The Combination of the Label Efficient and the Multi-Armed Bandit Problem in Adversarial Setting

György Ottucsák[1] and András György[2]

[1] Department of Computer Science and Information Theory,
Budapest University of Technology and Economics,
Stoczek u. 2, Budapest, Hungary H-1521
oti@szit.bme.hu
[2] Informatics Laboratory, Computer and Automation Research Institute
of the Hungarian Academy of Sciences,
Lágymányosi u. 11, Budapest, Hungary, H-1111
gya@szit.bme.hu

Abstract. In sequential prediction (decision) problems in general, a forecaster has to make a sequence of decisions. After each decision the forecaster suffers some loss, depending on the response of the environment, and the goal of the forecaster is to minimize its cumulative loss over a sufficiently long period of time. In the adversarial setting no probabilistic assumption is made on how the loss of the forecaster is generated, and the goal of the forecaster is to perform well relative to a set of experts. To solve this problem, the forecaster has access to the decisions of the experts before making his own. Although it is impossible to know in advance the performance of the experts, under general conditions the experts' advice can be combined such that the forecaster's average loss is asymptotically not larger than that of the best expert.
These combination algorithms are usually based on the past performance of the experts. However, in certain type of problems it is not possible to obtain all the losses corresponding to the decisions of the experts. In the so called multi-armed bandit problem the forecaster has only information on the loss of the chosen action, and no information is available about the loss it would have suffered had it made a different decision. Another example is label efficient prediction where it is expensive to obtain the losses of the experts, and therefore the forecaster has the option to query this information. In this paper we investigate the combination of the label efficient and the multi-armed bandit problem, where after choosing a decision, the forecaster learns its own loss if and only if it asks for it, which cannot be done too often. This combination is motivated by adaptive routing applications in certain packet networks, such as cognitive packet networks.

1 Introduction

In sequential decision (prediction) problems in general, a decision maker has to make a sequence of decisions. After each decision the decision maker suffers some

loss, depending on the response of the environment, and the goal of the decision maker is to minimize its cumulative loss over a sufficiently long period of time. In the adversarial setting no probabilistic assumption is made on how the loss of the decision maker is generated, and the goal of the decision maker is to perform well relative to a set of experts. More precisely, the aim of the decision maker is to achieve asymptotically the same average loss as the best experts. To solve this problem, the decision maker has access to the decisions of the experts before making his own, and hence can combine them. However, it is impossible to know in advance the performance of the experts.

The first theoretical results concerning sequential prediction are due to Blackwell [1] and Hannan [2], but they were rediscovered by the learning community only in the 1990's, see, for example, Vovk [3], Littlestone and Warmuth [4] and Cesa-Bianchi *et al.* [5]. These results show that it is possible the construct algorithms for sequential (online) prediction that predict almost as well as the best expert. The main idea of these algorithms is the same: after observing the past performance of the experts, in each step the decision of a randomly chosen expert is followed such that experts with superior past performance are chosen with higher probability.

However, in certain type of problems it is not possible to obtain full information on the past performance of the experts. For example, in many situations the decision maker has only information on the loss of the chosen action, and no information is available about the loss it would have suffered had it made a different decision. This is called the *multi-armed bandit problem*. Another example is when it is expensive to obtain the losses of the experts, and therefore the decision maker has the option to query this information. In typical cases this corresponds to the response of the environment, also called as outcome or label, from which it is possible to compute the loss of each expert. This type of problem is called *label efficient prediction*. In all of these problems, including the full information case, when the loss of each expert is revealed after the decision of the decision maker, algorithms whose cumulative loss in n steps exceed the cumulative loss of the best of N experts by an amount of $O(\sqrt{nN \log N/\overline{m}})$, where $\overline{m}$ is the average number of the experts whose loss is revealed to the decision maker in one round. That is, in the full information case this bound becomes $O(\sqrt{n \log N})$, for the multi-armed bandit problem it is $O(\sqrt{nN \log N})$, and for the label efficient prediction problem with m query in n rounds is $O(n\sqrt{\log N/m})$ (for a good survey on this topic, the reader is referred to, e.g., the recent book of Cesa-Bianchi and Lugosi [6]).

The routing problem in communication networks can naturally be fitted in the above prediction framework.

Example 1. The Cognitive Packet Networks (CPN) is introduced by Gelenbe *et al.* [7, 8] in which capabilities for routing and flow control are concentrated in the packets, rather in the nodes and protocols there are three types of the packets: smart packets, dumb packets and acknowledgements. The *smart packets* explore the network (only the chosen path) but they do not transport any data. The *dumb packets* are given the path to follow to their destination by the source (no

gather information) but they transport data. Let the possible paths be between two dedicated nodes the experts and the decision maker who would like to find the path with the smallest delay to the destination.

