Regret minimization under partial monitoring

Nicolo Cesa-Bianchi, Gabor Lugosi, Gilles Stoltz

Regret Minimization Under Partial Monitoring *

Nicolò Cesa-Bianchi
DSI, Università di Milano
via Comelico 39, 20135 Milano, Italy
cesa-bianchi@dsi.unimi.it

Gábor Lugosi
Department of Economics,
Pompeu Fabra University
08005 Barcelona, Spain
lugosi@upf.es

Gilles Stoltz
Département de Mathématiques et Applications,
Ecole Normale Supérieure,
45, rue d'Ulm, 75005 Paris, France
gilles.stoltz@ens.fr

April 4, 2005

Abstract

We consider repeated games in which the player, instead of observing the action chosen by the opponent in each game round, receives a feedback generated by the combined choice of the two players. We study Hannan consistent players for these games, that is, randomized playing strategies whose per-round regret vanishes with probability one as the number n of game rounds goes to infinity. We prove a general lower bound of $\Omega(n^{-1/3})$ for the convergence rate of the regret, and exhibit a specific strategy that attains this rate for any game for which a Hannan consistent player exists.

*The first two authors acknowledge support by the PASCAL Network of Excellence under EC grant no. 506778. The work of the second author was supported by the Spanish Ministry of Science and Technology and FEDER, grant BMF2003-03324. Part of this work was done wile the third co-author was visiting Pompeu Fabra University.

1 A motivating example

A simple yet nontrivial example of partial monitoring is the following dynamic pricing problem. A vendor sells a product to a sequence of customers whom he attends one by one. To each customer, the seller offers the product at a price he selects, say, from the interval $[0, 1]$. The customer then decides to buy the product or not. No bargaining is possible and no other information is exchanged between buyer and seller. The goal of the seller is to achieve an income almost as large as if he knew the maximal price each customer is willing to pay for the product. Thus, if the price offered to the t-th customer is p_t and the highest price this customer is willing to pay is $y_t \in [0, 1]$, then the loss of the seller is $y_t - p_t$ if the product is sold and (say) a constant $c > 0$ if the product is not sold. Formally, the loss of the vendor at time t is

$$\ell(p_t, y_t) = (y_t - p_t)\mathbb{I}_{p_t \le y_t} + c\,\mathbb{I}_{p_t > y_t}$$

where $c \in [0, 1]$. (In another version of the problem the constant c may be replaced by y_t. In this case it is easy to see that all terms depending on y_t cancel out when considering the regret, and we obtain the bandit setting analyzed by Kleinberg and Leighton [28]—see below.) In either case, if the seller knew in advance the empirical distribution of the y_t's then he could set a constant price $q \in [0, 1]$ which minimizes his overall loss. A natural question is whether there exists a randomized strategy for the seller such that his average regret

$$\frac{1}{n}\sum_{t=1}^{n}\ell(p_t, y_t) - \min_{q\in[0,1]}\frac{1}{n}\sum_{t=1}^{n}\ell(q, y_t)$$

is guaranteed to converge to zero as $n \to \infty$ regardless of the sequence $y_1, y_2, \ldots$ of prices. The difficulty in this problem is that the only information the seller (i.e., the forecaster) has access to is whether $p_t > y_t$ but neither y_t nor $\ell(p_t, y_t)$ are revealed. One of the main results of this paper describes a simple strategy such that the average regret defined above is of the order of $n^{-1/5}$.

We treat such limited-feedback (or *partial monitoring*) prediction problems in a more general framework which we describe next. The dynamic pricing problem described above, which is a special case of this more general framework, has been also investigated by Kleinberg and Leighton [28] in a simpler setting where the reward of the seller is defined as $\rho(p_t, y_t) = p_t\,\mathbb{I}_{p_t \le y_t}$. Note that, by using the feedback information (i.e., whether the customer bought the product or not), here the seller can compute the value of $\rho(p_t, y_t)$. Therefore, their game reduces to an instance of the multi-armed bandit game (see Example 1 below) with a continuous action space.

2 Main definitions

We adopt a learning-theoretic viewpoint and describe partial monitoring as a repeated prediction game between a *forecaster* (the player) and the *environment* (the opponent). In the same spirit, we call *outcomes* the actions taken by the environment. At each round $t = 1, 2\ldots$ of the game, the forecaster chooses an action I_t from the set $\{1, \ldots, N\}$, and the environment chooses an action y_t from the set $\{1, \ldots, M\}$. The losses of the forecaster are summarized in the *loss matrix*

PREDICTION WITH PARTIAL MONITORING

Parameters: number of actions N, number of outcomes M, loss function ℓ, feedback function h.

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

(1) the environment chooses the next outcome $y_t \in \{1, \ldots, M\}$ 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 incurs loss $\ell(I_t, y_t)$ and each action i incurs loss $\ell(i, y_t)$, where none of these values is revealed to the forecaster;

(4) the feedback $h(I_t, y_t)$ is revealed to the forecaster.

$\mathbf{L} = [\ell(i,j)]_{N\times M}$. (This matrix is assumed to be known by the forecaster.) Without loss of generality, we rescale the losses so that they all lie in $[0, 1]$. If, at time t, the forecaster chooses an action $I_t \in \{1, \ldots, N\}$ and the outcome is $y_t \in \{1, \ldots, M\}$, then the forecaster's suffers loss $\ell(I_t, y_t)$. However, instead of the outcome y_t, the forecaster only observes the feedback $h(I_t, y_t)$, where h is a known *feedback function* that assigns, to each action/outcome pair in $\{1, \ldots, N\} \times \{1, \ldots, M\}$ an element of a finite set $\mathcal{S} = \{s_1, \ldots, s_m\}$ of *signals*. The values of h are collected in a *feedback matrix* $\mathbf{H} = [h(i,j)]_{N\times M}$.

Note that we do not make any restrictive assumption on the power of the opponent. The environment may choose action y_t at time t by considering the whole past, that is, the whole sequence of action/outcome pairs (I_s, y_s), $s = 1, \ldots, t-1$. Without loss of generality, we assume that the opponent uses a deterministic strategy, so that the value of y_t is fixed by the sequence $(I_1, \ldots, I_{t-1})$. In comparison, the forecaster has access to significantly less information, since he only knows the sequence of past feedbacks, $(h(I_1, y_1), \ldots, h(I_{t-1}, y_{t-1}))$.

We note here that some authors consider a more general setup in which the feedback may be random. For the sake of clarity we treat the simpler model described above and return to the more general case in Section 7.

It is an interesting and complex problem to investigate the possibilities of a predictor only supplied with the limited information of the feedback. In this paper we focus on the average regret

$$\frac{1}{n}\sum_{t=1}^{n} \ell(I_t, y_t) - \min_{i=1,\ldots,N} \frac{1}{n}\sum_{t=1}^{n} \ell(i, y_t) ,$$

that is, the difference between the average (per-round) loss of the forecaster and the average (per-round) loss of the best action. Forecasting strategies guaranteeing that the average regret converges to zero almost surely for all possible strategies of the environment are called *Hannan consistent* after James Hannan, who first proved the existence of a Hannan consistent strategy in the *full information* case [21] when $h(i,j) = j$ for all i,j (i.e., when the true outcome y_t is revealed to the forecaster after taking an action). The full information case has been studied extensively in the theory of repeated games, and in the fields of learning theory and information theory. A few key references and surveys include Blackwell [6], Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth [8], Cesa-Bianchi and Lugosi [9], Feder, Merhav, and Gutman [14], Foster and Vohra [18], Hart and Mas-Colell [23], Littlestone and Warmuth [29], Merhav and Feder [32], and Vovk [38, 37].

A natural question one may ask is under what conditions on the loss and feedback matrices it is possible to achieve Hannan consistency, that is, to guarantee that, asymptotically, the cumulative loss of the forecaster is not larger than that of the best constant action with probability one. Naturally, this depends on the relationship between the loss and feedback functions. An initial answer to this question has been provided by the work of Piccolboni and Schindelhauer [34]. However, since they are only concerned with expected performance, their results do not imply Hannan consistency. In addition, their bounds have suboptimal rates of convergence. Below, we extend those results by showing a forecaster that achieves Hannan consistency with optimal convergence rates.

Note that the forecaster is free to encode the values $h(i,j)$ of the feedback function by real numbers. The only restriction is that if $h(i,j) = h(i,j')$ then the corresponding real numbers should also coincide. To avoid ambiguities by trivial rescaling, we assume that $|h(i,j)| \leq 1$ for all pairs (i,j). Thus, in the sequel we assume that $\mathbf{H} = [h(i,j)]_{N\times M}$ is a matrix of real numbers between -1 and 1 and keep in mind that the forecaster may replace this matrix by $\mathbf{H}_\phi = [\phi_i(h(i,j))]_{N\times M}$ for arbitrary functions $\phi_i : [-1,1] \to [-1,1]$, $i = 1,\ldots,N$. Note that the set $\mathcal{S}$ of signals may be chosen such that it has $m \leq M$ elements, though after numerical encoding the matrix may have as many as MN distinct elements.

The problem of partial monitoring was considered by Mertens, Sorin, and Zamir [33], Rustichini [35], Piccolboni, and Schindelhauer [34], and Mannor and Shimkin [30]. The forecaster strategy studied in Section 3 is first introduced in [34], where its expected regret is shown to have a sub-linear growth. Rustichini [35] and Mannor and Shimkin [30] consider a more general setup in which the feedback is not necessarily a deterministic function of the pair outcome and forecaster's action, but it may be random with a distribution indexed by this pair. Based on Blackwell's approachability theorem, Rustichini [35] establishes a general existence result for strategies with asymptotically optimal performance in this more general framework. In this paper we answer Rustichini's question about the fastest achievable rate of convergence in the case when Hannan consistent strategies exist. Mannor and Shimkin also consider cases when Hannan consistency may not be achieved, give a partial solution, and point out important difficulties in such cases.

Before introducing a general prediction strategy and sufficient conditions for its Hannan consistency, we describe a few concrete examples of partial monitoring problems.

Example 1 (*Multi-armed bandit problem.*) A well-studied special case of the partial monitoring prediction problem is the so-called multi-armed bandit problem. Here the forecaster, after taking an action, is able to measure his loss (or reward) but does not have access to what would have happened had he chosen another possible action. Here $\mathbf{H} = \mathbf{L}$, that is, the feedback received by the forecaster is just his own loss. This problem has been widely studied both in a stochastic and in a worst-case setting. The worst-case or adversarial setting considered in this paper was first investigated by Baños [5] (see also Megiddo [31]). Hannan consistent strategies were constructed by Foster and Vohra [17], Auer, Cesa-Bianchi, Freund, and Schapire [2], and Hart and Mas Colell [22, 24] (see also Fudenberg and Levine [20]). Auer, Cesa-Bianchi, Freund, and Schapire [2] (see also Auer [1] and the refined analysis of Cesa-Bianchi and Lugosi [11]) define a strategy that guarantees a rate of convergence of the order $O(\sqrt{N(\log N)/n})$ for the regret, which is optimal up to the logarithmic factor.

