$,*

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

where K is the bound on the subgradients considered above. In particular, choosing $\eta \sim \sqrt{(\ln N)/n}$ yields $R_n = O(n^{-1/2}\sqrt{\ln(N/\delta)})$.

Remark 4.1. The optimal choice of η in the upper bound is $K\sqrt{2(\ln N)/n}$, which depends on the parameters K and n. While the bound $K \leq 1$ is available, this bound might be loose. Sometimes the forecaster does not necessarily know in advance the number of prediction rounds and/or the value of K may be difficult to compute. In such cases, one may estimate online both the number of time rounds and K, using the techniques of Auer et al. [1] and Cesa-Bianchi et al. [11] as follows. Writing

$$K_t = \max_{s \leq t-1} \|\tilde{r}(\mathbf{p}_s, \delta_{h(J_s)})\|_\infty,$$

and introducing a round-dependent choice of the tuning parameter $\eta = \eta_t = CK_t\sqrt{(\ln N)/t}$ for a properly chosen constant C, one may prove a regret bound that is a constant multiple of $K_n\sqrt{(\ln N/\delta)/n}$ (that holds with probability at least $1 - \delta$). Because the proof of this is a straightforward combination of the techniques of the above-mentioned papers and our proof, the details are omitted.

Proof. Note that because the feedback signals are deterministic, $H(\bar{\mathbf{q}}_n)$ takes the simple form $H(\bar{\mathbf{q}}_n) = (1/n)\sum_{t=1}^{n} \delta_{h(J_t)}$. Now, for any $\mathbf{p}$,

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

where at the last step we used the fact that the forecaster is just the exponentially weighted average predictor based on the rewards $(\tilde{r}(\mathbf{p}_s, \delta_{h(J_s)}))_i$ 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 $1-\delta$,

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

5. Random feedback signals depending only on the outcome. Next, we consider the case when the feedback signals do 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

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

anymore, but only observes the realizations s_t drawn at random according to $H(J_t)$. To overcome this problem, we group together several consecutive time rounds (say, 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 $\mathcal{F}$ (because the feedback signals depend only on the outcome, we may now view the set $\mathcal{F}$ of feasible distributions 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

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

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 3. 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 polynomial time. 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 signals). We do not know whether this rate can be improved significantly because we do not know of any nontrivial lower bound in this case.

THEOREM 5.1. *For all integers $m \geq 1$, for all $\eta > 0$, and for all $\delta > 0$, the regret for 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 $K \leq 1$ and L are constants that 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 $1-\delta$.

Parameters: Integer $m \geq 1$, real number $\eta > 0$.
Initialization: $\mathbf{w}^0 = (1, \ldots, 1)$.
For each round $t = 1, 2, \ldots$
(i) If $bm+1 \leq t < (b+1)m$ for some integer b, choose the distribution $\mathbf{p}_t = \mathbf{p}^b$ 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;
(ii) if $t = (b+1)m$ for some integer b, perform the update

$$w_k^{b+1} = w_k^b e^{\eta(\tilde{r}(\mathbf{p}^b, \hat{\Delta}^b))_k} \quad \text{for each } k = 1, \ldots, N,$$

where for all Δ, $\tilde{r}(\cdot, \Delta)$ is a subgradient of $\rho(\cdot, \Delta)$ and $\hat{\Delta}^b$ is defined in (3).

FIGURE 3. The forecaster for random feedback signals depending only on the outcome.

Remark 5.1. Here again, K and L may, in principle, be computed or bounded (see Lemma 3.1, Remark A.1) by the forecaster. If the horizon n is known in advance (as it is assumed in this paper), the values of η and m may be chosen to optimize the upper bound for the regret. Observe that while one always has $K \le 1$, the value of L (i.e., the Lipschitz constant of ρ in its second argument) can be arbitrarily large; see Example 2.1. If the horizon n is unknown at the start of the game, the situation is not as simple as in §4 (see Remark 4.1) because now a time-dependent choice of η needs to be accompanied by an adaptive choice of the parameter m as well. A simple, although not very attractive, solution is the so-called "doubling trick" (see, e.g., Cesa-Bianchi and Lugosi [9, p. 17]). According to this solution, time is divided into periods of exponentially growing length, and in each period the forecaster is used as if the horizon was the length of the actual period. At the end of each period, the forecaster is reset and started again with new parameter values. It is easy to see that this forecaster achieves the same regret bounds, up to a constant multiplier. We believe that a smoother solution should also work (as in Remark 4.1). Because this seems like a technical endeavor, we do not pursue this issue further.

