

Minimizing Regret with Label Efficient Prediction^{*}

Nicolò Cesa-Bianchi¹, Gábor Lugosi², and Gilles Stoltz³

¹ DSI, Università di Milano
via Comelico 39, 20135 Milano, Italy
`cesa-bianchi@dsi.unimi.it`

² Department of Economics, Universitat Pompeu Fabra
Ramon Trias Fargas 25-27, 08005 Barcelona, Spain
`lugosi@upf.es`

³ Laboratoire de Mathématiques, Université Paris-Sud,
91405 Orsay Cedex, France
`gilles.stoltz@math.u-psud.fr`

Abstract. We investigate label efficient prediction, a variant of the problem of prediction with expert advice, proposed by Helmbold and Panizza, in which the forecaster does not have access to the outcomes of the sequence to be predicted unless he asks for it, which he can do for a limited number of times. We determine matching upper and lower bounds for the best possible excess error when the number of allowed queries is a constant. We also prove that a query rate of order $(\ln n)(\ln \ln n)^2/n$ is sufficient for achieving Hannan consistency, a fundamental property in game-theoretic prediction models. Finally, we apply the label efficient framework to pattern classification and prove a label efficient mistake bound for a randomized variant of Littlestone’s zero-threshold Winnow algorithm.

1 Introduction

Prediction with expert advice, a framework introduced about fifteen years ago in learning theory, may be viewed as a direct generalization of the theory of repeated games, a field pioneered by Hannan in the mid-fifties. At a certain level of abstraction, the common subject of these studies is the problem of forecasting each element y_t of an unknown “target” sequence given the knowledge of the previous elements y_1, \dots, y_{t-1} . The forecaster’s goal is to predict the target sequence almost as well as any forecaster using the same guess all the times. We call this the sequential prediction problem. To provide a suitable parametrization of the problem, we assume that the set from which the forecaster picks its guesses is finite of size $N > 1$, while the set to which the target sequence elements belong may be of arbitrary cardinality. A real-valued bounded loss function ℓ is then used to quantify the discrepancy between each outcome y_t and the forecaster’s

^{*} The first two authors gratefully acknowledge partial support by the PASCAL Network of Excellence under EC grant no. 506778.

LABEL EFFICIENT PREDICTION

Parameters: number N of actions, outcome space \mathcal{Y} , loss function ℓ , time horizon n , budget m of queries.

For each round $t = 1, \dots, n$

- (1) the environment chooses the next outcome $y_t \in \mathcal{Y}$ without revealing it;
- (2) the forecaster chooses an action $I_t \in \{1, \dots, N\}$;
- (3) each action i incurs loss $\ell(i, y_t)$;
- (4) if less than m queries have been issued so far the forecaster may issue a new query to obtain y_t ; if no query is issued then y_t remains unknown.

Fig. 1. Label efficient prediction as a game between the forecaster and the environment.

guess for y_t . Hannan’s seminal result [7] showed that randomized forecasters exist whose excess cumulative loss (or regret), with respect to the loss of any constant forecaster, grows sublinearly in the length n of the target sequence, and this holds for any individual target sequence. In particular, Hannan found the optimal growth rate, $\Theta(\sqrt{n})$, of the regret as a function of the sequence length n when no other assumption other than boundedness is made on the loss ℓ . Only relatively recently, Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth [4] have revealed that the correct dependence on N in the minimax regret rate is $\Theta(\sqrt{n \ln N})$.

Game theorists and learning theorists, who independently studied the sequential prediction model, addressed the fundamental question of whether a sub-linear regret rate is achievable in case the past outcomes y_1, \dots, y_{t-1} are not entirely accessible when computing the guess for y_t . In this work we investigate a variant of sequential prediction known as *label efficient prediction*. In this model, originally proposed by Helmbold and Panizza [8], after choosing its guess at time t the forecaster decides whether to query the outcome y_t . However, the forecaster is limited in the number of queries he can issue within a given time horizon. We prove that a query rate of order $(\ln n)(\ln \ln n)^2/n$ is sufficient for achieving Hannan consistency (i.e., regret growing sub-linearly with probability one). Moreover, we show that any forecaster issuing at most m queries must suffer a regret of at least order $n\sqrt{(\ln N)/m}$ on some outcome sequence of length n , and we show a randomized forecaster achieving this regret to within constant factors. We conclude the paper by proving a label efficient mistake bound for a randomized variant of Littlestone’s zero-threshold Winnow, an algorithm based on exponential weights for binary pattern classification.

2 Sequential prediction and the label efficient model

The sequential prediction problem is parametrized by a number $N > 1$ of player actions, by a set \mathcal{Y} of outcomes, and by a loss function ℓ . The loss function

has domain $\{1, \dots, N\} \times \mathcal{Y}$ and takes values in a bounded real interval, say $[0, 1]$. Given an unknown mechanism adaptively generating a sequence y_1, y_2, \dots of elements from \mathcal{Y} , a prediction strategy, or forecaster, chooses an action $I_t \in \{1, \dots, N\}$ incurring a loss $\ell(I_t, y_t)$. A crucial assumption in this model is that the forecaster can choose I_t only based on information related to the past outcomes y_1, \dots, y_{t-1} . That is, the forecaster's decision must not depend on any of the future outcomes. In the label efficient model, after choosing I_t the forecaster decides whether to issue a query to access y_t . If no query is issued, then y_t remains unknown. In other words, I_t does not depend on all the past outcomes y_1, \dots, y_{t-1} , but only on the queried ones. The label efficient model is best described as a repeated game between the forecaster, choosing actions, and the environment, choosing outcomes (see Figure 1).