Example 2 (*Dynamic pricing.*) Consider the dynamic pricing problem described in the introduction of the section under the additional restriction that all prices take their values from the finite set $\{0, 1/N, \ldots, (N-1)/N\}$ where N is a positive integer (see Example 6 for a non-discretized version). Clearly, if N is sufficiently large, this discrete version approximates arbitrarily the original problem. Now one may take $M = N$ and the loss matrix is

$$\mathbf{L} = [\ell(i,j)]_{N\times N} \qquad \text{where} \qquad \ell(i,j) = \frac{j-i}{N}\mathbb{I}_{i\le j} + c\,\mathbb{I}_{i>j}\ .$$

The information the forecaster (i.e., the vendor) receives is simply whether the predicted value I_t is greater than the outcome y_t or not. Thus, the entries of the feedback matrix $\mathbf{H}$ may be taken to be $h(i,j) = \mathbb{I}_{i>j}$ or, after an appropriate re-encoding,

$$h(i,j) = a\,\mathbb{I}_{i\le j} + b\,\mathbb{I}_{i>j} \quad i,j = 1,\ldots,N$$

where a and b are constants chosen by the forecaster satisfying $a, b \in [-1, 1]$.

Example 3 (*Apple tasting.*) This problem was first considered by Helmbold, Littlestone, and Long [26] in a somewhat more restrictive setting. In this example $N = M = 2$ and the loss and feedback matrices are given by

$$\mathbf{L} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \text{and} \qquad \mathbf{H} = \begin{bmatrix} a & a \\ b & c \end{bmatrix}\ .$$

Thus, the forecaster only receives feedback about the outcome y_t when he chooses the first action. (Imagine that apples are to be classified as “good for sale” or “rotten”. An apple classified as “rotten” may be opened to check whether its classification was correct. On the other hand, since apples that have been checked cannot be put on sale, an apple classified “good for sale” is never checked.)

Example 4 (*Label efficient prediction.*) In the problem of label efficient prediction (see Helmbold and Panizza [25] and also Cesa-Bianchi, Lugosi, and Stoltz [12]) the forecaster, after choosing its prediction for round t, decides whether to query the outcome y_t, which he can only do for a limited number of times. In [12] matching upper and lower bounds are given for the regret in terms of the number of available labels, total number of rounds, and number of actions. A variant of the label efficient prediction problem may also be cast as a partial monitoring problem. Let $N = 3$, $M = 2$, and consider loss and feedback matrices of the form

$$\mathbf{L} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{H} = \begin{bmatrix} a & b \\ c & c \\ c & c \end{bmatrix} .$$

In this example the only times useful feedback is received are when the first action is played but in this case a maximal loss is incurred regardless of the outcome. Thus, just like in the problem of label efficient prediction, playing the "informative" action has to be limited, otherwise there is no hope for Hannan consistency.

3 General upper bounds on the regret

The purpose of this section is to derive general upper bounds for the rate of convergence of the regret achievable under partial monitoring. This will be done by analyzing a forecasting strategy inspired by Piccolboni and Schindelhauer [34]. This strategy is based on the exponentially weighted average forecaster, a thoroughly studied predictor in the full information case, see, for example, Auer, Cesa-Bianchi, and Gentile [3], Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth [8], Littlestone and Warmuth [29], Vovk [38, 37]. In the special case of the multi-armed bandit problem, the forecaster reduces to the strategy of Auer, Cesa-Bianchi, Freund, and Schapire [2] (see also Hart and Mas-Colell [24] for a closely related method).

The crucial assumption under which the strategy is defined is that there exists an $N \times N$ matrix $\mathbf{K} = [k(i,j)]_{N \times N}$ such that

$$\mathbf{L} = \mathbf{K}\,\mathbf{H} ,$$

that is,

$$\mathbf{H} \quad \text{and} \quad \begin{bmatrix} \mathbf{H} \\ \mathbf{L} \end{bmatrix}$$

have the same rank. In other words we may write, for all $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, M\}$,

$$\ell(i,j) = \sum_{l=1}^{N} k(i,l)\, h(l,j) .$$

In this case one may define the estimated losses $\widetilde{\ell}$ by

$$\widetilde{\ell}(i, y_t) = \frac{k(i, I_t)\, h(I_t, y_t)}{p_{I_t,t}} , \quad i = 1, \ldots, N . \tag{1}$$

(Note that the estimates proposed above are real-valued, and may even be negative.) We denote the cumulative estimated losses at round t and for action i by $\widetilde{L}_{i,t} = \sum_{s=1}^{t} \widetilde{\ell}(i, y_t)$.

Consider the forecaster defined in Figure 1, where k^* is defined in Theorem 1. Roughly speaking, the two terms in the expression of $p_{i,t}$ correspond to "exploitation" and "exploration".

Parameters: matrix $\mathbf{L}$ of losses, feedback matrix $\mathbf{H}$, matrix $\mathbf{K}$ such that $\mathbf{L} = \mathbf{K}\,\mathbf{H}$
Initialization: $\widetilde{L}_{1,0} = \cdots = \widetilde{L}_{N,0} = 0$.
For each round $t = 1, 2, \ldots$

(1) let $\eta_t = (k^*)^{-2/3}((\ln N)/N)^{2/3} t^{-2/3}$ and $\gamma_t = (k^*)^{2/3} N^{2/3} (\ln N)^{1/3} t^{-1/3}$;

(2) choose an action I_t from the set of actions $\{1, \ldots, N\}$ at random, according to the distribution $\boldsymbol{p}_t$ defined by

$$p_{i,t} = (1 - \gamma_t) \frac{e^{-\eta_t \widetilde{L}_{i,t-1}}}{\sum_{k=1}^N e^{-\eta_t \widetilde{L}_{k,t-1}}} + \frac{\gamma_t}{N} \ ;$$

(3) let $\widetilde{L}_{i,t} = \widetilde{L}_{i,t-1} + \widetilde{\ell}(i, y_t)$ for all $i = 1, \ldots, N$.

Figure 1: The randomized forecaster for prediction under partial monitoring.

The first term assigns exponentially decreasing weights to the actions depending on their estimated cumulative losses, while the second term ensures sufficient exploration to guarantee accurate estimates of the losses.

A key property of the loss estimates is their unbiasedness in the following sense. Denoting by $\mathbb{E}_t$ the conditional expectation given $I_1, \ldots, I_{t-1}$ (i.e., the expectation with respect to the distribution $\boldsymbol{p}_t$ of the random variable I_t), observe that this conditioning fixes the value of y_t, and thus,

$$\begin{aligned} \mathbb{E}_t \widetilde{\ell}(i, y_t) &= \sum_{k=1}^N \frac{k(i,k)\, h(k, y_t)}{p_{k,t}} p_{k,t} \\ &= \sum_{k=1}^N k(i,k)\, h(k, y_t) = \ell(i, y_t) \ , \quad i = 1, \ldots, N \ , \end{aligned}$$

and therefore $\widetilde{\ell}(i, y_t)$ is an unbiased estimate of the loss $\ell(i, y_t)$.

The main performance bound of this section is summarized in the next theorem. Note that the average regret

$$\frac{1}{n} \left(\sum_{t=1}^n \ell(I_t, y_t) - \min_{i=1,\ldots,N} \sum_{t=1}^n \ell(i, y_t) \right)$$

decreases to zero at a rate $n^{-1/3}$. This is significantly slower than the best rate $n^{-1/2}$ obtained in the "full information" case. In the next section we show that this rate cannot be improved in general. Thus, the price paid for having access only to some feedback except for the actual outcomes is the deterioration in the rate of convergence. However, Hannan consistency is still achievable whenever the conditions of the theorem are satisfied.

Theorem 1 *Consider any partial monitoring problem such that the loss and feedback matrices satisfy* $\mathbf{L} = \mathbf{K}\,\mathbf{H}$ *for some* $N \times N$ *matrix* $\mathbf{K}$ *with* $k^* = \max\{1, \max_{i,j} |k(i,j)|\}$*, and consider the forecaster of Figure 1. Let* $\delta \in (0,1)$*. Then, for all strategies of the opponent, for all* n*, with probability at least* $1-\delta$*,*

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} \ell(I_t, y_t) - \min_{i=1,\dots,N} \frac{1}{n}\sum_{t=1}^{n} \ell(i, y_t) \\
&\leq \; 5\left(\frac{(k^*N)^2 \ln N}{n}\right)^{1/3}\left(1 + \sqrt{\frac{3}{2}\frac{\ln((N+4)/\delta)}{\ln N}}\right) \\
&\quad + \sqrt{\frac{1}{2n}\ln\frac{N+4}{\delta}} \; + \; 5(k^*N)^{4/3} n^{-2/3} (\ln N)^{-1/3} \ln\frac{N+4}{\delta} \\
&\quad + \frac{1}{n}\left(1 + ((k^*N)^2 \ln N)^{1/3} + k^*N\right)\ln\frac{N+4}{\delta} \; .
\end{aligned}$$

The main term in the performance bound has the order of magnitude $n^{-1/3}(k^*N)^{2/3}(\ln N)^{1/3}$. Observe that this theorem directly implies Hannan consistency, by a simple application of the Borel-Cantelli lemma.

Proof. The starting point of the proof of the theorem is an application of Theorem 5 (shown in the Appendix) to the estimated losses. Since $\widetilde{\ell}_{i,t}$ lies between $-B_t$ and B_t, where $B_t = k^*N/\gamma_t$, the proposed values of γ_t and η_t imply that $\eta_t B_t \leqslant 1$ if and only if $t \geqslant (\ln N)/(Nk^*)$, that is, for all $t \geqslant 1$. Therefore, defining for $t = 1, \dots, n$, the probability vector $\widetilde{\boldsymbol{p}}_t$ by its components

$$\widetilde{p}_{i,t} = \frac{e^{-\eta_t \widetilde{L}_{i,t-1}}}{\sum_{k=1}^{N} e^{-\eta_t \widetilde{L}_{k,t-1}}} \qquad i = 1, \dots, N \; ,$$

we may apply Theorem 5 to obtain

$$\sum_{t=1}^{n}\sum_{i=1}^{N} \widetilde{p}_{i,t}\widetilde{\ell}(i, y_t) - \min_{j=1,\dots,N} \widetilde{L}_{j,n} \leqslant \frac{2\ln N}{\eta_{n+1}} + \sum_{t=1}^{n} \eta_t \sum_{i=1}^{N} \widetilde{p}_{i,t}\widetilde{\ell}(i, y_t)^2 \; .$$

Since $p_{i,t} = (1-\gamma_t)\widetilde{p}_{i,t} + \gamma_t/N$, the inequality above yields, after some simple bounding,

$$\sum_{t=1}^{n}\sum_{i=1}^{N} p_{i,t}\widetilde{\ell}(i, y_t) - \min_{j=1,\dots,N} \widetilde{L}_{j,n} \leqslant \frac{2\ln N}{\eta_{n+1}} + \sum_{t=1}^{n} \eta_t \sum_{i=1}^{N} \widetilde{p}_{i,t}\widetilde{\ell}(i, y_t)^2 + \sum_{t=1}^{n} \gamma_t \sum_{i=1}^{N} \frac{1}{N}\widetilde{\ell}(i, y_t) \; . \quad (2)$$

Introduce the notation

$$\widehat{L}_n = \sum_{t=1}^{n} \ell(I_t, y_t) \quad \text{and} \quad L_{j,n} = \sum_{t=1}^{n} \ell(j, y_t), \;\; j = 1, \dots, N \; .$$

Next we show that, with an overwhelming probability, the right-hand side of the inequality (2) is less than something of the order $n^{2/3}$, and that the left-hand side is close to the actual regret

$$\sum_{t=1}^{n} \ell(I_t, y_t) - \min_{j=1,\dots,N} L_{j,n} \; .$$

Our main tool is Bernstein's inequality for martingales, see Lemma 7 in the Appendix. This inequality implies the following four lemmas, whose proofs are similar, so we omit some of them.