Proof. We start by grouping time rounds m by m. For simplicity, we assume that $n = (B+1)m$ for some integer B; if this is not the case, we consider the lower integer part of n and bound the regret suffered in the last at most $m-1$ rounds by m (this accounts for the m/n term in the bound). For all $\mathbf{p}$,

$$\begin{aligned} n\rho(\mathbf{p}, H(\bar{\mathbf{q}}_n)) - \sum_{t=1}^{n} r(\mathbf{p}_t, J_t) &= n\rho(\mathbf{p}, H(\bar{\mathbf{q}}_n)) - \sum_{b=0}^{B} mr\left(\mathbf{p}^b, \frac{1}{m}\sum_{t=bm+1}^{(b+1)m} \delta_{J_t}\right) \\ &\le \sum_{b=0}^{B}\left(m\rho(\mathbf{p}, \Delta^b) - mr\left(\mathbf{p}^b, \frac{1}{m}\sum_{t=bm+1}^{(b+1)m}\delta_{J_t}\right)\right) \\ &\le m\sum_{b=0}^{B}(\rho(\mathbf{p}, \Delta^b) - \rho(\mathbf{p}^b, \Delta^b)), \end{aligned}$$

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 $\hat{\Delta}^b$. By a version of the Hoeffding-Azuma inequality for sums of Hilbert space-valued martingale differences proved by Chen and White [13, Lemma 3.2], and because the l_2 projection can only help, for all b, with probability at least $1-\delta$,

$$\|\Delta^b - \hat{\Delta}^b\|_2 \le \sqrt{\frac{2\ln(2/\delta)}{m}}.$$

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

$$\begin{aligned} m\sum_{b=0}^{B}(\rho(\mathbf{p}, \Delta^b) - \rho(\mathbf{p}^b, \Delta^b)) &\le m\sum_{b=0}^{B}\left(\rho(\mathbf{p}, \hat{\Delta}^b) - \rho(\mathbf{p}^b, \hat{\Delta}^b) + 2L\sqrt{\frac{2\ln(2/\delta)}{m}}\right) \\ &= m\sum_{b=0}^{B}(\rho(\mathbf{p}, \hat{\Delta}^b) - \rho(\mathbf{p}^b, \hat{\Delta}^b)) + 2L(B+1)\sqrt{2m\ln\frac{2}{\delta}}. \end{aligned}$$

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 $\tilde{r}$ denotes a subgradient),

$$\begin{aligned} m\sum_{b=0}^{B}(\rho(\mathbf{p}, \hat{\Delta}^b) - \rho(\mathbf{p}^b, \hat{\Delta}^b)) &\le m\sum_{b=0}^{B}\tilde{r}(\mathbf{p}^b, \hat{\Delta}^b)\cdot(\mathbf{p}-\mathbf{p}^b) \\ &\le m\left(\frac{\ln N}{\eta} + \frac{(B+1)K^2\eta}{2}\right) = \frac{m\ln N}{\eta} + \frac{nK^2\eta}{2}, \end{aligned}$$

where we used Theorem 1 and the boundedness of the function $\tilde{r}$ 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 $1-(B+2)\delta \ge 1-(n/m+1)\delta$. □

6. Random feedback signals depending on the action-outcome pair. We now turn to the general case, where the feedback signals are random and depend on the action-outcome pairs (I_t, J_t). The key is, again, to exhibit efficient estimators of the (unobserved) $H(\cdot, \bar{\mathbf{q}}_n)$.

Denote by Π 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

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

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

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

(This form of estimators is reminiscent of those presented, e.g., in Auer et al. [2], Piccolboni and Schindelhauer [25], Cesa-Bianchi et al. [10].) We prove that the $\hat{h}_{i,t}$ 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 $\mathbf{p}_t$ and J_t. Thus,

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

For the sake of the analysis, introduce

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

The proposed forecasting strategy is described in Figure 4. The mixing with the uniform distribution is needed, similarly to the forecasters presented in Auer et al. [2], Piccolboni and Schindelhauer [25], and Cesa-Bianchi et al. [10], to ensure sufficient exploration of all actions. Mathematically, such a mixing lower bounds the probability of pulling each action, which will turn to be crucial in the proof of Theorem 6.1.

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 do not know whether this bound can be improved significantly. We recall that K denotes an upper bound on the infinity norm of the subgradients (see Lemma 3.1). The issues concerning the tuning of the parameters considered in the following theorem are similar to those discussed after the statement of Theorem 5.1; in particular, the simplest way of being adaptive in all parameters is to use the "doubling trick."