3 Regret and Hannan consistency

The cumulative loss of the forecaster on a sequence y_1, y_2, \dots of outcomes is denoted by

$$\widehat{L}_n = \sum_{t=1}^n \ell(I_t, y_t) \quad \text{for } n \geq 1.$$

As our forecasting strategies are randomized, each I_t is viewed as a random variable whose distribution over $\{1, \dots, N\}$ must be fully determined at time t . Without further specifications, all probabilities and expectations will be understood with respect to the σ -algebra of events generated by the sequence I_1, I_2, \dots of the forecaster's random choices. We compare the forecaster's loss \widehat{L}_n with the cumulative losses of the N constant forecasters, $L_{i,n} = \sum_{t=1}^n \ell(i, y_t)$, $i = 1, \dots, N$.

In particular, we devise label efficient forecasting strategies whose expected regret $\mathbb{E} \widehat{L}_n - \min_{i=1, \dots, N} L_{i,n}$ grows sublinearly in n for any individual sequence y_1, y_2, \dots of outcomes. Via a more refined analysis, we also prove the stronger result

$$\widehat{L}_n - \min_{i=1, \dots, N} L_{i,n} = o(n) \quad \text{a.s. ,}$$

for any sequence y_1, y_2, \dots of outcomes, almost surely with respect to the auxiliary randomization the forecaster has access to. This property, known as *Hannan consistency* in game theory, rules out the possibility that the regret is much larger than its expected value with a significant probability.

4 A label efficient forecaster

We start by introducing a simple forecaster whose expected regret is bounded by $n\sqrt{2(\ln N)/m}$, where m is the bound on the number of queries. Thus, if $m = n$ we recover the order of the optimal experts bound. It is easy to see that in order to achieve a nontrivial performance, a forecaster must use randomization in determining whether a label should be revealed or not. It turns out that a simple biased coin does the job. The strategy we propose, sketched in Figure 2,

Parameters: Real numbers $\eta > 0$ and $0 \leq \varepsilon \leq 1$.

Initialization: $\mathbf{w}_1 = (1, \dots, 1)$.

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

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

$$p_{i,t} = \frac{w_{i,t}}{\sum_{j=1}^N w_{j,t}} \quad i = 1, \dots, N$$

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

(3) if $Z_t = 1$ then obtain y_t and compute

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

else, let $\mathbf{w}_{t+1} = \mathbf{w}_t$.

Fig. 2. The label efficient exponentially weighted average forecaster.

uses an i.i.d. sequence Z_1, Z_2, \dots, Z_n of Bernoulli random variables such that $\mathbb{P}[Z_i = 1] = 1 - \mathbb{P}[Z_i = 0] = \varepsilon$ and asks the label y_t to be revealed whenever $Z_t = 1$. Here $\varepsilon > 0$ is a parameter of the strategy. (Typically, we take $\varepsilon \approx m/n$ so that the number of solicited labels during n rounds is about m . Note that this way the forecaster may ask the value of more than m labels but we ignore this detail as it can be dealt with by a simple adjustment.) Our label efficient forecaster uses the *estimated losses*

$$\tilde{\ell}(i, y_t) \stackrel{\text{def}}{=} \begin{cases} \ell(i, y_t)/\varepsilon & \text{if } Z_t = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\mathbb{E}[\tilde{\ell}(i, y_t) \mid Z_1^{t-1}, I_1^{t-1}] = \ell(i, y_t)$, where $Z_1^t = (Z_1, \dots, Z_{t-1})$ and $I_1^t = (I_1, \dots, I_{t-1})$. (The conditioning on Z_1^{t-1} and I_1^{t-1} is merely needed to fix the value of y_t , which may depend on the forecaster's past actions.) Therefore, $\tilde{\ell}(i, y_t)$ may be considered as an unbiased estimate of the true loss $\ell(i, y_t)$. The label efficient forecaster then uses the estimated losses to form an exponentially weighted average forecaster. The expected performance of this strategy may be bounded as follows.

Theorem 1. *Consider the label efficient forecaster of Figure 2 run with $\varepsilon = m/n$ and $\eta = (\sqrt{2m \ln N})/n$. Then, the expected number of revealed labels equals m and*

$$\mathbb{E} \hat{L}_n - \min_{i=1, \dots, N} L_{i,n} \leq n \sqrt{\frac{2 \ln N}{m}}.$$

In the sequel we write \mathbf{p}_t for the N -vector of components $p_{i,t}$. We also use the notation

$$\ell(\mathbf{p}_t, y_t) = \sum_{i=1}^N p_{i,t} \ell(i, y_t) \quad \text{and} \quad \tilde{\ell}(\mathbf{p}_t, y_t) = \sum_{i=1}^N p_{i,t} \tilde{\ell}(i, y_t).$$

Finally, we denote for $i = 1, \dots, N$,

$$\tilde{L}_{i,n} = \sum_{t=1}^n \tilde{\ell}(i, y_t) .$$

Proof. It is enough to adapt the proof of [1, Theorem 3.1], in the following way. First, we note that we have an upper bound over the regret in terms of squares of the losses, see also [12, Theorem 1],

$$\sum_{t=1}^n \tilde{\ell}(\mathbf{p}_t, y_t) - \min_{i=1, \dots, N} \tilde{L}_{i,n} \leq \frac{\ln N}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{j=1}^N p_{j,t} \tilde{\ell}(j, y_t)^2 .$$

Since $\tilde{\ell}(j, y_t) \in [0, 1/\varepsilon]$ for all j and y_t , we finally get

$$\sum_{t=1}^n \tilde{\ell}(\mathbf{p}_t, y_t) \left(1 - \frac{\eta}{2\varepsilon}\right) \leq \tilde{L}_{i,n} + \frac{\ln N}{\eta} \quad i = 1, \dots, N . \quad (1)$$

Taking expectations on both sides and substituting the values of η and ε yields the desired result.

Theorem 1 guarantees that the expected per-round regret converges to zero whenever $m \rightarrow \infty$ as $n \rightarrow \infty$. The next result shows that in fact this regret is, with overwhelming probability, bounded by a quantity proportional to $n\sqrt{(\ln N)/m}$.