Lemma 1 *With probability at least* $1-\delta/(N+4)$,

$$\sum_{t=1}^{n}\sum_{i=1}^{N} p_{i,t}\ell(i,y_t) \leqslant \sum_{t=1}^{n}\sum_{i=1}^{N} p_{i,t}\widetilde{\ell}(i,y_t) + \sqrt{2(k^*N)^2\left(\sum_{t=1}^{n}\frac{1}{\gamma_t}\right)\ln\frac{N+4}{\delta}} + \frac{\sqrt{2}}{3}\left(1+\frac{k^*N}{\gamma_n}\right)\ln\frac{N+4}{\delta}\ .$$

Proof. Define $Z_t = -\sum_{i=1}^{N} p_{i,t}\widetilde{\ell}(i,y_t)$ so that $\mathbb{E}_t[Z_t] = -\sum_{i=1}^{N} p_{i,t}\ell(i,y_t)$, and consider $X_t = Z_t - \mathbb{E}_t[Z_t]$. We note that

$$\begin{aligned}\mathbb{E}_t[X_t^2] \leqslant \mathbb{E}_t[Z_t^2] &= \sum_{i,j} p_{i,t}p_{j,t}\mathbb{E}_t\left[\widetilde{\ell}(i,y_t)\widetilde{\ell}(j,y_t)\right] \\ &= \sum_{i,j} p_{i,t}p_{j,t}\sum_{k=1}^{N} p_{k,t}\frac{k(i,k)k(j,k)h(k,y_t)^2}{p_{k,t}^2} \leqslant \frac{(k^*N)^2}{\gamma_t}\ ,\end{aligned}$$

and therefore,

$$V_n = \sum_{t=1}^{n}\mathbb{E}_t[X_t^2] \leqslant (k^*N)^2\sum_{t=1}^{n}\frac{1}{\gamma_t}\ .$$

On the other hand, $|X_t|$ is bounded by $K = 1+(k^*N)/\gamma_n$. Bernstein's inequality (see Lemma 7) thus concludes the proof. ∎

Lemma 2 *For each fixed* j*, with probability at least* $1-\delta/(N+4)$,

$$\widetilde{L}_{j,n} \leqslant L_{j,n} + + \sqrt{2(k^*N)^2\left(\sum_{t=1}^{n}\frac{1}{\gamma_t}\right)\ln\frac{N+4}{\delta}} + \frac{\sqrt{2}}{3}\left(1+\frac{k^*N}{\gamma_n}\right)\ln\frac{N+4}{\delta}\ .$$

Lemma 3 *With probability at least* $1-\delta/(N+4)$,

$$\sum_{t=1}^{n}\eta_t\sum_{i=1}^{N}\widetilde{p}_{i,t}\widetilde{\ell}(i,y_t)^2 \leqslant \sum_{t=1}^{n}\eta_t\frac{(k^*N)^2}{\gamma_t} + \sqrt{2(k^*N)^4\left(\sum_{t=1}^{n}\frac{\eta_t^2}{\gamma_t^3}\right)\ln\frac{N+4}{\delta}} + \frac{\sqrt{2}}{3}\ln\frac{N+4}{\delta}\ .$$

Proof. Let $Z_t = \eta_t\sum_{i=1}^{N}\widetilde{p}_{i,t}\widetilde{\ell}(i,y_t)^2$, and $X_t = Z_t - \mathbb{E}_t[Z_t]$. All $|X_t|$ are bounded by

$$K = \max_{t=1,\dots,n}\eta_t\frac{(k^*N)^2}{\gamma_t^2} = 1\ .$$

On the other hand,

$$V_n = \sum_{t=1}^{n} \mathbb{E}_t[X_t^2] \leqslant (k^* N)^4 \sum_{t=1}^{n} \frac{\eta_t^2}{\gamma_t^3} \,.$$

Bernstein's inequality (see Lemma 7) now concludes the proof, together with the inequality

$$\mathbb{E}_t[Z_t] \leqslant \eta_t \frac{(k^* N)^2}{\gamma_t} \,.$$

■

Lemma 4 *With probability at least* $1 - \delta/(N+4)$,

$$\sum_{t=1}^{n} \gamma_t \sum_{i=1}^{N} \frac{1}{N} \widetilde{\ell}(i, y_t) \leqslant \sum_{t=1}^{n} \gamma_t + \sqrt{2(k^* N)^2 \left(\sum_{t=1}^{n} \gamma_t \right) \ln \frac{N+4}{\delta}} + \frac{\sqrt{2}}{3} \left(k^* N + \gamma_1 \right) \ln \frac{N+4}{\delta} \,.$$

The next lemma is an easy consequence of the Hoeffding-Azuma inequality for sums of bounded martingale differences (see Hoeffding [27], Azuma [4]).

Lemma 5 *With probability at least* $1 - \delta/(N+3)$,

$$\sum_{t=1}^{n} \ell(I_t, y_t) \leqslant \sum_{t=1}^{n} \sum_{i=1}^{N} p_{i,t} \ell(i, y_t) + \sqrt{\frac{n}{2} \ln \frac{N+4}{\delta}} \,.$$

The proof of the main result follows now from a combination of Lemmas 1 to 5 with (2) (where Lemma 2 is applied N times). Using a union-of-event bound, we see that, with probability $1 - \delta$,

$$\begin{aligned}
&\sum_{t=1}^{n} \ell(I_t, y_t) - \min_{j=1,\ldots,N} L_{j,n} \\
&\leqslant \frac{2 \ln N}{\eta_{n+1}} \\
&\quad + 2 \left(\sqrt{2(k^* N)^2 \left(\sum_{t=1}^{n} \frac{1}{\gamma_t} \right) \ln \frac{N+4}{\delta}} + \frac{\sqrt{2}}{3} \left(1 + \frac{k^* N}{\gamma_n} \right) \ln \frac{N+4}{\delta} \right) \\
&\quad + \sum_{t=1}^{n} \eta_t \frac{(k^* N)^2}{\gamma_t} + \sqrt{2(k^* N)^4 \left(\sum_{t=1}^{n} \frac{\eta_t^2}{\gamma_t^3} \right) \ln \frac{N+4}{\delta}} + \frac{\sqrt{2}}{3} \ln \frac{N+4}{\delta} \\
&\quad + \sum_{t=1}^{n} \gamma_t + \sqrt{2(k^* N)^2 \left(\sum_{t=1}^{n} \gamma_t \right) \ln \frac{N+4}{\delta}} + \frac{\sqrt{2}}{3} \left(k^* N + \gamma_1 \right) \ln \frac{N+4}{\delta} \\
&\quad + \sqrt{\frac{n}{2} \ln \frac{N+4}{\delta}} \,.
\end{aligned}$$

Substituting the proposed values of γ_t and η_t, and using that for $-1 < \alpha \leqslant 0$

$$\sum_{t=1}^{n} t^{\alpha} \leqslant \frac{1}{\alpha+1} n^{\alpha+1} ,$$

we obtain the claimed result with a simple calculation. ∎

We close this section by considering the implications of Theorem 1 to the special cases mentioned in the introduction.

Example 5 (*Multi-armed bandit problem.*) Recall that in the case of the multi-armed bandit problem $\mathbf{H} = \mathbf{L}$ and the condition of the theorem is trivially satisfied. Indeed, one may take $\mathbf{K}$ to be the identity matrix so that $k^* = 1$. Thus, Theorem 1 implies a bound of the order of $((N^2 \ln N)/n)^{1/3}$. Even though, as it is shown in the next section, the rate $O(n^{-1/3})$ cannot be improved in general, faster rates of convergence are achievable for the special case of the bandit problem. Indeed, for the bandit problem Auer, Cesa-Bianchi, Freund, and Schapire [2], Auer [1], and Cesa-Bianchi and Lugosi [11] describe careful modifications of the forecaster of Theorem 1 that achieves an upper bound of the order of $\sqrt{N(\ln N)/n}$. It remains a challenging problem to characterize the class of problems that admit rates of convergence faster than $O(n^{-1/3})$.

Example 6 (*Dynamic pricing.*) In the discretized version of the dynamic pricing problem (i.e., when all prices are restricted to the set $\{0, 1/N, \ldots, (N-1)/N\}$), the feedback matrix is given by $h(i,j) = a\,\mathbb{I}_{i \le j} + b\,\mathbb{I}_{i>j}$ for some arbitrarily chosen values of a and b. By choosing, for example, $a = 1$ and $b = 0$, it is clear that $\mathbf{H}$ is an invertible matrix and therefore one may choose $\mathbf{K} = \mathbf{L}\,\mathbf{H}^{-1}$ and obtain a Hannan-consistent strategy with average regret of the order of $n^{-1/3}$. Thus, the seller has a way of selecting the prices I_t such that his loss is not much larger than what he could have achieved had he known the values y_t of all costumers and offered the best constant price. Note that with this choice of a and b, the value of k^* equals 1 (i.e., does not depend on N) and therefore the upper bound has the form $C((N^2 \log N)/n)^{1/3}\sqrt{\ln(1/\delta)}$ for some constant C. By choosing $N \approx n^{1/5}$ and running the forecaster into stages of doubling lengths the effect of discretization decreases at about the same rate as the average regret, and for the original problem with unrestricted price range one may obtain a regret bound of the form

$$\frac{1}{n}\sum_{t=1}^{n} \ell(p_t, y_t) - \min_{q \in [0,1]} \frac{1}{n}\sum_{t=1}^{n} \ell(q, y_t) = O(n^{-1/5} \ln n) .$$

We leave out the simple but tedious details of the proof. We simply note here that the discretization to N prices is done by the mapping y_t to $Y_N(y_t) = \lfloor N y_t \rfloor / N$.

Example 7 (*Apple tasting.*) In the apple tasting problem described above, one may choose the feedback values $a = b = 1$ and $c = 0$. Then, the feedback matrix is invertible and, once again, Theorem 1 applies.

Example 8 (*Label efficient prediction.*) Recall next the variant of the label efficient prediction problem described in the previous section. Here the rank of $\mathbf{L}$ equals two, so it is necessary (and sufficient) to encode the feedback matrix such that its rank equals two. One possibility is to choose $a = 1$, $b = 1/2$, and $c = 1/4$. Then we have $\mathbf{L} = \mathbf{K}\,\mathbf{H}$ for

$$\mathbf{K} = \begin{bmatrix} 0 & 2 & 2 \\ 2 & -2 & -2 \\ -2 & 4 & 4 \end{bmatrix} .$$

The obtained rate of convergence $O(n^{-1/3})$ may be shown to be optimal. In fact, it is this example that we use in Section 5 to show that this rate of convergence cannot be improved in general.