Theorem 6.1. *For all integers $m \ge 1$, for all $\eta > 0$, $\gamma \in (0, 1)$, and $\delta > 0$, the regret for any strategy of the environment is bounded, with probability at least $1 - (n/m + 1)\delta$, as*

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

Parameters: Integer $m \ge 1$, real numbers $\eta, \gamma > 0$.
Initialization: $\mathbf{w}^0 = (1, \ldots, 1)$.
For each round $t = 1, 2, \ldots$

(i) if $bm + 1 \le t < (b+1)m$ for some integer b, choose the distribution $\mathbf{p}_t = \mathbf{p}^b = (1-\gamma)\tilde{\mathbf{p}}^b + \gamma\mathbf{u}$, where $\tilde{\mathbf{p}}^b$ is defined componentwise as

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

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

$$w_k^{b+1} = w_k^b e^{\eta(\tilde{r}(\mathbf{p}^b, \hat{\Delta}^b))_k} \quad \text{for each } k = 1, \ldots, N,$$

where for all $\Delta \in \mathcal{F}$, $\tilde{r}(\cdot, \Delta)$ is a subgradient of $\rho(\cdot, \Delta)$ and $\hat{\Delta}^b$ is defined in (4).

Figure 4. The forecaster for random feedback signals depending on the action-outcome pair.

where L and $K \le 1$ are constants that 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 $1-\delta$, $R_n = O(n^{-1/5} N \sqrt{\ln(Nn/\delta)} + n^{-2/5} N^{3/2} \ln(Nn/\delta))$.

Proof. The proof is similar to that of Theorem 5.1. 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 [15] rather than the Hoeffding-Azuma inequality. Also, the mixing with the uniform distribution in the first step of the definition of the forecaster in Figure 4 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, again, for the m/n term in the bound). As before, we get that, for all $\mathbf{p}$,

$$n\rho(\mathbf{p}, H(\cdot, \bar{\mathbf{q}}_n)) - \sum_{t=1}^{n} r(\mathbf{p}_t, J_t) \le m \sum_{b=0}^{B} (\rho(\mathbf{p}, \Delta^b) - \rho(\mathbf{p}^b, \Delta^b)) \tag{5}$$

and proceed by estimating Δ^b by $\hat{\Delta}^b$. Freedman's inequality [15] (see, also, Cesa-Bianchi et al. [10, Lemma A.1]) implies that for all $b = 0, 1, \ldots, B$, $i = 1, \ldots, N$, $s \in \mathcal{S}$, and $\delta > 0$,

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

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

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

where the second inequality follows from the lower bound γ/N on the components of $\mathbf{p}_t$ (ensured by the mixing step in the definition of the forecaster). Summing (because the l_2 projection can only help), the union bound shows that for all b, with probability at least $1-\delta$,

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

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

$$\begin{aligned} m \sum_{b=0}^{B} (\rho(\mathbf{p}, \Delta^b) - \rho(\mathbf{p}^b, \Delta^b)) &\le m \sum_{b=0}^{B} (\rho(\mathbf{p}, \hat{\Delta}^b) - \rho(\mathbf{p}^b, \hat{\Delta}^b) + 2Ld) \\ &= m \sum_{b=0}^{B} (\rho(\mathbf{p}, \hat{\Delta}^b) - \rho(\mathbf{p}^b, \hat{\Delta}^b)) + 2m(B+1)Ld. \end{aligned}$$

The terms $2m(B+1)Ld = 2nLd$ 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 $\tilde{r}$ denotes a subgradient bounded between $-K$ and K):

$$m \sum_{b=0}^{B} (\rho(\mathbf{p}, \hat{\Delta}^b) - \rho(\mathbf{p}^b, \hat{\Delta}^b)) \le m \sum_{b=0}^{B} \tilde{r}(\mathbf{p}^b, \hat{\Delta}^b) \cdot (\mathbf{p} - \mathbf{p}^b).$$