Theorem 2. *Let $\delta \in (0, 1)$ and consider the label efficient forecaster of Figure 2 run with parameters*

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

Then, with probability at least $1 - \delta$ the number of revealed labels is at most m and

$$\widehat{L}_n - \min_{i=1, \dots, N} L_{i,n} \leq 2n\sqrt{\frac{\ln N}{m}} + 7n\sqrt{\frac{\ln(4N/\delta)}{m}} .$$

In the full paper, we will prove a more refined bound in which the factors $n\sqrt{(\ln N)/m}$ are replaced by $(1 + o(1))\sqrt{nL^*(\ln N)/m}$ in all cases where L^* , the cumulative loss of the best action, is $\Omega((n/m)\ln N)$. In the cases when L^* is small, then the quantity replacing the above terms is of the order of $(n/m)\ln N$. In particular, we recover the behavior already observed by Helmbold and Panizza [8] in the case $L^* = 0$ (the best expert makes no mistakes).

Even though the label efficient forecaster investigated above assumes the preliminary knowledge of the time horizon n (just note that both η and ε depend on the value of the parameters n and m), using standard adaptive techniques—such as those described in [2]—, a label efficient forecaster may be constructed without knowing n in advance. By letting the query budget m depend on n , one can then achieve Hannan consistency, as stated in the next result.

Corollary 1. *There exists a randomized label efficient forecaster that achieves Hannan consistency while issuing, for all $n > 1$, at most $O((\ln \ln n)^2 \ln n)$ queries in the first n prediction steps.*

Proof. An algorithm that achieves Hannan consistency divides time into consecutive blocks of exponentially increasing length $1, 2, 4, 8, 16, \dots$. In the r -th block (of length 2^{r-1}) it uses the forecaster of Theorem 2 with parameters $n = 2^{r-1}$, $m = (\ln r)(\ln \ln r)$ and $\delta = 1/r^3$. Then, using the bound of Theorem 2 it is easy to see that, with probability one, for all n , the algorithm does not ask for more than $O((\ln \ln n)^2 \ln n)$ labels and the cumulative regret is $o(n)$. Details are omitted. Just note that it is sufficient to prove the statement for $n = 2^{r-1}$ for $r \geq 1$.

Before proving Theorem 2, note that if $\delta \leq 4Ne^{-m/8}$, then the right-hand side of the inequality is greater than n and therefore the statement is trivial. Thus, we may assume throughout the proof that $\delta > 4Ne^{-m/8}$. This also ensures that $\varepsilon > 0$. We need a number of preliminary lemmas. The first is obtained by a simple application of Bernstein's inequality.

Lemma 1. *The probability that the strategy asks for more than m labels is at most $\delta/4$.*

Lemma 2. *With probability at least $1 - \delta/4$,*

$$\sum_{t=1}^n \ell(\mathbf{p}_t, y_t) \leq \sum_{t=1}^n \tilde{\ell}(\mathbf{p}_t, y_t) + 2n\sqrt{\frac{2}{m} \ln \frac{4}{\delta}}.$$

Furthermore, with probability at least $1 - \delta/4$, for all $i = 1, \dots, N$,

$$\tilde{L}_{i,n} \leq L_{i,n} + 2\sqrt{2}n\sqrt{\frac{\ln(4N/\delta)}{m}}.$$

Proof. The proofs of both inequalities rely on Chernoff's bounding. We therefore only prove the first one. Let $s \leq 1$ be a positive number. Define $u = 2\sqrt{\frac{n}{\varepsilon} \ln \frac{4}{\delta}}$ and observe that since $n/m \geq 1/(2\varepsilon)$ (which is implied by the above assumption on δ),

$$\begin{aligned} & \mathbb{P} \left[\sum_{t=1}^n \ell(\mathbf{p}_t, y_t) > \sum_{t=1}^n \tilde{\ell}(\mathbf{p}_t, y_t) + u \right] \\ & \leq \mathbb{E} \left[\exp \left(s \sum_{t=1}^n (\ell(\mathbf{p}_t, y_t) - \tilde{\ell}(\mathbf{p}_t, y_t)) \right) \right] e^{-su} \quad (\text{by Markov's inequality}) \\ & = \mathbb{E} \left[\exp \left(s \sum_{t=1}^{n-1} (\ell(\mathbf{p}_t, y_t) - \tilde{\ell}(\mathbf{p}_t, y_t)) \right) \right. \\ & \quad \left. \times \mathbb{E} \left[\exp \left(s(\ell(\mathbf{p}_n, y_n) - \tilde{\ell}(\mathbf{p}_n, y_n)) \right) \mid Z_1^{n-1}, I_1^{n-1} \right] \right] e^{-su}. \end{aligned}$$

To bound the right-hand side, note that $\ell(\mathbf{p}_n, y_n) - \tilde{\ell}(\mathbf{p}_n, y_n) \leq 1$ and therefore, since we assumed $s \leq 1$,

$$\begin{aligned}
& \mathbb{E} \left[\exp \left(s(\ell(\mathbf{p}_n, y_n) - \tilde{\ell}(\mathbf{p}_n, y_n)) \right) \mid Z_1^{n-1}, I_1^{n-1} \right] \\
& \leq \mathbb{E} \left[1 + s(\ell(\mathbf{p}_n, y_n) - \tilde{\ell}(\mathbf{p}_n, y_n)) + s^2(\ell(\mathbf{p}_n, y_n) - \tilde{\ell}(\mathbf{p}_n, y_n))^2 \mid Z_1^{n-1}, I_1^{n-1} \right] \\
& \quad (\text{since } e^x \leq 1 + x + x^2 \text{ for all } x \leq 1) \\
& = 1 + \mathbb{E} \left[s^2(\ell(\mathbf{p}_n, y_n) - \tilde{\ell}(\mathbf{p}_n, y_n))^2 \mid Z_1^{n-1}, I_1^{n-1} \right] \\
& \quad (\text{since } \mathbb{E}[(\ell(\mathbf{p}_n, y_n) - \tilde{\ell}(\mathbf{p}_n, y_n)) \mid Z_1^{n-1}, I_1^{n-1}] = 0) \\
& \leq 1 + \frac{s^2}{\varepsilon}
\end{aligned}$$

where the last step holds because

$$\mathbb{E} \left[(\ell(\mathbf{p}_n, y_n) - \tilde{\ell}(\mathbf{p}_n, y_n))^2 \mid Z_1^{n-1}, I_1^{n-1} \right] \leq \mathbb{E} \left[\tilde{\ell}(\mathbf{p}_n, y_n)^2 \mid Z_1^{n-1}, I_1^{n-1} \right] \leq 1/\varepsilon .$$