Remark 1 It is interesting to point out that the bound of Theorem 1 does not depend explicitly on the value of the cardinality M of the set of outcomes. Of course, in some problems the value k^* may depend on M. However, in some important special cases, such as the multi-armed bandit problem for which $k^* = 1$, this value is independent of M. In such cases the result extends easily to an infinite set of outcomes. In particular, the case when the loss matrix may change with time can be encoded this way.

4 Other regret-minimizing strategies

In the previous section we saw a forecasting strategy that guarantees that the average regret is of the order of $n^{-1/3}$ whenever the loss matrix $\mathbf{L}$ can be expressed as $\mathbf{K}\,\mathbf{H}$ for some matrix $\mathbf{K}$. In this section we discuss some alternative strategies that yield small regret under different conditions.

First note that it is not true that the existence of a Hannan consistent predictor is guaranteed if and only the loss matrix $\mathbf{L}$ can be expressed as $\mathbf{K}\,\mathbf{H}$. The following example describes such a situation.

Example 9 Let $N = M = 3$,

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{H} = \begin{bmatrix} a & b & c \\ d & d & d \\ e & e & e \end{bmatrix} .$$

Clearly, for all choices of the numbers a, b, c, d, e, the rank of the feedback matrix is at most two and therefore there is no matrix $\mathbf{K}$ for which $\mathbf{L} = \mathbf{K}\,\mathbf{H}$. However, note that whenever the first action is played, the forecaster has full information about the outcome y_t. Formally, an action $i \in \{1, \ldots, N\}$ is said to be *revealing* for a feedback matrix $\mathbf{H}$ if all entries in the i-th row of $\mathbf{H}$ are different. Below we prove the existence of a Hannan consistent forecaster for all problems in which there exists a revealing action.

Theorem 2 *Consider an arbitrary partial monitoring problem* $(\mathbf{L}, \mathbf{H})$ *such that* $\mathbf{L}$ *has a revealing action. Let* $\delta \in (0, 1)$. *If the randomized forecasting strategy of Figure 2 is run with parameters*

$$\varepsilon = \max\left\{0, \frac{m - \sqrt{2m\ln(4/\delta)}}{n}\right\} \qquad \text{and} \qquad \eta = \sqrt{\frac{2\varepsilon \ln N}{n}}$$

Parameters: $0 \le \varepsilon \le 1$ and $\eta > 0$. Action r is revealing.
Initialization: $w_{1,0} = \cdots = w_{N,0} = 1$.
For each round $t = 1, 2, \ldots$

(1) draw an action J_t from $\{1, \ldots, N\}$ according to the distribution

$$p_{i,t} = \frac{w_{i,t-1}}{\sum_{j=1}^{N} w_{j,t-1}} \ , \qquad i = 1, \ldots, N \ ,$$

(2) draw a Bernoulli random variable Z_t such that $\mathbb{P}[Z_t = 1] = \varepsilon$;

(3) if $Z_t = 1$ then play a revealing action, $I_t = r$, observe y_t, and compute

$$w_{i,t} = w_{i,t-1} e^{-\eta\, \ell(i,y_t)/\varepsilon} \qquad \text{for each } i = 1, \ldots, N \ ;$$

(4) otherwise, if $Z_t = 0$, play $I_t = J_t$ and let $w_{i,t} = w_{i,t-i}$ for each $i = 1, \ldots, N$.

Figure 2: The randomized forecaster for feedback matrices with a revealing action.

where $m = (4n)^{2/3}(\ln(4N/\delta))^{1/3}$*, then*

$$\frac{1}{n}\left(\sum_{t=1}^{n} \ell(I_t, y_t) - \min_{i=1,\ldots,N} L_{1,n}\right) \le 8n^{-1/3}\left(\ln\frac{4N}{\delta}\right)^{1/3}$$

holds with probability at least $1-\delta$ *for any strategy of the opponent.*

Proof. The forecaster of Figure 2 chooses at each round a revealing action with a small probability $\varepsilon \approx m/n$ (of the order of $n^{-1/3}$). At these m stages where a revealing action is chosen, the forecaster suffers a total loss of about $m = O(n^{2/3})$ but gets full information about the outcome y_t. This situation is a modification of the problem of *label efficient prediction* studied in Helmbold and Panizza [25], and in Cesa-Bianchi, Lugosi, and Stoltz [12]. In particular, the algorithm proposed in Figure 2 coincides with that of of Theorem 2 of [12]–except maybe at those rounds when $Z_t = 1$. Indeed, Theorem 2 of [12] ensures that, with probability at least $1 - \delta$, not more than m among the Z_t have value 1, and that

$$\sum_{t=1}^{n} \ell(J_t, y_t) - \min_{j=1,\ldots,N} \sum_{t=1}^{n} \ell(j, y_t) \leqslant 8n\sqrt{\frac{\ln(4N/\delta)}{m}} \ .$$

This in turn implies that

$$\sum_{t=1}^{n} \ell(I_t, y_t) - \min_{j=1,\ldots,N} \sum_{t=1}^{n} \ell(j, y_t) \leqslant m + 8n\sqrt{\frac{\ln(4N/\delta)}{m}} \ ,$$

and substituting the proposed value for the parameter m concludes the proof. ∎

Remark 2 (*Dependence on N.*) Observe that, even when the condition of Theorem 1 is satisfied, the bound of Theorem 2 is considerably tighter. Indeed, even though the dependence on the time horizon n is identical in both bounds (of the order of $n^{-1/3}$), the bound of Theorem 2 depends on the number of actions N in a logarithmic way only. As an example, consider the case of the multi-armed bandit problem. Recall that here $\mathbf{H} = \mathbf{L}$ and there is a revealing action if and only if the loss matrix has a row whose elements are all different. In such a case Theorem 2 provides a bound of the order of $((\ln N)/n)^{1/3}$. On the other hand, there exist bandit problems for which, if $N \le n$, it is impossible to achieve a regret smaller than $(1/20)(N/n)^{1/2}$ (see [2]). If N is large, the logarithmic dependence of Theorem 2 gives a considerable advantage.

Interestingly, even if $\mathbf{L}$ cannot be expressed as $\mathbf{K}\,\mathbf{H}$, if a revealing action exists, the strategy of Section 3 may be used to achieve a small regret. This may be done by using a trick of Piccolboni and Schindelhauer [34] to first convert the problem into another partial-monitoring problem for which the strategy of Section 3 can be used. The basic step of this conversion is to replace the pair of $N \times M$ matrices $(\mathbf{L}, \mathbf{H})$ by a pair of $mN \times M$ matrices $(\mathbf{L}', \mathbf{H}')$ where $m \le M$ denotes the cardinality of the set $\mathcal{S} = \{s_1, \ldots, s_m\}$ of signals (i.e., the number of distinct elements of the matrix $\mathbf{H}$). In the obtained prediction problem the forecaster chooses among mN actions at each time instance. The converted loss matrix $\mathbf{L}'$ is obtained simply by repeating each row of the original loss matrix m times. The new feedback matrix $\mathbf{H}'$ is binary and is defined by

$$H'(m(i-1)+k, j) = \mathbb{I}_{h(i,j)=s_k}\,, \quad i = 1, \ldots, N,\ k = 1, \ldots, m,\ j = 1, \ldots, M\,.$$

Note that this way we get rid of the inconvenient problem of how to encode in a natural way the feedback symbols. If the matrices

$$\mathbf{H}' \quad \text{and} \quad \begin{bmatrix} \mathbf{H}' \\ \mathbf{L}' \end{bmatrix}$$

have the same rank, then there exists a matrix $\mathbf{K}'$ such that $\mathbf{L}' = \mathbf{K}'\,\mathbf{H}'$ and the forecaster of Section 3 may be applied to obtain a forecaster that has an average regret of the order of $n^{-1/3}$ for the converted problem. However, it is easy to see that any forecaster A with such a bounded regret for the converted problem may be trivially transformed into a forecaster A' for the original problem with the same regret bound: A' simply takes an action i whenever A takes an action of the form $m(i-1)+k$ for any $k = 1, \ldots, m$.

The above conversion procedure guarantees Hannan consistency for a large class of partial monitoring problems. For example, if the original problem has a revealing action i, then $m = M$ and the $M \times M$ sub-matrix formed by the rows $M(i-1)+1, \ldots, Mi$ of $\mathbf{H}'$ is the identity matrix (up to some permutations over the rows), and therefore has full rank. Then obviously a matrix $\mathbf{K}'$ with the desired property exists and the procedure described above leads to a forecaster with an average regret of the order of $n^{-1/3}$.

This last statement may be generalized, in a straightforward way, to an even larger class of problems as follows.

Corollary 1 (Distinguishing actions) *Assume that the feedback matrix* $\mathbf{H}$ *is such that for each outcome* $j = 1, \ldots, M$ *there exists an action* $i \in \{1, \ldots, N\}$ *such that for all outcomes* $j' \neq j$, $h(i,j) \neq h(i,j')$. *Then the conversion procedure described above leads to a Hannan consistent forecaster with an average regret of the order of* $n^{-1/3}$.

The rank of $\mathbf{H}'$ may be considered as a measure of the information provided by the feedback. The highest possible value is achieved by matrices $\mathbf{H}'$ with rank M. For such feedback matrices, Hannan consistency may be achieved for all associated loss matrices $\mathbf{L}'$.

Even though the above conversion strategy applies to a large class of problems, the associated condition fails to characterize the set of pairs $(\mathbf{L}, \mathbf{H})$ for which a Hannan consistent forecaster exists. Indeed, Piccolboni and Schindelhauer [34] show a second simple conversion of the pair $(\mathbf{L}', \mathbf{H}')$ that can be applied in situations when there is no matrix $\mathbf{K}'$ with the property $\mathbf{L}' = \mathbf{K}'\,\mathbf{L}'$. (This second conversion basically deals with some actions which they define as "non-exploitable" and which correspond to Pareto-dominated actions.) In these situations a Hannan consistent procedure may be constructed based on the forecaster of Section 3. On the other hand, Piccolboni and Schindelhauer also show that if the condition of Theorem 1 is not satisfied after the second step of conversion, then there exists an external randomization over the sequences of outcomes such that the sequence of expected regrets grows at least as n, where the expectations are understood with respect to the forecaster's auxiliary randomization and the external randomization. Thus, a proof by contradiction using the dominated-convergence theorem shows that Hannan consistency is impossible to achieve in these cases. This result combined with Theorem 1 implies the following gap theorem.

Corollary 2 *Consider a partial monitoring forecasting problem with loss and feedback matrices* $\mathbf{L}$ *and* $\mathbf{H}$. *If Hannan consistency can be achieved for this problem, then there exists a Hannan consistent forecaster whose average regret vanishes at rate* $n^{-1/3}$.

Thus, whenever it is possible to force the average regret to converge to zero, a convergence rate of the order of $n^{-1/3}$ is also possible. In some special cases, such as the multi-armed bandit problem, even faster rates of the order of $n^{-1/2}$ may be achieved (see Auer, Cesa-Bianchi, Freund, and Schapire [2] and Auer [1]). However, as it is shown in Section 5 below, for certain problems in which Hannan consistency is achievable, it can only be achieved with rate of convergence not faster than $n^{-1/3}$.

5 A lower bound on the regret

Next we show that the order of magnitude (in terms of the length of the play n) of the bound of Theorem 1 is, in general, not improvable. A closely related idea in a somewhat different context appears in Mertens, Sorin and Zamir [33, page 290].