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

$$\begin{aligned} m \sum_{b=0}^{B} \tilde{r}(\mathbf{p}^b, \hat{\Delta}^b) \cdot (\mathbf{p} - \mathbf{p}^b) &\le (1-\gamma) m \sum_{b=0}^{B} \tilde{r}(\mathbf{p}^b, \hat{\Delta}^b) \cdot (\mathbf{p} - \tilde{\mathbf{p}}^b) + 2K\gamma m(B+1) \\ &\qquad \text{(because by definition, } \mathbf{p}^b = (1-\gamma)\tilde{\mathbf{p}}^b + \gamma \mathbf{u}) \\ &\le (1-\gamma) m \left(\frac{\ln N}{\eta} + \frac{(B+1)K^2 \eta}{2} \right) + 2K\gamma m(B+1) \quad \text{(by (1))} \\ &\le \frac{m \ln N}{\eta} + \frac{nK^2 \eta}{2} + 2K\gamma n. \end{aligned}$$

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 holds with probability at least $1-(B+2)\delta \ge 1-(n/m+1)\delta$. $\square$

7. Deterministic feedback signals depending on the action-outcome pair. In this last section, we explain how in the case of deterministic feedback signals, 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 signals, ρ may not be linear and it is because of this fact that we needed to group rounds of size m. If the feedback signals are deterministic, such grouping is not needed and the rate $n^{-1/3}$ is obtained as a trade-off between an exploration term (γ) and the cost payed for estimating the feedback signals ($\sqrt{1/(\gamma n)}$). This rate of convergence has been shown to be optimal in Cesa-Bianchi et al. [10], even in the Hannan-consistent case. The key property is summarized in the next technical lemma, whose proof can be found in the appendix.

LEMMA 7.1. *For every fixed* $\mathbf{p}$, *the function* $\rho(\mathbf{p}, \cdot)$ *is linear on* $\mathcal{F}$.

REMARK 7.1. The fact that the forecaster does not need to group rounds in the case of deterministic feedback signals has an interesting consequence. It is easy to see from the proofs of Proposition 4.1 and Theorem 7.1, through the linearity property stated above, that the results presented there are still valid when the payoff function r may change with time (even when the environment can set it). The definition of the regret is then generalized as

$$R_n = \max_{\mathbf{p}} \min_{z_1^n:\, H(\cdot, \bar{\mathbf{z}}_n) = H(\cdot, \bar{\mathbf{q}}_n)} \frac{1}{n}\sum_{t=1}^{n} r_t(\mathbf{p}, z_t) - \frac{1}{n}\sum_{t=1}^{n} r_t(I_t, J_t),$$

where $\bar{\mathbf{z}}_n$ is the empirical distribution of the sequence of outcomes $z_1^n = (z_1, \ldots, z_n)$, and the same bounds hold. This may model some more complex situations, including Markov decision processes. Note that choosing time-varying reward functions was not possible with the forecasters of Piccolboni and Schindelhauer [25] and Cesa-Bianchi et al. [10] because these relied on a crucial structural assumption on the relation between r and h.

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 $\mathcal{H}$. Because the $\rho(\mathbf{p}, \cdot)$ 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 $\rho(\mathbf{p}, \Pi(\cdot))$ defined on $\mathbb{R}^{|\mathcal{S}|N}$ are linear and coincide with the original definition on $\mathcal{F}$. We denote by $\tilde{r}$ a subgradient (i.e., for all $\Delta \in \mathbb{R}^{|\mathcal{S}|N}$, $\tilde{r}(\cdot, \Delta)$ is a subgradient of $\rho(\cdot, \Pi(\Delta))$).

The subgradients are evaluated at the following points. (Recall that because the feedback signals are deterministic, $s_t = h(I_t, J_t)$.) For $t = 1, 2, \ldots$, let

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

The $\hat{h}_{i,t}$ estimate the feedback signals $H(i, J_t) = \delta_{h(i, J_t)}$ received by action i at round t. They are still conditionally unbiased estimators of the $h(i, J_t)$, and so is $\hat{h}_t$ for $H(\cdot, J_t)$. The proposed forecaster is defined in Figure 5 and the regret bound is established in Theorem 7.1.

THEOREM 7.1. *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 for any strategy of the environment is bounded, with probability at least $1 - \delta$, as*