Therefore, using $1 + s^2/\varepsilon \leq e^{s^2/\varepsilon}$, we have

$$\begin{aligned}
& \mathbb{P} \left[\sum_{t=1}^n \ell(\mathbf{p}_t, y_t) > \sum_{t=1}^n \tilde{\ell}(\mathbf{p}_t, y_t) + u \right] \\
& \leq \mathbb{E} \left[\exp \left(s \sum_{t=1}^{n-1} (\ell(\mathbf{p}_t, y_t) - \tilde{\ell}(\mathbf{p}_t, y_t)) \right) \right] e^{s^2/\varepsilon} e^{-su} \\
& \leq e^{ns^2/\varepsilon} e^{-su}
\end{aligned}$$

by repeating the previous argument $n - 1$ times. The value of s minimizing the obtained upper bound is $s = u\varepsilon/2n$ which satisfies the condition $s \leq 1$ because $n \geq m \geq u\varepsilon/2$ due to our assumption on δ . Resubstituting this choice for s we get

$$\mathbb{P} \left[\sum_{t=1}^n \ell(\mathbf{p}_t, y_t) > \sum_{t=1}^n \tilde{\ell}(\mathbf{p}_t, y_t) + u \right] \leq e^{-u^2\varepsilon/(4n)} = \frac{\delta}{4} ,$$

and the proof is completed.

Proof (of Theorem 2). We start again from (1). It remains to show that $\tilde{L}_{i,n}$ is close, with large probability to its expected value $L_{i,n}$ and that $\sum_{t=1}^n \tilde{\ell}(\mathbf{p}_t, y_t)$ is close to $\sum_{t=1}^n \ell(\mathbf{p}_t, y_t) = \tilde{L}_n$.

A straightforward combination of Lemmas 1 and 2 with (1) shows that with probability at least $1 - 3\delta/4$, the strategy asks for at most m labels and has an expected cumulative loss

$$\sum_{t=1}^n \ell(\mathbf{p}_t, y_t) \left(1 - \frac{\eta}{2\varepsilon} \right) \leq \min_{i=1, \dots, N} L_{i,n} + 4\sqrt{2}n \sqrt{\frac{1}{m} \ln \frac{4N}{\delta}} + \frac{\ln N}{\eta} ,$$

which, since $\sum_{t=1}^n \ell(\mathbf{p}_t, y_t) \leq n$, implies

$$\begin{aligned} \sum_{t=1}^n \ell(\mathbf{p}_t, y_t) - \min_{i=1, \dots, n} L_{i,n} &\leq \frac{n\eta}{2\varepsilon} + 4\sqrt{2}n\sqrt{\frac{1}{m} \ln \frac{4N}{\delta}} + \frac{\ln N}{\eta} \\ &= 2n\sqrt{\frac{\ln N}{m}} + 4\sqrt{2}n\sqrt{\frac{1}{m} \ln \frac{4N}{\delta}} \end{aligned}$$

by our choice of η and using $1/(2\varepsilon) \leq n/m$ derived, once more, from our assumption $\delta > 4N e^{-m/8}$. The proof is finished by noting that the Hoeffding-Azuma inequality implies that, with probability at least $1 - \delta/4$,

$$\widehat{L}_n = \sum_{t=1}^n \ell(I_t, y_t) \leq \sum_{t=1}^n \ell(\mathbf{p}_t, y_t) + \sqrt{\frac{n}{2} \ln \frac{4}{\delta}} \leq \sum_{t=1}^n \ell(\mathbf{p}_t, y_t) + n\sqrt{\frac{1}{2m} \ln \frac{4N}{\delta}}$$

since $m \leq n$.

5 A lower bound for label efficient prediction

Here we show that the performance bounds proved in the previous section for the label efficient exponentially weighted average forecaster are essentially unimprovable in the strong sense that no other label efficient forecasting strategy can have a significantly better performance for all problems. Denote the set of natural numbers by $\mathbb{N} = \{1, 2, \dots\}$.

Theorem 3. *There exist an outcome space \mathcal{Y} , a loss function $\ell : \mathbb{N} \times \mathcal{Y} \rightarrow [0, 1]$, and a universal constant $c > 0$ such that, for all $N \geq 2$ and for all $n \geq m \geq 20 \frac{e}{1+e} \ln(N-1)$, the cumulative (expected) loss of any (randomized) forecaster that uses actions in $\{1, \dots, N\}$ and asks for at most m labels while predicting a sequence of n outcomes satisfies the inequality*

$$\sup_{y_1, \dots, y_n \in \mathcal{Y}} \left(\mathbb{E} \left[\sum_{t=1}^n \ell(I_t, y_t) \right] - \min_{i=1, \dots, N} \sum_{t=1}^n \ell(i, y_t) \right) \geq cn \sqrt{\frac{\ln(N-1)}{m}}.$$

In particular, we prove the theorem for $c = \frac{\sqrt{e}}{(1+e)\sqrt{5(1+e)}}$.

Proof. First, we define $\mathcal{Y} = [0, 1]$ and ℓ . Given $y \in [0, 1]$, we denote by (y_1, y_2, \dots) its dyadic expansion, that is, the unique sequence not ending with infinitely many zeros such that

$$y = \sum_{k \geq 1} y_k 2^{-k}.$$

Now, the loss function is defined as $\ell(k, y) = y_k$ for all $y \in \mathcal{Y}$ and $k \in \mathbb{N}$.