Theorem 3 *Consider the partial monitoring problem of label efficient prediction introduced in Example 4 and defined by the pair of loss and feedback matrices*

$$\mathbf{L} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \textit{and} \qquad \mathbf{H} = \begin{bmatrix} a & b \\ c & c \\ c & c \end{bmatrix} .$$

Then, for any $n \geqslant 8$ and for any (randomized) forecasting strategy there exists a sequence $y_1, \ldots, y_n$ of outcomes such that

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\ell(I_t, y_t)\right] - \min_{i=1,2,3}\frac{1}{n}\sum_{t=1}^{n}\ell(i, y_t) \geqslant \frac{n^{-1/3}}{5} ,$$

where $\mathbb{E}$ denotes the expectation with respect to the auxiliary randomization of the forecaster.

Remark 3 Using techniques as in [12], it is easy to extend the theorem above to get a lower bound of the order of $((\ln N)/n)^{1/3}$. In view of the upper bound obtained in Theorem 2, this lower bound is the best possible for the variant of label efficient prediction described in Example 4, extended to the case of $N+1$ actions and N outcomes. However, we conjecture that for many other prediction problems with partial monitoring, significantly larger lower bounds (as a function of N) hold.

Proof. The proof proceeds by constructing a random sequence of outcomes and showing that, for any (possibly randomized) forecaster, the expected value of the regret with respect both to the random choice of the outcome sequence and to the forecaster's random choices is bounded from below by the claimed quantity.

More precisely, fix $n \geq 8$ and denote by $U_1, \ldots, U_n$ the auxiliary randomization which the forecaster has access to. Without loss of generality, it can be taken as an i.i.d. sequence of uniform random variables in $[0, 1]$. The underlying probability space is equipped with the σ-algebra of events generated by the random sequence of outcomes $Y_1, \ldots, Y_n$ and by the randomization $U_1, \ldots, U_n$. The random sequence of outcomes is independent of the auxiliary randomization, whose associated probability distribution is denoted by $\mathbb{P}_A$.

We define three different probability distributions, $\mathbb{P} \otimes \mathbb{P}_A$, $\mathbb{Q} \otimes \mathbb{P}_A$, and $\mathbb{R} \otimes \mathbb{P}_A$, formed by the product of the auxiliary randomization and one of the three probability distributions $\mathbb{P}$, $\mathbb{Q}$, and $\mathbb{R}$ over the sequence of outcomes defined as follows. Under $\mathbb{P}$ the sequence $Y_1, Y_2, \ldots, Y_n$ is formed by independent, identically distributed $\{1, 2\}$-valued random variables with parameter $1/2$. Under $\mathbb{Q}$ (respectively $\mathbb{R}$) the Y_i are also i.i.d. and $\{1, 2\}$-valued but with parameter $1/2 - \varepsilon$ (respectively $1/2 + \varepsilon$), where $\varepsilon > 0$ is chosen below.

We denote by $\mathbb{E}_{\mathbb{A}}$ (respectively, $\mathbb{E}_{\mathbb{P}}$, $\mathbb{E}_{\mathbb{Q}}$, $\mathbb{E}_{\mathbb{R}}$, $\mathbb{E}_{\mathbb{P}\otimes\mathbb{P}_A}$, $\mathbb{E}_{\mathbb{Q}\otimes\mathbb{P}_A}$, $\mathbb{E}_{\mathbb{R}\otimes\mathbb{P}_A}$) the expectation with respect to $\mathbb{P}_{\mathbb{A}}$ (respectively, $\mathbb{P}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{P} \otimes \mathbb{P}_A$, $\mathbb{Q} \otimes \mathbb{P}_A$, $\mathbb{R} \otimes \mathbb{P}_A$). Obviously,

$$\sup_{y_1^n}\left(\mathbb{E}_{\mathbb{A}}\left[\widehat{L}_n\right] - \min_{j=1,2,3} L_{j,n}\right) \geqslant \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{A}}\left[\widehat{L}_n\right] - \min_{j=1,2,3} L_{j,n}\right] . \tag{3}$$

Now,

$$\mathbb{E}_{\mathbb{Q}}\left[\min_{j=1,2,3} L_{j,n}\right] \leqslant \min_{j=1,2,3} \mathbb{E}_{\mathbb{Q}}\left[L_{j,n}\right] = \frac{n}{2} - n\varepsilon ,$$

whereas

$$\mathbb{E}_{\mathbb{Q}}\left[\widehat{L}_n\right] = \frac{n}{2} + \frac{1}{2}\mathbb{E}_{\mathbb{Q}}\left[N_1\right] + \varepsilon\mathbb{E}_{\mathbb{Q}}\left[N_3\right] - \varepsilon\mathbb{E}_{\mathbb{Q}}\left[N_2\right] ,$$

where N_j is the random variable denoting the number of times the forecaster chooses the action j over the sequence $Y_1, \ldots, Y_n$, given the state $U_1, \ldots, U_n$ of the auxiliary randomization, for $j = 1,\ 2,\ 3$. Thus, using Fubini's theorem,

$$\mathbb{E}_{\mathbb{Q}}\left[\mathbb{E}_{\mathbb{A}}\left[\widehat{L}_n\right] - \min_{j=1,2,3} L_{j,n}\right] \geqslant \frac{1}{2}\mathbb{E}_{\mathbb{Q}\otimes\mathbb{P}_A}\left[N_1\right] + \varepsilon\left(n - \mathbb{E}_{\mathbb{Q}\otimes\mathbb{P}_A}\left[N_2\right]\right) .$$

A similar argument shows that

$$\mathbb{E}_{\mathbb{R}}\left[\mathbb{E}_{\mathbb{A}}\left[\widehat{L}_n\right] - \min_{j=1,2,3} L_{j,n}\right] \geqslant \frac{1}{2}\mathbb{E}_{\mathbb{R}\otimes\mathbb{P}_A}\left[N_1\right] + \varepsilon\left(n - \mathbb{E}_{\mathbb{R}\otimes\mathbb{P}_A}\left[N_3\right]\right) .$$

Averaging the two inequalities we get

$$\mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{A}}\left[\widehat{L}_n\right] - \min_{j=1,2,3} L_{j,n}\right] \geqslant \frac{1}{2}\mathbb{E}_{\mathbb{P}\otimes\mathbb{P}_A}\left[N_1\right] + \varepsilon\left(n - \frac{1}{2}\left(\mathbb{E}_{\mathbb{Q}\otimes\mathbb{P}_A}\left[N_2\right] + \mathbb{E}_{\mathbb{R}\otimes\mathbb{P}_A}\left[N_3\right]\right)\right) . \quad (4)$$

Consider first a *deterministic* forecaster. Denote by $T_1, \ldots, T_{N_1} \in \{1, \ldots, n\}$ the times when the forecaster chose action 1. Since action 1 is revealing, we know the outcomes at these times, and denote them by $Z_{n+1} = (Y_{T_1}, \ldots, Y_{T_{N_1}})$. Denote by K_t the (random) index of the largest integer j such that $T_j \leqslant t-1$. Each action I_t of the forecaster is determined by the random vector (of random length) $Z_t = \left(Y_1, \ldots, Y_{T_{K_t}}\right)$. Since the forecaster we consider is deterministic, K_t is fully determined by Z_{n+1}. Hence, I_t may be seen as a function of Z_{n+1} rather than a function of Z_t only. This implies that, denoting by $\mathbb{P}_n$ (respectively $\mathbb{Q}_n$) the distribution of Z_{n+1} under $\mathbb{P}$ (respectively $\mathbb{Q}$), we have $\mathbb{Q}\left[I_t = 2\right] = \mathbb{Q}_n\left[I_t = 2\right]$ and $\mathbb{P}\left[I_t = 2\right] = \mathbb{P}_n\left[I_t = 2\right]$. Pinsker's inequality (see, e.g., [13, Lemma 12.6.1]) then ensures that, for all t,

$$\mathbb{Q}\left[I_t = 2\right] \leqslant \mathbb{P}\left[I_t = 2\right] + \sqrt{\frac{1}{2}\mathcal{K}\left(\mathbb{P}_n, \mathbb{Q}_n\right)} , \quad (5)$$

where $\mathcal{K}$ denotes the Kullback-Leibler divergence. The right-hand side may be further bounded using the following lemma.

Lemma 6 *Consider a deterministic forecaster. For* $0 \leqslant \varepsilon \leqslant 1/\sqrt{6}$,

$$\mathcal{K}\left(\mathbb{P}_n, \mathbb{Q}_n\right) \leqslant 6\mathbb{E}_{\mathbb{P}}\left[N_1\right]\varepsilon^2 .$$

Proof. We note that $Z_{n+1} = Z_n$, except when $I_n = 1$. In this case, $Z_{n+1} = (Z_n, Y_n)$. Therefore, using the chain rule for relative entropy (see, e.g., [13, Lemma 2.5.3]),

$$\begin{aligned}\mathcal{K}\left(\mathbb{P}_n, \mathbb{Q}_n\right) &\leqslant \mathcal{K}\left(\mathbb{P}_{n-1}, \mathbb{Q}_{n-1}\right) + \mathbb{P}\left[I_n = 1\right]\mathcal{K}\left(\mathbb{B}_{1/2}, \mathbb{B}_{1/2-\varepsilon}\right) \\ &\leqslant \mathcal{K}\left(\mathbb{P}_{n-1}, \mathbb{Q}_{n-1}\right) + \mathbb{P}\left[I_n = 1\right]\frac{2\varepsilon^2}{1-4\varepsilon^2} ,\end{aligned}$$

where $\mathbb{B}_p$ denotes the Bernoulli distribution with parameter p. We conclude by iterating the argument and using that $1 - 4\varepsilon^2 \geqslant 1/3$ for $0 \leqslant \varepsilon \leqslant 1/\sqrt{6}$. ∎

Summing (5) over $t = 1, \dots, n$, we have proved that

$$\mathbb{E}_{\mathbb{Q}}\left[N_2\right] \leqslant \mathbb{E}_{\mathbb{P}}\left[N_2\right] + n\varepsilon\sqrt{3\mathbb{E}_{\mathbb{P}}\left[N_1\right]}\,,$$

and this holds for any deterministic strategy. (Note that considering a deterministic strategy amounts to conditioning on the auxiliary randomization $U_1, \dots, U_n$.)