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

Parameters: Real numbers $\eta, \gamma > 0$.
Initialization: $\mathbf{w}_1 = (1, \ldots, 1)$.
For each round $t = 1, 2, \ldots$
(i) choose the distribution $\mathbf{p}_t = (1-\gamma)\tilde{\mathbf{p}}_t + \gamma\mathbf{u}$, where $\tilde{\mathbf{p}}_t$ is defined componentwise as

$$\tilde{p}_{k,t} = \frac{w_{k,t}}{\sum_{j=1}^{N} w_{j,t}}$$

and $\mathbf{u}$ denotes the uniform distribution, $\mathbf{u} = (1/N, \ldots, 1/N)$; then draw an action I_t from $\{1, \ldots, N\}$ according to $\mathbf{p}_t$;
(ii) perform the update

$$w_{k,t+1} = w_{k,t}\, e^{\eta(\tilde{r}(\mathbf{p}_t, \hat{h}_t))_k} \quad \text{for each } k = 1, \ldots, N,$$

where Π is the projection operator defined after the statement of Lemma 7.1, for all $\Delta \in \mathbb{R}^{|\mathcal{S}|N}$, $\tilde{r}(\cdot, \Delta)$ is a subgradient of $\rho(\cdot, \Pi(\Delta))$, and $\hat{h}_t$ is defined in (6).

FIGURE 5. The forecaster for deterministic feedback signals depending on the action-outcome pair.

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

Note that here, as in §4 (see Remark 4.1), the tuning of the parameters can be done efficiently online without resorting to the "doubling trick." The optimization of the upper bound (in both γ and η) requires the knowledge of N, C, K, and n. The first three parameters only depend on the game and are known or may be calculated beforehand (the proof indicates an explicit expression for C and the bound on the subgradients may be computed as explained in §3). If n and/or K are unknown, their tuning may be dealt with by taking time-dependent γ_t and η_t.

Proof. The proof is similar to the one of Theorem 6.1, except that we do not have to consider the grouping steps and that we do not apply the Hoeffding-Azuma inequality to the estimated feedback signals but to the estimated rewards. By the bound on r in terms of ρ and convexity (linearity) of ρ in its second argument,

$$n\rho(\mathbf{p}, H(\cdot, \bar{\mathbf{q}}_n)) - \sum_{t=1}^{n} r(\mathbf{p}_t, J_t) \le \sum_{t=1}^{n} (\rho(\mathbf{p}, H(\cdot, J_t)) - \rho(\mathbf{p}_t, H(\cdot, J_t))).$$

Next, we estimate

$$\rho(\mathbf{p}, H(\cdot, J_t)) - \rho(\mathbf{p}_t, H(\cdot, J_t)) \quad \text{by } \rho(\mathbf{p}, \Pi(\hat{h}_t)) - \rho(\mathbf{p}_t, \Pi(\hat{h}_t)).$$

By Freedman's inequality (see, again, Cesa-Bianchi et al. [10, Lemma A.1]), because $\hat{h}_t$ 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 $1-\delta/2$,

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

where, denoting by $\mathbf{e}_i(\delta_{h(i,j)})$ the column vector whose ith component is $\delta_{h(i,j)}$ and all other components equal zero,

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

(A more precise look at the definition of C shows that it is less than the maximal l_1 norm of the barycentric coordinates of the points $\Pi[\mathbf{e}_i(\delta_{h(i,j)})]$ with respect to the $h(\cdot, j)$.) This is because for all t, the conditional variances are bounded as follows. For all $\mathbf{p}'$,

$$\begin{aligned}\mathbb{E}_t[\rho(\mathbf{p}', \Pi(\hat{h}_t))^2] &= \sum_{i=1}^{N} p_{i,t}\rho(\mathbf{p}', \Pi[\mathbf{e}_i(\delta_{h(i,j)}/p_{i,t})])^2 \\ &= \sum_{i=1}^{N}\frac{1}{p_{i,t}}\rho(\mathbf{p}', \Pi[\mathbf{e}_i(\delta_{h(i,j)}/p_{i,t})])^2 \le \sum_{i=1}^{N}\frac{C^2}{p_{i,t}} \le \frac{C^2N^2}{\gamma}.\end{aligned}$$

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

$$\sum_{t=1}^{n}(\rho(\mathbf{p}, \Pi(\hat{h}_t)) - \rho(\mathbf{p}_t, \Pi(\hat{h}_t))) \le \sum_{t=1}^{n}\tilde{r}(\mathbf{p}_t, \hat{h}_t)\cdot(\mathbf{p} - \mathbf{p}_t).$$

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

$$\begin{aligned}\sum_{t=1}^{n}\tilde{r}(\mathbf{p}, \hat{h}_t)\cdot(\mathbf{p} - \mathbf{p}) &\le (1-\gamma)\sum_{t=1}^{n}\tilde{r}(\mathbf{p}_t, \hat{h}_t)\cdot(\mathbf{p} - \tilde{\mathbf{p}}_t) + 2K\gamma n \\ &\qquad \text{(because by definition, } \mathbf{p}_t = (1-\gamma)\tilde{\mathbf{p}}_t + \gamma\mathbf{u}) \\ &\le (1-\gamma)\left(\frac{\ln N}{\eta} + \frac{n\eta K^2}{2}\right) + 2K\gamma n \quad \text{(by (1))}.\end{aligned}$$

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

Appendix A. Uniform Lipschitzness of ρ.