We construct a random outcome sequence and show that the expected value of the regret (with respect both to the random choice of the outcome sequence

and to the forecaster's possibly random choices) for any possibly randomized forecaster is bounded from below by the claimed quantity.

More precisely, we denote by U_1, \dots, 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 uniformly distributed random variables over $[0, 1]$. Our underlying probability space is equipped with the σ -algebra of events generated by the random outcome sequence Y_1, \dots, Y_n and by the randomization U_1, \dots, U_n . The random outcome sequence is independent of the auxiliary randomization: we define N different probability distributions, $\mathbb{P}_i \otimes \mathbb{P}_A$, $i = 1, \dots, N$, formed by the product of the auxiliary randomization (whose associated probability distribution is denoted by \mathbb{P}_A) and one of the N different probability distributions $\mathbb{P}_1, \dots, \mathbb{P}_N$ over the outcome sequence defined as follows.

For $i = 1, \dots, N$, \mathbb{Q}_i is defined as the distribution (over $[0, 1]$) of

$$Z^* 2^{-i} + \sum_{k=1, \dots, N, k \neq i} Z_k 2^{-k} + 2^{-(N+1)} U,$$

where U, Z^*, Z_1, \dots, Z_N are independent random variables such that U has uniform distribution, and Z^* and the Z_k have Bernoulli distribution with parameter $1/2 - \epsilon$ for Z^* and $1/2$ for the Z_k . Now, the randomization is such that under \mathbb{P}_i , the outcome sequence Y_1, \dots, Y_n is i.i.d. with common distribution \mathbb{Q}_i .

Then, under each \mathbb{P}_i (for $i = 1, \dots, N$), the losses $\ell(k, Y_t)$, $k = 1, \dots, N$, $t = 1, \dots, n$, are i.i.d. Bernoulli random variables. In addition, $\ell(i, Y_t) = 1$ with probability $1/2 - \epsilon$ and $\ell(k, Y_t) = 1$ with probability $1/2$ for each $k \neq i$, where ϵ is a positive number specified below.

We have

$$\begin{aligned} \max_{y_1, \dots, y_n} \left(\mathbb{E}_A \widehat{L}_n - \min_{i=1, \dots, N} L_{i,n} \right) &= \max_{y_1, \dots, y_n} \max_{i=1, \dots, N} \left(\mathbb{E}_A \widehat{L}_n - L_{i,n} \right) \\ &\geq \max_{i=1, \dots, N} \mathbb{E}_i \left[\mathbb{E}_A \widehat{L}_n - L_{i,n} \right], \end{aligned}$$

where \mathbb{E}_i (resp. \mathbb{E}_A) denotes expectation with respect to \mathbb{P}_i (resp. \mathbb{P}_A).

Now, we use the following decomposition lemma, which states that a randomized algorithm performs, on the average, just as a convex combination of deterministic algorithms. The simple but cumbersome proof is omitted from this extended abstract.

Lemma 3. *For any given randomized forecaster, there exists an integer D , a point $\alpha = (\alpha_1, \dots, \alpha_D) \in \mathbb{R}^D$ in the probability simplex, and D deterministic algorithms (indexed by a superscript $d = 1, \dots, D$) such that, for every t and every possible outcome sequence $y_1^{t-1} = (y_1, \dots, y_{t-1})$,*

$$\mathbb{P}_A [I_t = i | y_1^{t-1}] = \sum_{d=1}^D \alpha_d \mathbb{I}_{[I_t^d = i | y_1^{t-1}]},$$

where $\mathbb{I}_{[I_t^d = i | y_1^{t-1}]}$ is the indicator function that the d -th deterministic algorithm chooses action i when the sequence of past outcomes is formed by y_1^{t-1} .

Using this lemma, we have that there exist D , α and D deterministic sub-algorithms such that

$$\begin{aligned} \max_{i=1,\dots,N} \mathbb{E}_i \left[\mathbb{E}_A \widehat{L}_n - L_{i,n} \right] &= \max_{i=1,\dots,N} \mathbb{E}_i \left[\sum_{t=1}^n \sum_{d=1}^D \alpha_d \sum_{k=1}^N \mathbb{I}_{[I_t^d=k | Y_1^{t-1}]} \ell(k, Y_t) - L_{i,n} \right] \\ &= \max_{i=1,\dots,N} \sum_{d=1}^D \alpha_d \mathbb{E}_i \left[\sum_{t=1}^n \sum_{k=1}^N \mathbb{I}_{[I_t^d=k | Y_1^{t-1}]} \ell(k, Y_t) - L_{i,n} \right] \end{aligned}$$

Now, under \mathbb{P}_i the regret grows by ε whenever an action different from i is chosen and remains the same otherwise. Hence,

$$\begin{aligned} \max_{i=1,\dots,N} \mathbb{E}_i \left[\mathbb{E}_A \widehat{L}_n - L_{i,n} \right] &= \max_{i=1,\dots,N} \sum_{d=1}^D \alpha_d \mathbb{E}_i \left[\sum_{t=1}^n \sum_{k=1}^N \mathbb{I}_{[I_t^d=k | Y_1^{t-1}]} \ell(k, Y_t) - L_{i,n} \right] \\ &= \varepsilon \max_{i=1,\dots,N} \sum_{d=1}^D \alpha_d \sum_{t=1}^n \mathbb{P}_i [I_t^d \neq i] \\ &= \varepsilon n \left(1 - \min_{i=1,\dots,N} \sum_{d=1}^D \sum_{t=1}^n \frac{\alpha_d}{n} \mathbb{P}_i [I_t^d = i] \right). \end{aligned}$$

For the d -th deterministic subalgorithm, let $1 \leq T_1^d < \dots < T_m^d \leq n$ be the times when the m queries were issued. Then T_1^d, \dots, T_m^d are finite stopping times with respect to the i.i.d. process Y_1, \dots, Y_n . Hence, by a well-known fact in probability theory (see, e.g., [5, Lemma 2, page 138]), the revealed outcomes $Y_{T_1^d}, \dots, Y_{T_m^d}$ are independent and identically distributed as Y_1 .