Consider now an arbitrary (possibly randomized) forecaster. Using Fubini's theorem and Jensen's inequality, we get

$$\mathbb{E}_{\mathbb{Q}\otimes\mathbb{P}_A}\left[N_2\right] \leqslant \mathbb{E}_{\mathbb{P}\otimes\mathbb{P}_A}\left[N_2\right] + n\varepsilon\sqrt{3\mathbb{E}_{\mathbb{P}\otimes\mathbb{P}_A}\left[N_1\right]}\,. \tag{6}$$

Symmetrically,

$$\mathbb{E}_{\mathbb{R}\otimes\mathbb{P}_A}\left[N_3\right] \leqslant \mathbb{E}_{\mathbb{P}\otimes\mathbb{P}_A}\left[N_3\right] + n\varepsilon\sqrt{3\mathbb{E}_{\mathbb{P}\otimes\mathbb{P}_A}\left[N_1\right]}\,. \tag{7}$$

Using $\mathbb{E}_{\mathbb{P}\otimes\mathbb{P}_A}\left[N_2\right] + \mathbb{E}_{\mathbb{P}\otimes\mathbb{P}_A}\left[N_3\right] \leqslant n$, and substituting (6) and (7) into (4) yield

$$\mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{A}}\left[\widehat{L}_n\right] - \min_{j=1,2,3} L_{j,n}\right] \geqslant \frac{1}{2}m_0 + n\varepsilon\left(\frac{1}{2} - \varepsilon\sqrt{3m_0}\right)\,, \tag{8}$$

where m_0 denotes $\mathbb{E}_{\mathbb{P}\otimes\mathbb{P}_A}\left[N_1\right]$. If $m_0 \leqslant 1/8$ then for $\varepsilon = 1/\sqrt{6}$ the right-hand side of (8) is at least $n/10$, which is greater than $n^{2/3}/5$ for $n \geqslant 8$. Otherwise, if $m_0 \geqslant 1/8$, we set $\varepsilon = \left(4\sqrt{3m_0}\right)^{-1}$, which still satisfies $0 \leqslant \varepsilon \leqslant 1/\sqrt{6}$. The lower bound then becomes

$$\mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{A}}\left[\widehat{L}_n\right] - \min_{j=1,2,3} L_{j,n}\right] \geqslant \frac{1}{2}m_0 + \frac{n}{16\sqrt{3m_0}}$$

and the right-hand side may be seen to be always bigger than $n^{2/3}/5$. An application of (3) concludes the proof. ∎

6 Internal regret

In this section we deal with the stronger notion of internal (or conditional) regret. Internal regret is concerned with consistent modifications of the forecasting strategy. Each of these possible modifications is parameterized by a departure function $\Phi : \{1, \dots, N\} \to \{1, \dots, N\}$. After round n, the cumulative loss of the forecaster is compared to the cumulative loss that would have been accumulated had the forecaster chosen action $\Phi(I_t)$ instead of action I_t at round t, $t = 1, \dots, n$. If such a consistent modification does not result in a much smaller accumulated loss then the strategy is said to have small internal regret. Formally, we seek strategies achieving

$$\frac{1}{n}\sum_{t=1}^{n} \ell(I_t, y_t) - \frac{1}{n}\min_{\Phi}\sum_{t=1}^{n} \ell(\Phi(I_t), y_t) = o(1)$$

where the minimization is over all possible functions Φ. We can extend the notion of Hannan consistency to internal regret by requiring that the above average regret vanishes with probability 1 as $n \to \infty$.

The notion of internal regret has been shown to be useful in the theory of equilibria of repeated games. Foster and Vohra [16, 18] showed that if all players of a finite game choose a strategy that is Hannan consistent with respect to the internal regret, then the joint empirical frequencies of play converge to the set of correlated equilibria of the game (see also Fudenberg and Levine [19], Hart and Mas-Colell [22]). Foster and Vohra [16, 18] proposed internal regret minimizing strategies for the full-information case, see also Cesa-Bianchi and Lugosi [10]. We design here such a procedure in the setting of partial monitoring. The key tool is a conversion trick described in Stoltz and Lugosi [36] (see also Blum and Mansour [7] for a similar procedure). This trick essentially converts external regret minimizing strategies into internal regret minimizing strategies, under full information. We extend it here to prediction under partial monitoring.

The forecaster we propose is formed by a sub-algorithm and a master algorithm. The parameters η_t and γ_t used below are tuned as in Section 3. At each round t the sub-algorithm outputs a probability distribution

$$\boldsymbol{u}_t = \left(u_t^{i \to j}\right)_{(i,j)\,:\,i \neq j}$$

over the set of pairs of different actions; with the help of $\boldsymbol{u}_t$ the master algorithm computes a probability distribution $\boldsymbol{p}_t$ over the actions.

Consider the loss estimates $\widetilde{\ell}(i, y_t)$ defined in (1). For a given distribution $\boldsymbol{p}$ over $\{1, \ldots, N\}$, denote

$$\widetilde{\ell}(\boldsymbol{p}, y) = \sum_{k=1}^{N} p_k \, \widetilde{\ell}(k, y) \ .$$

Now introduce the cumulative losses

$$\widetilde{L}_{t-1}^{i \to j} = \sum_{s=1}^{t-1} \widetilde{\ell}(\boldsymbol{p}_s^{i \to j}, y_s)$$

where $\boldsymbol{p}_s^{i \to j}$ denotes the probability distribution obtained from $\boldsymbol{p}_s$ by moving the probability mass $p_{i,s}$ from i to j; that is, we set $p_{s,i}^{i \to j} = 0$ and $p_{s,j}^{i \to j} = p_{s,j} + p_{s,i}$. The distribution $\boldsymbol{u}_t$ computed by the sub-algorithm is an exponentially weighted average associated to the cumulative losses $\widetilde{L}_{t-1}^{i \to j}$, that is,

$$u_t^{i \to j} = \frac{\exp\left(-\eta_t \widetilde{L}_{t-1}^{i \to j}\right)}{\sum_{k \neq l} \exp\left(-\eta_t \widetilde{L}_{t-1}^{k \to l}\right)} \ .$$

Now let $\widetilde{\boldsymbol{p}}_t$ be the probability distribution over the set of actions defined by the equation

$$\sum_{(i,j)\,:\,i \neq j} u_t^{i \to j} \widetilde{\boldsymbol{p}}_t^{i \to j} = \widetilde{\boldsymbol{p}}_t \ . \tag{9}$$

Such a distribution exists, and can be computed by a simple Gaussian elimination (see e.g., Foster and Vohra [18], or Stoltz and Lugosi [36]). The master algorithm then chooses, at round t, the action I_t drawn according to the probability distribution

$$\boldsymbol{p}_t = (1 - \gamma_t) \widetilde{\boldsymbol{p}}_t + \frac{\gamma_t}{N} \boldsymbol{1} \tag{10}$$

where $\boldsymbol{1} = (1, \ldots, 1)$.

Theorem 4 *Consider any partial monitoring problem such that the loss and feedback matrices satisfy* $\mathbf{L} = \mathbf{K}\,\mathbf{H}$ *for some* $N \times N$ *matrix* $\mathbf{K}$ *with* $k^* = \max\{1, \max_{i,j} |k(i,j)|\}$*, and consider the forecaster described above. Let* $\delta \in (0,1)$*. Then, for all* n*, with probability at least* $1-\delta$*, the cumulative internal regret is bounded as*

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} \ell(I_t, y_t) - \min_{\Phi} \frac{1}{n}\sum_{t=1}^{n} \ell(\Phi(I_t), y_t) \\
&\quad\leq\ 9\left(\frac{(k^*)^2 N^5 \ln N}{n}\right)^{1/3}\left(1 + \sqrt{\frac{3}{2}\frac{\ln(2N^2)/\delta)}{\ln N}}\right) \\
&\qquad + N\sqrt{\frac{1}{2n}\ln\frac{2N^2}{\delta}} + 4(k^*N)^{4/3} n^{-2/3} (\ln N)^{-1/3} \ln \frac{2N^2}{\delta} \\
&\qquad + \frac{1}{n}\left(2N + ((k^*N)^2 \ln N)^{1/3} + k^*N\right) \ln\frac{2N^2}{\delta}
\end{aligned}$$

where the minimum is taken over all functions $\Phi : \{1, \ldots, N\} \to \{1, \ldots, N\}$.

Note that with the help of Borel-Cantelli lemma, Theorem 4 shows that, under the same conditions on $\mathbf{L}$ and $\mathbf{H}$, the forecaster decribed above achieves Hannan consistency with respect to internal regret.

Proof. First observe that it suffices to consider departure functions Φ that differ from the identity function in only one point of their domain. This follows simply from

$$\sum_{t=1}^{n} \ell(I_t, y_t) - \min_{\Phi} \sum_{t=1}^{n} \ell(\Phi(I_t), y_t) \leqslant N \left(\max_{i \neq j} \sum_{t=1}^{n} \mathbb{I}_{I_t = i} \left(\ell(i, y_t) - \ell(j, y_t) \right) \right) .$$

We now bound the right-hand side of the latter inequality.

For a given t, the estimated losses $\widetilde{\ell}(\boldsymbol{p}_t^{i\to j}, y_t)$, $i \neq j$, fall in the interval $[-k^*N/\gamma_t,\ k^*N/\gamma_t]$. Since γ_t and η_t are tuned as in Theorem 1, $k^*N\eta_t/\gamma_t \leqslant 1$, and we may apply Theorem 5 to derive

$$\begin{aligned}
\sum_{t=1}^{n} \sum_{i \neq j} u_t^{i\to j} \widetilde{\ell}(\boldsymbol{p}_t^{i\to j}, y_t) - \min_{i \neq j} \sum_{t=1}^{n} \widetilde{\ell}(\boldsymbol{p}_t^{i\to j}, y_t)& \\
\leqslant \frac{2\ln N(N-1)}{\eta_{n+1}} + \sum_{t=1}^{n} \eta_t \sum_{i \neq j} u_t^{i\to j} \left(\widetilde{\ell}(\boldsymbol{p}_t^{i\to j}, y_t)\right)^2 \ .& \qquad (11)
\end{aligned}$$

For $i \neq j$, $\mathbf{1}^{i\to j}$ is the vector $\boldsymbol{v}$ such that $v_i = 0$, $v_j = 2$, and $v_k = 1$ for all $k \neq i$ and $k \neq j$. Use first (10) and then (9) to rewrite the first term of the left-hand side of (11) as

$$\begin{aligned}
\sum_{t=1}^{n} \sum_{i \neq j} u_t^{i \to j} \widetilde{\ell}(\boldsymbol{p}_t^{i \to j}, y_t) &= \sum_{t=1}^{n} \sum_{i \neq j} u_t^{i \to j} \left((1-\gamma_t) \widetilde{\ell}(\widetilde{\boldsymbol{p}}_t^{i \to j}, y_t) + \frac{\gamma_t}{N} \widetilde{\ell}(\mathbf{1}^{i \to j}, y_t) \right) \\
&= \sum_{t=1}^{n} (1-\gamma_t) \widetilde{\ell}(\widetilde{\boldsymbol{p}}_t, y_t) + \sum_{t=1}^{n} \frac{\gamma_t}{N} \sum_{i \neq j} u_t^{i \to j} \widetilde{\ell}(\mathbf{1}^{i \to j}, y_t) \\
&= \sum_{t=1}^{n} \widetilde{\ell}(\boldsymbol{p}_t, y_t) + \sum_{t=1}^{n} \frac{\gamma_t}{N} \sum_{i \neq j} u_t^{i \to j} \left(\widetilde{\ell}(\mathbf{1}^{i \to j}, y_t) - \widetilde{\ell}(\mathbf{1}, y_t) \right) \\
&= \sum_{t=1}^{n} \widetilde{\ell}(\boldsymbol{p}_t, y_t) + \sum_{t=1}^{n} \frac{\gamma_t}{N} \sum_{i \neq j} u_t^{i \to j} \left(\widetilde{\ell}(j, y_t) - \widetilde{\ell}(i, y_t) \right) .
\end{aligned}$$