Proposition A.1. *The function* $(\mathbf{p}, \Delta) \mapsto \rho(\mathbf{p}, \Delta)$ *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 $\rho(\mathbf{p}, \Delta)$ as the solution of the following linear program (we denote $\Delta = (\Delta_1, \dots, \Delta_N) \in \mathscr{F} \subset \mathscr{P}(S)^N$, where, as usual, we identify each Δ_j with a $|\mathscr{S}|$-dimensional vector):

$$\begin{aligned}\rho(\mathbf{p}, \Delta) = \min_{\mathbf{q}\in\mathbb{R}^{|\mathscr{S}|}} \quad & r(\mathbf{p}, \cdot)^\top \mathbf{q} \\ \text{s.t.} \quad & H^k \mathbf{q} = \Delta_k, \quad k = 1, 2, \dots, N, \\ & \mathbf{e}_M^\top \mathbf{q} = 1, \\ & \mathbf{q} \ge 0,\end{aligned}$$

where $r(\mathbf{p}, \cdot) = (r(\mathbf{p}, j))_j$ is an M-dimensional vector, $\mathbf{e}_M$ is an M-dimensional vector of ones, and $H^k = H(k, \cdot)$ is the $|\mathscr{S}| \times M$ 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 $\Delta \in \mathscr{F}$, so by the duality theorem,

$$\begin{aligned}\rho(\mathbf{p}, \Delta) = \max_{\mathbf{y}\in\mathbb{R}^{N|\mathscr{S}|+1}} \quad & [\Delta_1^\top \Delta_2^\top \cdots \Delta_N^\top 1]\mathbf{y} \\ \text{s.t.} \quad & [H^1(\cdot, j)^\top H^2(\cdot, j)^\top \cdots H^N(\cdot, j)^\top 1]\mathbf{y} \le r(\mathbf{p}, j), \quad j = 1, 2, \dots, M, \\ & \mathbf{y} \ge 0,\end{aligned} \tag{A1}$$

where we recall that $H^k(\cdot, j)$ is the $|\mathscr{S}|$-dimensional vector, whose sth entry is the probability of observing signal s if the action is k and the outcome is j.

We first claim that $\Delta \mapsto \rho(\mathbf{p}, \Delta)$ is Lipschitz for every fixed $\mathbf{p}$. Indeed, for every fixed $\mathbf{p}$, 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 $\mathbf{y}$ (the vertices of the feasible cone defined by the constraints of program (A1)). (More precisely, the obtained cone may be unbounded if there are some unconstrained components of $\mathbf{y}$. This happens when there exists an s such that $H^k(s, j) = 0$ for all j. But then, $\Delta_k(s) = 0$ as well and we do not care about the unbounded component $(k-1)N + s$ of $\mathbf{y}$.) Because $\rho(\mathbf{p}, \cdot)$ is a maximum of finitely many linear functions, we obtain that it is Lipschitz, with Lipschitz constant bounded by the maximal l_1 norm of the vertices of the feasible cone of (A1).

We now prove that the Lipschitz constant is uniform with respect to $\mathbf{p}$. It suffices to consider the polytope defined by

$$\left\{\mathbf{y} \in \mathbb{R}^{N|\mathscr{S}|+1}\colon \mathbf{y} \ge 0, [H^1(\cdot, j)^\top H^2(\cdot, j)^\top \cdots H^N(\cdot, j)^\top 1]\mathbf{y} \le 1, \ j = 1, 2, \dots, M\right\}.$$

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

Remark A.1. Observe from the proof that an upper bound on the uniform Lipschitz constant can be easily computed by solving the following linear program:

$$\begin{aligned}\max_{\mathbf{y}\in\mathbb{R}^{N|\mathscr{S}|+1}} \quad & \mathbf{e}_{NS+1}^\top \mathbf{y} \\ \text{s.t.} \quad & [H^1(\cdot, j)^\top H^2(\cdot, j)^\top \cdots H^N(\cdot, j)^\top 1]\mathbf{y} \le 1, \quad j = 1, 2, \dots, M, \\ & \mathbf{y} \ge 0.\end{aligned}$$

Appendix B. Proof of Lemma 7.1. It is equivalent to prove that for all fixed $\mathbf{p}$, the function $\mathbf{q} \mapsto \rho(\mathbf{p}, H(\cdot, \mathbf{q}))$ is linear on the simplex. Actually, the proof exhibits a simpler expression for ρ.