Let R_t^d be the number of revealed outcomes at time t and note that R_t^d is measurable with respect to the random outcome sequence. Now, as the subalgorithm we consider is deterministic, R_t^d is fully determined by $Y_{T_1^d}, \dots, Y_{T_m^d}$. Hence, I_t^d may be seen as a function of $Y_{T_1^d}, \dots, Y_{T_m^d}$ rather than a function of $Y_{T_1^d}, \dots, Y_{R_t^d}$ only. This essentially means that the knowledge of the extra values cannot hurt in the sense that it cannot lead the forecaster to choose different actions. As the joint distribution of $Y_{T_1^d}, \dots, Y_{T_m^d}$ under \mathbb{P}_i is \mathbb{Q}_i^m , we have proven indeed that

$$\mathbb{P}_i [I_t^d = i] = \mathbb{Q}_i^m [I_t^d = i].$$

Consequently, our lower bound rewrites as

$$\max_{i=1,\dots,N} \mathbb{E}_i \left[\mathbb{E}_A \widehat{L}_n - L_{i,n} \right] = \varepsilon n \left(1 - \min_{i=1,\dots,N} \sum_{d=1}^D \sum_{t=1}^n \frac{\alpha_d}{n} \mathbb{Q}_i^m [I_t^d = i] \right).$$

By the generalized Fano's lemma (see Lemma 5 in the Appendix), it is guaranteed that

$$\min_{i=1,\dots,N} \sum_{d=1}^D \sum_{t=1}^n \frac{\alpha_d}{n} \mathbb{Q}_i^m [I_t^d = i] \leq \max \left\{ \frac{e}{1+e}, \frac{\bar{K}}{\ln(N-1)} \right\},$$

where

$$\bar{K} = \sum_{t=1}^n \sum_{d=1}^D \sum_{i=2}^N \frac{\alpha_d}{n(N-1)} \text{KL}(\mathbb{Q}_i^m, \mathbb{Q}_1^m) = \frac{1}{N-1} \sum_{i=2}^N \text{KL}(\mathbb{Q}_i^m, \mathbb{Q}_1^m),$$

and KL is the Kullback-Leibler divergence (or relative entropy) between two probability distributions.

Moreover, \mathbb{B}_p denoting the Bernoulli distribution with parameter p ,

$$\begin{aligned} \text{KL}(\mathbb{Q}_i^m, \mathbb{Q}_1^m) &= m \text{KL}(\mathbb{Q}_i, \mathbb{Q}_1) \leq m (\text{KL}(\mathbb{B}_{1/2-\varepsilon}, \mathbb{B}_{1/2}) + \text{KL}(\mathbb{B}_{1/2}, \mathbb{B}_{1/2-\varepsilon})) \\ &= m \varepsilon \ln \left(1 + \frac{4\varepsilon}{1-2\varepsilon} \right) \leq 5m \varepsilon^2 \end{aligned}$$

for $0 \leq \varepsilon \leq 1/10$, where the first inequality holds by noting that the definition of the \mathbb{Q}_i implies that the considered Kullback-Leibler divergence is upper bounded by the Kullback-Leibler divergence between $(Z_1, \dots, Z^*, \dots, Z_n, U)$, where Z^* is in the i -th position, and $(Z^*, Z_2, \dots, Z_n, U)$. Therefore,

$$\max_{y_1, \dots, y_n} \left(\mathbb{E}_A \widehat{L}_n - \min_{i=1, \dots, N} L_{i,n} \right) \geq \varepsilon n \left(1 - \max \left\{ \frac{e}{1+e}, \frac{5m \varepsilon^2}{\ln(N-1)} \right\} \right).$$

The choice

$$\varepsilon = \sqrt{\frac{e \ln(N-1)}{5(1+e)m}}$$

yields the claimed bound.

6 A label efficient algorithm for pattern classification

So far, we have shown that exponentially weighted average forecasters can be made label efficient without losing important properties, such as Hannan consistency. In this section we move away from the abstract sequential decision problem defined in Section 2 and show that the idea of label efficient prediction finds interesting applications in more concrete pattern classification problems. More specifically, consider the problem of predicting the binary labels of an arbitrarily chosen sequence $\mathbf{x}_1, \mathbf{x}_2, \dots \in \mathbb{R}^d$ of instances where, for each $t = 1, 2, \dots$, the label $y_t \in \{-1, 1\}$ of \mathbf{x}_t satisfies $y_t \mathbf{u} \cdot \mathbf{x}_t > 0$. Here $\mathbf{u} \in \mathbb{R}^d$ is a fixed but unknown linear separator for the labeled sequence. In this framework, we show that the zero-threshold Winnow algorithm of Littlestone [10], a natural extension to pattern classification of the exponentially weighted average forecaster, can be made label efficient. In particular, for the label efficient variant of this algorithm (described in Figure 3) we prove an expected mistake bound exactly equal to the mistake bound of the original zero-threshold Winnow. In addition, unlike the algorithms shown in previous sections, in our variant the probability of querying a label is a function of the previously observed instances and previously queried labels.

Algorithm Label efficient zero-threshold Winnow

Parameters $\eta > 0$

Initialization $w_{i,1} = 1$ for $i = 1, \dots, N$

For $t = 1, 2, \dots$

1. get $\mathbf{x}_t \in \mathbb{R}^d$, define \mathbf{p}_t by $p_{i,t} = w_{i,t}/W_t$, where $W_t = \sum_{i=1}^N w_{i,t}$, and let $q_t = \mathbf{p}_t \cdot \mathbf{x}_t$
2. predict with $\hat{y}_t = \text{sgn}(q_t)$
3. draw a Bernoulli variable Z_t of parameter $(2|q_t|/\gamma + 1)^{-1}$.
4. if $Z_t = 1$, then
 - (a) get $y_t \in \{-1, 1\}$.
 - (b) if $\hat{y}_t \neq y_t$, then let $w_{i,t+1} = w_{i,t} e^{\eta y_t x_{i,t}}$ for all $i = 1, \dots, N$
5. else, $w_{i,t+1} = w_{i,t}$ for all $i = 1, \dots, N$.