Substituting into (11), we have

$$\begin{aligned}
& \max_{i \neq j} \sum_{t=1}^{n} p_{i,t} \left(\widetilde{\ell}(i, y_t) - \widetilde{\ell}(j, y_t) \right) \\
= \; & \sum_{t=1}^{n} \widetilde{\ell}(\boldsymbol{p}_t, y_t) - \min_{i \neq j} \sum_{t=1}^{n} \widetilde{\ell}(\boldsymbol{p}_t^{i \to j}, y_t) \qquad (12) \\
\leqslant \; & \frac{4 \ln N}{\eta_{n+1}} + \sum_{t=1}^{n} \eta_t \sum_{i \neq j} u_t^{i \to j} \left(\widetilde{\ell}(\boldsymbol{p}_t^{i \to j}, y_t) \right)^2 + \sum_{t=1}^{n} \frac{\gamma_t}{N} \sum_{i \neq j} u_t^{i \to j} \left(\widetilde{\ell}(i, y_t) - \widetilde{\ell}(j, y_t) \right) .
\end{aligned}$$

Now, we apply Bernstein's inequality (see Lemma 7) several times again and mimic the proofs of Lemmas 1 and 2. For all pairs $i \neq j$, with probability at least $1 - \delta/(2N(N-1)+2)$,

$$\begin{aligned}
& \sum_{t=1}^{n} p_{i,t} \left(\widetilde{\ell}(i, y_t) - \widetilde{\ell}(j, y_t) \right) \geqslant \sum_{t=1}^{n} p_{i,t} \left(\ell(i, y_t) - \ell(j, y_t) \right) \\
& - \left(\sqrt{4(k^* N)^2 \left(\sum_{t=1}^{n} \frac{1}{\gamma_t} \right) \ln \frac{2N(N-1)+2}{\delta}} + \frac{2\sqrt{2}}{3} \left(1 + \frac{k^* N}{\gamma_n} \right) \ln \frac{2N(N-1)+2}{\delta} \right) . \qquad (13)
\end{aligned}$$

Similarly to Lemma 3, we also have, with probability at least $1 - \delta/(2N(N-1)+2)$,

$$\begin{aligned}
& \sum_{t=1}^{n} \eta_t \sum_{i \neq j} u_t^{i \to j} \left(\widetilde{\ell}(\boldsymbol{p}_t^{i \to j}, y_t) \right)^2 \leq \sum_{t=1}^{n} \eta_t \frac{(k^* N)^2}{\gamma_t} \\
& \qquad + \sqrt{2(k^* N)^4 \left(\sum_{t=1}^{n} \frac{\eta_t^2}{\gamma_t^3} \right) \ln \frac{2N(N-1)+2}{\delta}} + \frac{\sqrt{2}}{3} \ln \frac{2N(N-1)+2}{\delta} \qquad (14)
\end{aligned}$$

whereas, similarly to Lemma 4, with probability at least $1-\delta/(2N(N-1)+2)$,

$$\sum_{t=1}^{n} \frac{\gamma_t}{N} \sum_{i \neq j} u_t^{i\to j} \left(\widetilde{\ell}(i,y_t) - \widetilde{\ell}(j,y_t)\right) \le \frac{1}{N} \sum_{t=1}^{n} \gamma_t + \sqrt{4(k^*)^2 \left(\sum_{t=1}^{n} \gamma_t\right) \ln \frac{2N(N-1)+2}{\delta}} + \frac{\sqrt{2}}{3}\left(k^* + \frac{\gamma_1}{N}\right) \ln \frac{2N(N-1)+2}{\delta} \,. \tag{15}$$

We then use the Hoeffding-Azuma inequality (see Hoeffding [27], Azuma [4]) $N(N-1)$ times to show that for every pair $i \neq j$, with probability at least $1-\delta/(2N(N-1)+2)$,

$$\sum_{t=1}^{n} p_{i,t} \left(\ell(i,y_t) - \ell(j,y_t)\right) \ge \sum_{t=1}^{n} \mathbb{I}_{I_t=i} \left(\ell(i,y_t) - \ell(j,y_t)\right) - \sqrt{2n \ln \frac{N(N-1)+3}{\delta}} \,. \tag{16}$$

Finally, we substitute inequalities (13)–(16) into (12) and use a union-of-event bound to obtain that, with probability at least $1-\delta$,

$$\begin{aligned} &\max_{i \neq j} \sum_{t=1}^{n} \mathbb{I}_{I_t=i} \left(\ell(i,y_t) - \ell(j,y_t)\right) \\ &\leqslant \;\; \frac{4 \ln N}{\eta_{n+1}} \\ &\quad + \sqrt{4(k^*N)^2 \left(\sum_{t=1}^{n} \frac{1}{\gamma_t}\right) \ln \frac{1}{\delta'}} + \frac{2\sqrt{2}}{3}\left(1 + \frac{k^*N}{\gamma_n}\right) \ln \frac{1}{\delta'} \\ &\quad + \sum_{t=1}^{n} \eta_t \frac{(k^*N)^2}{\gamma_t} + \sqrt{2(k^*N)^4 \left(\sum_{t=1}^{n} \frac{\eta_t^2}{\gamma_t^3}\right) \ln \frac{1}{\delta'}} + \frac{\sqrt{2}}{3} \ln \frac{1}{\delta'} \\ &\quad + \frac{1}{N} \sum_{t=1}^{n} \gamma_t + \sqrt{4(k^*)^2 \left(\sum_{t=1}^{n} \gamma_t\right) \ln \frac{1}{\delta'}} + \frac{\sqrt{2}}{3}\left(k^* + \frac{\gamma_1}{N}\right) \ln \frac{1}{\delta'} \\ &\quad + \sqrt{2n \ln \frac{1}{\delta'}} \,, \end{aligned}$$

where we used the notation $\delta' = \delta/(2N(N-1)+2)$, with $\delta' \geqslant \delta/(2N^2)$ when $N \geqslant 2$. The proof is now concluded as that of Theorem 1. ∎

7 Random feedback

Several authors consider an extended setup in which the feedbacks are random variables. See Rustichini [35], Mannor and Shimkin [30], Weissman and Merhav [39], Weissman, Merhav, and

Somekh-Baruch [40] for examples. In this section we briefly point out that most of the results of this paper extend effortlessly to this more general case.

To describe the model, denote by $\Delta(\mathcal{S})$ the set of all probability distributions over the set of signals $\mathcal{S}$. The signaling structure is formed by a collection of NM probability distributions $\mu_{(i,j)}$ over $\mathcal{S}$, for $i = 1, \ldots, N$ and $j = 1, \ldots, M$. At each round, the forecaster now observes a random variable $H(I_t, y_t)$, drawn independently from all the other random variables, with distribution $\mu_{(I_t,y_t)}$.

We may easily generalize the results of Theorems 1 and 4 to the case of random feedbacks. As above, each element of $\mathcal{S}$ is encoded by a real number in $[-1, 1]$. Let $\mathbf{E}$ be the $N \times M$ matrix whose elements are given by the expectations of the random variables $H(i, j)$. Theorems 1 and 4 remain true under the condition that there exists a matrix $\mathbf{K}$ such that $\mathbf{L} = \mathbf{K}\,\mathbf{E}$. The only necessary modification is how the losses are estimated. Here the forecaster uses the estimates

$$\breve{\ell}(i, y_t) = \frac{k(i, I_t)H(I_t, y_t)}{p_{I_t,t}} \quad i = 1, \ldots, N$$

instead of the estimates defined in Section 3. Conditioned on $I_1, \ldots, I_{t-1}$, the expectation of $\breve{\ell}(i, y_t)$ is the loss $\ell(i, y_t)$. Since this, together with boundedness, are the only conditions that were needed in the proofs, the extension of the results to this more general framework is immediate.

The results of Section 4 may be generalized to the case of random feedbacks as well. For example, to construct $\mathbf{H}'$ when $\mathbf{H}$ is a matrix of probability distributions over $\mathcal{S}$, we proceed as follows: for $1 \leqslant i \leqslant N$ and $s \in \mathcal{S}$, denote by $H_{(i,s)}$ the row vector of elements in $[0, 1]$, such that the k-th element of $H_{(i,s)}$ is $\mu_{(i,k)}(s)$. Now, the $((k_1 - 1)m + k_2)$-th row of $\mathbf{H}'$, $1 \leqslant k_1 \leqslant N$, $1 \leqslant k_2 \leqslant m$, is $H_{(k_1,s_{k_2})}$. All the other details of the construction and the proofs go through.

Appendix: Bernstein's inequality

Bernstein's inequality (see, e.g. [15]) is used several times in the proofs.

Lemma 7 (Bernstein's inequality) *Let $X_1, X_2, \ldots, X_n$ be a bounded martingale difference sequence (with respect to the filtration $\mathcal{F} = (\mathcal{F}_t)_{1 \leqslant t \leqslant n}$)), with increments bounded in absolute values by K, and*

$$M_n = \sum_{t=1}^{n} X_t$$

the associated martingale. Denote its predictable quadratic variation by

$$V_n = \sum_{t=1}^{n} \mathbb{E}\left[X_t^2 \,|\, \mathcal{F}_{t-1}\right]$$

and assume that $V_n \leqslant v$ for some constant v. Then, for all $u > 0$,

$$\mathbb{P}\left[M_n > u\right] \leqslant \exp\left(-\frac{u^2}{2\left(v + Ku/3\right)}\right)$$

and in particular, for all $x > 0$,

$$\mathbb{P}\left[M_n > \sqrt{2vx} + (\sqrt{2}/3)Kx\right] \leqslant e^{-x} \ .$$

Appendix: basic lemmas

Theorem 5 *Consider any sequence of losses $\ell_{i,t} \in [-B_t, B_t]$, $i = 1, \ldots, N$, $B_t > 0$, $t = 1, \ldots, n$, and any non-increasing sequence of tuning parameters $\eta_t > 0$, $t = 1, \ldots, n$, such that $\eta_t B_t \leqslant 1$ for all t. Then, the forecaster which uses the exponentially weighted averages*

$$q_{i,t} = \frac{w_{i,t}}{\sum_{j=1}^N w_{j,t}}, \qquad i = 1, \ldots, N,$$

where

$$w_{i,t} = \exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_{i,s}\right) \ ,$$

satisfies

$$\sum_{t=1}^n \sum_{i=1}^N q_{i,t}\ell_{i,t} - \min_{j=1,\ldots,N} \sum_{t=1}^n \ell_{j,t} \leqslant \left(\frac{2}{\eta_{n+1}} - \frac{1}{\eta_1}\right) \ln N + \sum_{t=1}^n \eta_t \sum_{i=1}^N q_{i,t}\, \ell_{i,t}^2 \ .$$

The proof below is a simple modification of an argument first proposed in [3]. Denote the numerator of the defining expression of $q_{i,t}$ by $w_{i,t} = e^{-\eta_t L_{i,t-1}}$, where $L_{i,t-1} = \ell_{i,1} + \ldots + \ell_{i,t-1}$, and use $w'_{i,t} = e^{-\eta_{t-1} L_{i,t-1}}$ to denote the weight $w_{i,t}$ where the parameter η_t is replaced by η_{t-1}. The normalization factors will be denoted by $W_t = \sum_{j=1}^N w_{j,t}$ and $W'_t = \sum_{j=1}^N w'_{j,t}$. Finally, we use k_t to denote the expert whose loss after the first t rounds is the lowest (ties are broken by choosing the expert with smallest index). That is, $L_{k_t,t} = \min_{i \le N} L_{i,t}$.