To this end, we first group together the outcomes with same feedback signals and define a mapping

$$T\colon \mathscr{P}(\{1, \dots, M\}) \to \mathscr{P}(\{1, \dots, M\}),$$

where $\mathscr{P}(\{1, \dots, M\})$ is the set of all probability distributions $\mathbf{q}$ on the outcomes. Formally, consider the binary relation defined by $j \equiv j'$ if and only if $h(\cdot, j) = h(\cdot, j')$. (We use here the notation h to emphasize that we deal with deterministic feedback signals.) Denote by $F_1, \dots, F_{M'}$ the partition of the outcomes $\{1, \dots, M\}$ obtained so, and pick in every F_j the outcome y_j with minimal reward $r(\mathbf{p}, y_j)$ against $\mathbf{p}$ (ties can be broken arbitrarily, e.g., by choosing the outcome with lowest index). Then, for every $\mathbf{q}$, the distribution $\mathbf{q}' = T(\mathbf{q})$ is defined as $q'_{y_j} = \sum_{y \in F_j} q_y$ for $j = 1, \ldots, M'$, and $q'_k = 0$ if $k \neq y_j$ for all j.

T is a linear projection (i.e., $T \circ T = T$). It is easy to see that in the case of deterministic feedback signals, $H(\cdot, \mathbf{q}) = H(\cdot, \mathbf{q}')$ if and only if $T(\mathbf{q}) = T(\mathbf{q}')$. This implies that

$$\rho(\mathbf{p}, H(\cdot, \mathbf{q})) = \min_{\mathbf{q}' \colon T(\mathbf{q}') = T(\mathbf{q})} r(\mathbf{p}, \mathbf{q}') = r(\mathbf{p}, T(\mathbf{q})), \tag{B1}$$

where the last equality follows from the fact that, by choices of the y_j, $r(\mathbf{p}, \mathbf{q}') \geq r(\mathbf{p}, T(\mathbf{q}'))$ for all $\mathbf{q}'$, with equality for $\mathbf{q}' = T(\mathbf{q}) = T^2(\mathbf{q})$. By linearity of T, $\mathbf{q} \mapsto r(\mathbf{p}, T(\mathbf{q})) = \rho(\mathbf{p}, H(\cdot, \mathbf{q}))$ is therefore linear itself, as claimed.

Note that the equivalence of $H(\cdot, \mathbf{q}) = H(\cdot, \mathbf{q}')$ and $T(\mathbf{q}) = T(\mathbf{q}')$, together with (B1), implies the following sufficient condition for Hannan consistency (for necessary and sufficient conditions, see Piccolboni and Schindelhauer [25] and Cesa-Bianchi et al. [10]). It is more general than the distinguishing actions condition of Cesa-Bianchi et al. [10].

Remark B.1. Whenever H has no two identical columns in the case of deterministic feedback, i.e., $h(\cdot, j) \neq h(\cdot, j')$ for all $j \neq j'$, one has that for all $\mathbf{p}$ and $\mathbf{q}$,

$$\rho(\mathbf{p}, H(\cdot, \mathbf{q})) = r(\mathbf{p}, \mathbf{q}).$$

The condition is satisfied, for example, for multi-armed bandit problems, where $h = r$ (provided that we identify outcomes yielding the same rewards against all decision maker's actions).

Acknowledgments. The authors thank Tristan Tomala for helpful discussions. Shie Mannor was partially supported by the Canada Research Chairs Program and by the Natural Sciences and Engineering Research Council of Canada. Gábor Lugosi acknowledges the support of the Spanish Ministry of Science and Technology grant MTM2006-05650 and of Fundación BBVA. Gilles Stoltz was partially supported by the French "Agence Nationale pour la Recherche" under grant JCJC06-137444 "From applications to theory in learning and adaptive statistics." Gábor Lugosi and Gilles Stoltz acknowledge the PASCAL Network of Excellence under EC Grant 506778. An extended abstract of this paper appeared in the *Proceedings of the 20th Annual Conference on Learning Theory*, Springer, 2007.