We assume that the decision maker knows the topology of the network, but the delays on the edges are dynamically changes. In that case the physical size of smart packet is the cost of the information.

If only smart packets are sent in each rounds then the decision maker obtain the bandit setting, but does not send any useful information through the network. So in practice the proportion to the smart packets have to be decreased, i.e., the decision maker sends smart packet only if it queries the "label". Thus, the CPN is the combination of the label efficient and bandit setting.

2 The model

The sequential prediction problem is characterized by a set $\mathcal{Y}$ of outcomes, by $\mathcal{D}$ the decision space, by the experts $\{f_{i,t}\}_{1\leq i\leq N, 1\leq t\leq n}$ and by a loss function ℓ. The advice of expert i is $f_{i,t} \in \mathcal{D}$ at t for all $i = 1, \ldots, N$. The performance of the decision maker and the experts are scored using a loss function $\ell : \mathcal{D} \times \mathcal{Y} \to [0, 1]$. The loss of expert i is $\ell(f_{i,t}, y_t)$ and loss of the decision maker is $\ell(f_{I_t,t}, y_t)$, where $I_t \in 1, \ldots, N$ is a random variable (an expert is chosen by the decision maker at t). I_t only depends on the past outcomes $y_{t-1}, \ldots, y_1$ and the earlier choice of the decision maker $I_{t-1}, \ldots, I_1$.

Throughout the paper we assume that the experts are static, i.e., $f_{i,t} = i \in \{1, \ldots, N\}$ for all t. For convenience we use the notations $\ell_{i,t}$ instead of $\ell(f_{i,t}, y_t)$ and $\ell_{I_t,t}$ instead of $\ell(f_{I_t,t}, y_t)$. The cumulative loss of the decision maker is

$$\widehat{L}_n = \sum_{t=1}^{n} \ell_{I_t,t},$$

and the cumulative loss of the expert i up to n is

$$L_{i,n} = \sum_{t=1}^{n} \ell_{i,t}.$$

The aim of the learning algorithm is to find a decision maker for which

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n}$$

is small universally for all possible sequence of $\{y_t\}$.

2.1 The case of full information: exponentially weighted average algorithm

In full information case the decision maker allows to observe the performance of each experts after each rounds (Algorithm 1).

Algorithm 1 *(Full Information) Fix* $\eta > 0$.
Initialization*:* $w_{i,0} = 1$ *and* $p_{i,1} = 1/N$ *for* $i = 1, \ldots, N$.
For each round $t = 1, 2, \ldots$

(1) Select an expert $I_t \in \{1, \ldots, N\}$ *according to the probability distribution* $\mathbf{p}_t = (p_{1,t}, \ldots, p_{N,t})$*;*
(2) Update the weights $w_{i,t} = w_{i,t-1} e^{-\eta \ell_{i,t}}$*;*
(3) Calculate the updated probability distribution

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

Fig. 1. Exponentially weighted average algorithm in full information case

The maximum difference between the cumulative loss of the decision maker and cumulative loss of the best expert is $O(\sqrt{n \ln N})$ was proved by Kivinen and Warmuth [4]:

Theorem 1. *Let* $n, N \geq 1$ *and* $0 < \delta < 1$*. The exponentially weighted average algorithm (Algorithm 1) with* $\eta = \sqrt{8 \ln N / n}$ *satisfies, with probability at least* $1 - \delta$*,*

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq \sqrt{\frac{n \ln N}{2}} + \sqrt{\frac{n}{2} \ln \frac{1}{\delta}}.$$

2.2 Partial information: label efficient prediction

In case of label efficient prediction after choosing its action at time t, the decision maker decides whether to query the "label" y_t. For query a label the decision maker uses i.i.d. sequence $S_1, S_2, \ldots, S_n$ of Bernoulli random variables such that $\mathbb{P}\{S_t = 1\} = \epsilon$ and asks label y_t if $S_t = 1$. With knowing y_t one can calculate all $\ell_{i,t}$ for all $i = 1, \ldots, N$. Typically, $\epsilon \approx m/n$, so the number of the revealed labels during n rounds is approximately m, where $m \leq n$.