Fig. 3. The randomized label-efficient zero-threshold Winnow.

Theorem 4. *Pick any sequence $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$ such that, for all $t = 1, \dots, n$, $y_t \mathbf{u} \cdot \mathbf{x}_t \geq \gamma$ for some $\gamma > 0$ and some vector \mathbf{u} from the probability simplex in \mathbb{R}^d . Let X_∞ be any number such that $\max_t \|\mathbf{x}_t\|_\infty \leq X_\infty$. Then the randomized label efficient zero-threshold Winnow algorithm of Figure 3, run with parameter $\eta = \gamma/X_\infty^2$, makes an expected number of mistakes bounded by $(2X_\infty^2 \ln N)/\gamma^2$ while querying an expected number of labels equal to $\sum_{t=1}^n (2|q_t|/\gamma + 1)^{-1}$.*

The dependence of η on γ is inherited from the original Winnow algorithm and is not caused by the label efficient framework. Note also that, while the expected mistake bound is the same as the mistake bound for the original zero-threshold Winnow, the probability of querying a label at step t attains 1 as the “margin” $|q_t|$ shrinks to 0, and attains $(2X_\infty/\gamma + 1)^{-1}$ as $|q_t|$ grows to its maximum value X_∞ . Obtaining an explicit bound on the expected number of queried labels appears hard as q_t depends in a complicated way on the structure of the labeled sequence. Hence, the result demonstrates that the label efficient framework in this case does provide an advantage (in expectation), even though the theoretical assessment of this advantage appears to be problematic.

Proof. Let M_t be the indicator function for a mistake in step t . Pick a step t such that M_t and Z_t are both 1. Then,

$$\ln \frac{W_{t+1}}{W_t} = \ln \left(\sum_{i=1}^N p_{i,t} e^{\eta y_t x_{i,t}} \right) \leq \eta y_t \mathbf{p}_t \cdot \mathbf{x}_t + \frac{\eta^2}{2} X_\infty^2 = -\eta |q_t| + \frac{\eta^2}{2} X_\infty^2$$

where the inequality is an application of the Hoeffding inequality [9] while the last equality holds because $M_t = 1$ implies $y_t q_t \leq 0$. On the other hand, if M_t or Z_t is 0 at step t , then $W_{t+1} = W_t$ and thus $\ln(W_{t+1}/W_t) = 0$. Summing for

$t = 1, \dots, n$ we get

$$\ln \frac{W_{n+1}}{W_1} \leq \eta \sum_{t=1}^n \left(\frac{\eta}{2} X_\infty^2 - |q_t| \right) M_t Z_t \quad (2)$$

Now consider any vector \mathbf{u} of convex coefficients such that $y_t \mathbf{u} \cdot \mathbf{x}_t \geq \gamma$ for all $t = 1, \dots, n$. Let

$$\mathbf{R} = \sum_{t=1}^n (y_t \mathbf{x}_t) M_t Z_t .$$

Using the log-sum inequality [6], and recalling that $y_t \mathbf{u} \cdot \mathbf{x}_t \geq \gamma$ for all t ,

$$\begin{aligned} \ln \frac{W_{n+1}}{W_1} &= -\ln N + \ln \sum_{i=1}^N e^{\eta R_i} \geq -\ln N + \eta \mathbf{R} \cdot \mathbf{u} + H(\mathbf{u}) \\ &\geq -\ln N + \eta \gamma \sum_{t=1}^n M_t Z_t + H(\mathbf{u}) . \end{aligned} \quad (3)$$

Dropping $H(\mathbf{u}) \geq 0$, the entropy of \mathbf{u} , from (2) and (3) we obtain

$$-\ln N + \eta \gamma \sum_{t=1}^n M_t Z_t \leq \eta \sum_{t=1}^n \left(\frac{\eta}{2} X_\infty^2 - |q_t| \right) M_t Z_t .$$

Dividing by $\eta > 0$ and rearranging yields

$$\sum_{t=1}^n \left(\gamma - \frac{\eta}{2} X_\infty^2 + |q_t| \right) M_t Z_t \leq \frac{\ln N}{\eta} .$$

Replacing η with γ/X_∞^2 gets us

$$\sum_{t=1}^n \left(\frac{\gamma}{2} + |q_t| \right) M_t Z_t \leq \frac{X_\infty^2 \ln N}{\gamma} . \quad (4)$$

Now recall that $\mathbb{E}[Z_t | Z_1, \dots, Z_{t-1}] = (2|q_t|/\gamma + 1)^{-1}$, where the conditioning is needed as q_t is a function of Z_1, \dots, Z_{t-1} . Taking expectation on both sides of (4) yields

$$\begin{aligned} \frac{X_\infty^2 \ln N}{\gamma} &\geq \mathbb{E} \left[\sum_{t=1}^n \left(\frac{\gamma}{2} + |q_t| \right) M_t Z_t \right] \\ &= \mathbb{E} \left[\sum_{t=1}^n \left(\frac{\gamma}{2} + |q_t| \right) M_t \mathbb{E}[Z_t | Z_1, \dots, Z_{t-1}] \right] \\ &= \mathbb{E} \left[\sum_{t=1}^n \left(\frac{\gamma}{2} + |q_t| \right) \frac{M_t}{2|q_t|/\gamma + 1} \right] = \frac{\gamma}{2} \mathbb{E} \left[\sum_{t=1}^n M_t \right] . \end{aligned}$$

Multiplying both sides by $2/\gamma$ gets us the desired result.