References

[1] Auer, P., N. Cesa-Bianchi, C. Gentile. 2002. Adaptive and self-confident on-line learning algorithms. *J. Comput. System Sci.* **64** 48–75.
[2] Auer, P., N. Cesa-Bianchi, Y. Freund, R. Schapire. 2002. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* **32** 48–77.
[3] Azuma, K. 1967. Weighted sums of certain dependent random variables. *Tohoku Math. J.* **68** 357–367.
[4] Baños, A. 1968. On pseudo-games. *Ann. Math. Statist.* **39** 1932–1945.
[5] Bertsekas, D. P. 1995. *Nonlinear Programming*. Athena Scientific, Belmont, MA.
[6] Blackwell, D. 1956. Controlled random walks. *Proc. Internat. Congress of Mathematicians, 1954*, Vol. III. North-Holland, Amsterdam, 336–338.
[7] Cesa-Bianchi, N. 1999. Analysis of two gradient-based algorithms for on-line regression. *J. Comput. System Sci.* **59**(3) 392–411.
[8] Cesa-Bianchi, N., G. Lugosi. 1999. On prediction of individual sequences. *Ann. Statist.* **27** 1865–1895.
[9] Cesa-Bianchi, N., G. Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge University Press, New York.
[10] Cesa-Bianchi, N., G. Lugosi, G. Stoltz. 2006. Regret minimization under partial monitoring. *Math. Oper. Res.* **31** 562–580.
[11] Cesa-Bianchi, N., Y. Mansour, G. Stoltz. 2007. Improved second-order bounds in prediction with expert advice. *Machine Learning* **66** 321–352.
[12] Cesa-Bianchi, N., Y. Freund, D. Haussler, D. P. Helmbold, R. Schapire, M. Warmuth. 1997. How to use expert advice. *J. ACM* **44**(3) 427–485.
[13] Chen, X., H. White. 1996. Laws of large numbers for Hilbert space-valued mixingales with applications. *Econometric Theory* **12**(2) 284–304.
[14] Foster, D., R. Vohra. 1998. Asymptotic calibration. *Biometrika* **85** 379–390.
[15] Freedman, D. A. 1975. On tail probabilities for martingales. *Ann. Probab.* **3** 100–118.
[16] Hannan, J. 1957. Approximation to Bayes risk in repeated play. *Contributions Theory Games* **3** 97–139.
[17] Hart, S., A. Mas-Colell. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* **68** 1127–1150.
[18] Hart, S., A. Mas-Colell. 2002. A reinforcement procedure leading to correlated equilibrium. *Economic Essays*: *A Festschrift for Werner Hildenbrand*. Springer, New York, 181–200.
[19] Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.
[20] Kivinen, J., M. Warmuth. 1997. Exponentiated gradient versus gradient descent for linear predictors. *Inform. Comput.* **132**(1) 1–63.
[21] Littlestone, N., M. Warmuth. 1994. The weighted majority algorithm. *Inform. Comput.* **108** 212–261.
[22] Mannor, S., N. Shimkin. 2003. On-line learning with imperfect monitoring. *Proc. 16th Annual Conf. Learn. Theory*, Springer, New York, 552–567.

[23] Megiddo, N. 1980. On repeated games with incomplete information played by non-Bayesian players. *Internat. J. Game Theory* **9** 157–167.

[24] Mertens, J.-F., S. Sorin, S. Zamir. 1994. Repeated games. CORE discussion paper 9420, 9421, 9422, Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium.

[25] Piccolboni, A., C. Schindelhauer. 2001. Discrete prediction games with arbitrary feedback and loss. *Proc. 14th Annual Conf. Computational Learn. Theory*. Springer, New York, 208–223.

[26] Rustichini, A. 1999. Minimizing regret: The general case. *Games Econom. Behav.* **29** 224–243.

[27] Vovk, V. 1990. Aggregating strategies. *Proc. 3rd Annual Workshop on Computational Learn. Theory*. Morgan Kaufmann Publishers Inc., San Francisco, 372–383.

[28] Vovk, V. 1998. A game of prediction with expert advice. *J. Comput. System Sci.* **56**(2) 153–173.

[29] Weissman, T., N. Merhav. 2001. Universal prediction of binary individual sequences in the presence of noise. *IEEE Trans. Inform. Theory* **47** 2151–2173.

[30] Weissman, T., N. Merhav, A. Somekh-Baruch. 2001. Twofold universal prediction schemes for achieving the finite state predictability of a noisy individual binary sequence. *IEEE Trans. Inform. Theory* **47** 1849–1866.