In the proof of the theorem, we also make use of the following technical lemma.

Lemma 8 *For all $N \geq 2$, for all $\beta \geq \alpha \geq 0$, and for all $d_1, \ldots, d_N \geq 0$ such that $\sum_{i=1}^N e^{-\alpha d_i} \geq 1$,*

$$\ln \frac{\sum_{i=1}^N e^{-\alpha d_i}}{\sum_{j=1}^N e^{-\beta d_j}} \leq \frac{\beta - \alpha}{\alpha} \ln N \ .$$

Proof. We begin by writing

$$\begin{aligned}
\ln \frac{\sum_{i=1}^N e^{-\alpha d_i}}{\sum_{j=1}^N e^{-\beta d_j}} &= \ln \frac{\sum_{i=1}^N e^{-\alpha d_i}}{\sum_{j=1}^N e^{(\alpha-\beta) d_j} e^{-\alpha d_j}} \\
&= -\ln \mathbb{E}\left[e^{(\alpha-\beta)D}\right] \\
&\leq (\beta - \alpha)\mathbb{E}\left[D\right]
\end{aligned}$$

where we applied Jensen inequality to the random variable D taking value d_i with probability $e^{-\alpha d_i} / \sum_{j=1}^N e^{-\alpha d_j}$ for each $j = 1, \ldots, N$. Since D takes at most N distinct values, its entropy $H(D)$ is at most $\ln N$. Therefore

$$\begin{aligned} \ln N \geq H(D) &= \frac{\sum_{i=1}^N e^{-\alpha d_i}}{\sum_{j=1}^N e^{-\beta d_j}} \left(\alpha d_i + \ln \sum_{j=1}^N e^{-\beta d_j} \right) \\ &= \alpha \mathbb{E}\left[D\right] + \ln \sum_{j=1}^N e^{-\beta d_j} \geq \alpha \mathbb{E}\left[D\right] \end{aligned}$$

where the last inequality holds since $\sum_{i=1}^N e^{-\alpha d_i} \geq 1$. Hence $\mathbb{E}\left[D\right] \leq (\ln N)/\alpha$. As $\beta > \alpha$ by hypothesis, we can plug the bound on $\mathbb{E}\left[D\right]$ in the upper bound above and conclude the proof. ∎

Proof. As it is usual in the analysis of the exponentially weighted average predictor, we study the evolution of $\ln(W_{t+1}/W_t)$. However, here we need to couple this term with $\ln(w_{k_{t-1},t}/w_{k_t,t+1})$ including in both terms the time-varying parameter η_t. Tracking the currently best expert k_t is used to lower bound the weight $\ln(w_{k_t,t+1}/W_{t+1})$. In fact, the weight of the overall best expert (after n rounds) could get arbitrarily small during the prediction process. We thus obtain the following

$$\begin{aligned} &\frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t}}{W_t} - \frac{1}{\eta_{t+1}} \ln \frac{w_{k_t,t+1}}{W_{t+1}} \\ &= \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln \frac{W_{t+1}}{w_{k_t,t+1}} + \frac{1}{\eta_t} \ln \frac{w'_{k_t,t+1}/W'_{t+1}}{w_{k_t,t+1}/W_{t+1}} + \frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t}/W_t}{w'_{k_t,t+1}/W'_{t+1}} \\ &= (A) + (B) + (C) \ . \end{aligned}$$

We now bound separately the three terms on the right-hand side. The term (A) is easily bounded by using $\eta_{t+1} \leqslant \eta_t$ and using the fact that k_t is the index of the expert with smallest loss after the first t rounds. Therefore, $w_{k_t,t+1}/W_{t+1}$ must be at least $1/N$. Thus we have

$$(A) = \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln \frac{W_{t+1}}{w_{k_t,t+1}} \leq \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln N \ .$$

We proceed to bounding the term (B) as follows

$$\begin{aligned} (B) &= \frac{1}{\eta_t} \ln \frac{w'_{k_t,t+1}/W'_{t+1}}{w_{k_t,t+1}/W_{t+1}} = \frac{1}{\eta_t} \ln \frac{\sum_{i=1}^N e^{-\eta_{t+1}(L_{i,t} - L_{k_t,t})}}{\sum_{j=1}^N e^{-\eta_t(L_{j,t} - L_{k_t,t})}} \\ &\leq \frac{\eta_t - \eta_{t+1}}{\eta_t \eta_{t+1}} \ln N = \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln N \end{aligned}$$

where the inequality is proven by applying Lemma 8 with $d_i = L_{i,t} - L_{k_t,t}$. Note that $d_i \geq 0$ since k_t is the index of the expert with smallest loss after the first t rounds and $\sum_{i=1}^N e^{-\eta_{t+1} d_i} \geq 1$ as for $i = k_t$ we have $d_i = 0$. The term (C) is first split as follows

$$(C) = \frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t}/W_t}{w'_{k_t,t+1}/W'_{t+1}} = \frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t}}{w'_{k_t,t+1}} + \frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} \ .$$

We bound separately each one of the two terms on the right-hand side. For the first one, we have

$$\frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t}}{w'_{k_t,t+1}} = \frac{1}{\eta_t} \ln \frac{e^{-\eta_t L_{k_{t-1},t-1}}}{e^{-\eta_t L_{k_t,t}}} = L_{k_t,t} - L_{k_{t-1},t-1} \, .$$

For the second term, we consider the random variable Z_t that takes value $\ell_{i,t}$ with probability $q_{i,t} = w_{i,t}/W_t$ for each $i = 1, \ldots, N$. As $\eta_t B_t \leqslant 1$, we have in particular $\eta_t \ell_{i,t} \leqslant 1$, so we may use the inequality $e^x \leqslant 1 + x + x^2$ for $x \leqslant 1$, and $\ln(1+u) \leqslant u$ for $u > -1$, to obtain

$$\begin{aligned} \frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} &= \frac{1}{\eta_t} \ln \frac{\sum_{i=1}^N w_{i,t} e^{-\eta_t \ell_{i,t}}}{W_t} = \frac{1}{\eta_t} \ln \sum_{i=1}^N q_{i,t} e^{-\eta_t \ell_{i,t}} \\ &\leq \frac{1}{\eta_t} \ln \left(\sum_{i=1}^N q_{i,t} \left(1 - \eta_t \ell_{i,t} + \eta_t^2 \ell_{i,t}^2\right) \right) \\ &\leqslant -\sum_{i=1}^N q_{i,t} \ell_{i,t} + \eta_t \sum_{i=1}^N q_{i,t} \ell_{i,t}^2 \, . \end{aligned}$$

Finally, we plug back in the main equation the bounds on the first two terms (A) and (B), and the bounds on the two parts of the term (C). After rearranging we obtain

$$\begin{aligned} \sum_{i=1}^N q_{i,t} \ell_{i,t} \leq \; & \left(L_{k_t,t} - L_{k_{t-1},t-1}\right) + \eta_t \sum_{i=1}^N q_{i,t} \ell_{i,t}^2 \\ & - \frac{1}{\eta_{t+1}} \ln \frac{w_{k_t,t+1}}{W_{t+1}} + \frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t}}{W_t} \\ & + 2 \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln N \, . \end{aligned}$$

We apply the above inequalities to each $t = 1, \ldots, n$ and sum up using

$$\begin{aligned} & \sum_{t=1}^n \left(L_{k_t,t} - L_{k_{t-1},t-1}\right) = \min_{j=1,\ldots,N} L_{j,n} \, , \\ & \sum_{t=1}^n \left(-\frac{1}{\eta_{t+1}} \ln \frac{w_{k_t,t+1}}{W_{t+1}} + \frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t}}{W_t} \right) \leq -\frac{1}{\eta_1} \ln \frac{w_{k_0,1}}{W_1} = \frac{\ln N}{\eta_1} \end{aligned}$$

to conclude the proof. ∎

References

[1] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002. A preliminary version has appeared in *Proc. of the 41th Annual Symposium on Foundations of Computer Science*.

[2] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32:48–77, 2002.

[3] P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64:48–75, 2002.

[4] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 68:357–367, 1967.

[5] A. Baños. On pseudo-games. *Annals of Mathematical Statistics*, 39:1932–1945, 1968.

[6] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.

[7] A. Blum and Y. Mansour. From external to internal regret. In *Proceedings of the 18th Annual Conference on Computational Learning Theory*, 2005. To appear.

[8] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, D. Haussler, R. Schapire, and M.K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[9] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Annals of Statistics*, 27, 1999.

[10] N. Cesa-Bianchi and G. Lugosi. Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51:239–261, 2003.

[11] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, to appear.

[12] N. Cesa-Bianchi, G. Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, to appear.

[13] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley, New York, 1991.

[14] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38:1258–1270, 1992.

[15] D. A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3:100–118, 1975.

[16] D. Foster and R. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21:40–55, 1997.

[17] D. Foster and R. Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1998.

[18] D. Foster and R. Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29:7–36, 1999.

[19] D. Fudenberg and D. Levine. Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19:1065–1089, 1995.

[20] D. Fudenberg and D.K. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

[21] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the theory of games*, 3:97–139, 1957.

[22] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.

[23] S. Hart and A. Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98:26–54, 2001.

[24] S. Hart and A. Mas-Colell. A reinforcement procedure leading to correlated equilibrium. In G. Debreu, W. Neuefeind, and W. Trockel, editors, *Economic Essays: A Festschrift for Werner Hildenbrand*, pages 181–200. Srpinger, New York, 2002.

[25] D. P. Helmbold and S. Panizza. Some label efficient learning results. In *Proceedings of the 10th Annual Conference on Computational Learning Theory*, pages 218–230. ACM Press, 1997.

[26] D.P. Helmbold, N. Littlestone, and P.M. Long. Apple tasting. *Information and Computation*, 161(2):85–139, 2000.

[27] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[28] R. Kleinberg and T. Leighton. The value of knowing a demand curve: Bounds on regret for on-line posted-price auctions. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, pages 594–605. IEEE Press, 2003.

[29] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[30] 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, New York, 2003.

[31] N. Megiddo. On repeated games with incomplete information played by non-Bayesian players. *International Journal of Game Theory*, 9:157–167, 1980.

[32] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44:2124–2147, 1998.

[33] J.-F. Mertens, S. Sorin, and S. Zamir. Repeated games. CORE Discussion paper, no. 9420,9421,9422, Louvain-la-Neuve, 1994.

[34] 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.

[35] A. Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29:224–243, 1999.

[36] G. Stoltz and G. Lugosi. Internal regret in on-line portfolio selection. *Machine Learning*, to appear.

[37] V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

[38] V.G. Vovk. Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pages 372–383, 1990.

[39] T. Weissman and N. Merhav. Universal prediction of binary individual sequences in the presence of noise. *IEEE Transactions on Information Theory*, 47:2151–2173, 2001.

[40] T. Weissman, N. Merhav, and Somekh-Baruch. Twofold universal prediction schemes for achieving the finite state predictability of a noisy individual binary sequence. *IEEE Transactions on Information Theory*, 47:1849–1866, 2001.