References

1. P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.
2. P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1), 2002.
3. L. Birgé. A new look at an old result: Fano’s lemma. Technical report, Université Paris 6. 2001.
4. N. Cesa-Bianchi, Y. Freund, D. Haussler, D.P. Helmbold, R. Schapire, and M.K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.
5. Y.S. Chow and H. Teicher. *Probability Theory*. Springer, 1988.
6. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, 1991.
7. J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the theory of games*, 3:97–139, 1957.
8. 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.
9. W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
10. N. Littlestone. *Mistake Bounds and Logarithmic Linear-threshold Learning Algorithms*. PhD thesis, University of California at Santa Cruz, 1989.
11. P. Massart. *Concentration inequalities and model selection*. Saint-Flour summer school lecture notes, 2003. To appear.
12. A. Piccolboni and C. Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Proceedings of the 14th Annual Conference on Computational Learning Theory*, pages 208–223, 2001.

A Technical lemmas

The crucial point in the proof of the lower bound theorem is an extension of Fano’s lemma to a convex combination of probability masses, which may be proved thanks to a straightforward modification of the techniques developed by Birgé [3] (see also Massart [11]). Recall first a consequence of the variational formula for entropy.

Lemma 4. *For arbitrary probability distributions \mathbb{P}, \mathbb{Q} and for each $\lambda > 0$,*

$$\lambda \mathbb{P}[A] - \psi_{\mathbb{Q}[A]}(\lambda) \leq \text{KL}(\mathbb{P}, \mathbb{Q})$$

where $\psi_p(\lambda) = \ln(p(e^\lambda - 1) + 1)$.

Lemma 5 (Generalized Fano). *Let $\{A_{s,j} : s = 1, \dots, S, j = 1, \dots, N\}$ be a family of subsets of a set Ω such that $A_{s,1}, \dots, A_{s,N}$ form a partition of Ω for each fixed s . Let $\alpha_1, \dots, \alpha_S$ be such that $\alpha_s \geq 0$ for $s = 1, \dots, S$ and $\alpha_1 + \dots + \alpha_S = 1$. Then, for all sets $\mathbb{P}_{s,1}, \dots, \mathbb{P}_{s,N}$, $s = 1, \dots, S$, of probability distributions on Ω ,*

$$\min_{j=1, \dots, N} \sum_{s=1}^S \alpha_s \mathbb{P}_{s,j}[A_{s,j}] \leq \max \left\{ \frac{e}{1+e}, \frac{\bar{K}}{\ln(N-1)} \right\},$$

where

$$\bar{K} = \sum_{s=1}^S \sum_{j=2}^N \frac{\alpha_s}{N-1} \text{KL}(\mathbb{P}_{s,j}, \mathbb{P}_{s,1}) .$$

Proof. Using Lemma 4, we have that

$$\begin{aligned} \sum_{s=1}^S \sum_{j=2}^N \frac{\alpha_s}{N-1} \lambda \mathbb{P}_{s,j}[A_{s,j}] - \sum_{s=1}^S \sum_{j=2}^N \frac{\alpha_s}{N-1} \psi_{\mathbb{P}_{s,1}[A_{s,j}]}(\lambda) \\ \leq \sum_{s=1}^S \sum_{j=2}^N \frac{\alpha_s}{N-1} \text{KL}(\mathbb{P}_{s,j}, \mathbb{P}_{s,1}) = \bar{K} . \end{aligned}$$

Now, for each fixed $\lambda > 0$, the function that maps p to $-\psi_p(\lambda)$ is convex. Hence, letting

$$p_1 = \sum_{s=1}^S \sum_{j=2}^N \frac{\alpha_s}{N-1} \mathbb{P}_{s,1}[A_{s,j}] = \frac{1}{N-1} \left(1 - \sum_{s=1}^S \alpha_s \mathbb{P}_{s,1}[A_{s,1}] \right) ,$$

by Jensen's inequality we get

$$\begin{aligned} \sum_{s=1}^S \sum_{j=2}^N \frac{\alpha_s}{N-1} \lambda \mathbb{P}_{s,j}[A_{s,j}] - \psi_{p_1}(\lambda) \\ \leq \sum_{s=1}^S \sum_{j=2}^N \frac{\alpha_s}{N-1} \lambda \mathbb{P}_{s,j}[A_{s,j}] - \sum_{s=1}^S \sum_{j=2}^N \frac{\alpha_s}{N-1} \psi_{\mathbb{P}_{s,1}[A_{s,j}]}(\lambda) . \end{aligned}$$

Recalling that the right-hand side of the above inequality above is less than \bar{K} , and introducing the quantities

$$a_j = \sum_{s=1}^S \alpha_s \mathbb{P}_{s,j}[A_{s,j}] \quad \text{for } j = 1, \dots, N,$$

we conclude

$$\lambda \min_{j=1, \dots, N} a_j - \psi_{\frac{1-a_j}{N-1}}(\lambda) \leq \lambda \frac{1}{N-1} \sum_{j=2}^N a_j - \psi_{\frac{1-a_1}{N-1}}(\lambda) \leq \bar{K} .$$

Denote by a the minimum of the a_j 's and let $p^* = (1-a)/(N-1) \geq p_1$. We only have to deal with the case when $a \geq e/(1+e)$. As for all $\lambda > 0$, the function that maps p to $-\psi_p$ is decreasing, we have

$$\bar{K} \geq \sup_{\lambda > 0} (\lambda a - \psi_{p^*}(\lambda)) \geq a \ln \frac{a}{e p^*} \geq a \ln \frac{a(N-1)}{(1-a)e} \geq a \ln(N-1) ,$$

whenever $p^* \leq a \leq 1$ for the second inequality to hold, and by using $a \geq e/(1+e)$ for the last one. As $p^* \leq 1/(N-1) \leq e/(1+e)$ whenever $N \geq 3$, the case $a < p^*$ may only happen when $N = 2$, but then the result is trivial.