In order to apply the exponential weighted average decision maker in this case the losses have to been modified namely if $S_t = 1$ then the decision maker gets all of the information, if $S_t = 0$ then gets no information. In the Algorithm 2 estimated losses are used instead of observed losses

$$\widetilde{\ell}_{i,t} = \begin{cases} \frac{\ell_{i,t}}{\epsilon}, \text{ if } S_t = 1, \\ 0, \quad \text{otherwise}, \end{cases}$$

which is an unbiased estimate of the true losses $(\ell_{i,t})$

$$\mathbb{E}\left[\widetilde{\ell}_{i,t} \middle| S_1^{t-1}, I_1^{t-1}\right] = \ell_{i,t}.$$

The upper bound of the difference of the best expert and label efficient decision maker is $O(n\sqrt{\ln(4N/\delta)/m})$ was proved by Cesa-Bianchi *et al.* [9]:

Algorithm 2 *(Label Efficient) Fix $\eta > 0$ and $0 < \epsilon \leq 1$.*
Initialization*: $w_{i,0} = 1$ and $p_{i,1} = 1/N$ for $i = 1, \ldots, N$.*
For each round $t = 1, 2, \ldots$

(1) Select an action $I_t \in \{1, \ldots, N\}$ according to the probability distribution $\mathbf{p}_t = (p_{1,t}, \ldots, p_{N,t})$;
(2) Draw a Bernoulli random variable S_t such that $\mathbb{P}\{S_t = 1\} = \epsilon$;
(3) if $S_t = 1$ then obtain $\ell_{i,t}$ for all i and compute the estimated loss $(\widetilde{\ell}_{i,t})$

$$\widetilde{\ell}_{i,t} = \begin{cases} \frac{\ell_{i,t}}{\epsilon}, \text{ if } S_t = 1, \\ 0, \quad \text{otherwise}; \end{cases}$$

(4) Update the weights $w_{i,t} = w_{i,t-1} e^{-\eta \widetilde{\ell}_{i,t}}$;
(5) Calculate the updated probability distribution

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

Fig. 2. Label efficient exponentially weighted average algorithm

Theorem 2. *Let $n, N \geq 1$ and $0 < \delta < 1$. The label efficient exponentially weighted average algorithm (Algorithm 2) with parameters*

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

Then, with probability at least $1 - \delta$,

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq 2n\sqrt{\frac{\ln N}{m}} + 6n\sqrt{\frac{\ln(4N/\delta)}{m}},$$

where m is the average number of the revealed labels.

Corollary 1. *Let $n, N \geq 1$ and $0 < \delta < 1$. For any $n \geq 2\ln(4/\delta)$ the label efficient exponentially weighted average forecaster (Algorithm 2) with parameters*

$$0 < \epsilon \leq 1 - \sqrt{\frac{2\ln(4/\delta)}{n}} \text{ and } \eta = \sqrt{\frac{2\epsilon \ln N}{n}}.$$

Then with probability at least $1 - \delta$,

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq 2\sqrt{\frac{n \ln N}{\epsilon}} + 6\sqrt{\frac{n\ln(4N/\delta)}{\epsilon}}.$$

2.3 Partial information: multi-armed bandit problem

In the multi-armed bandit problem the forecaster after choosing an expert, learns the own loss $\ell_{I_t,t}$, but not the other value of the losses $\ell_{i,t}$. Thus, the forecaster

does not have access to the losses it would have suffered if it had chosen a different expert. It means that the forecaster observes only a piece of information per rounds. The lack of the information implies a natural strategy namely at the beginning of the game the forecaster has to explore the losses of the experts (*exploration phase*) and then may keep choosing the expert with smallest estimated loss for the remaining time (*the exploitation phase*).

In the classical formulation of multi-armed bandit problems (see, e.g., Robbins [10]) it is assumed that, for each action, the losses are randomly and independently drawn with respect to a fixed but unknown distribution. This version is stochastic multi-armed bandit problem. Here we investigate a more challenge problem is analyzed by Auer *et al.* [11], when the outcomes are generated in an adversary opponent (non-stochastic or adversarial multi-armed bandit problem).

There are three modifications according to the full information case. First of all, the modified strategy uses estimated *gains* instead of losses. We introduce notation

$$g_{i,t} = 1 - \ell_{i,t},$$

and similarly to the label efficient prediction here is used the estimated gain:

$$\widetilde{g}_{i,t} = \begin{cases} \frac{g_{i,t}}{p_{i,t}}, \text{ if } I_t = i, \\ 0, \quad \text{otherwise.} \end{cases}$$

It is an unbiased estimation of the true gain, i.e

$$\mathbb{E}\big[\widetilde{g}_{i,t}\big|I_1^{t-1}\big] = g_{i,t}.$$

The second modification is that instead of an unbiased estimate, a slightly larger quantity is used by the strategy:

$$g'_{i,t} = \widetilde{g}_{i,t} + \frac{\beta}{p_{i,t}}.$$

The third change is a parameter γ which is taken into account in the exploration phase. With the γ it is guaranteed a lower bound for $p_{i,t}$ for all $i = 1, \ldots, N$:

$$p_{i,t+1} = (1-\gamma)\frac{w_{i,t}}{\sum_{j=1}^{N} w_{j,t}} + \frac{\gamma}{N}, \qquad i = 1, \ldots, N.$$

Instead of the pure probability distribution via weighted average, the forecaster uses a mixture of the weighted average and the uniform distribution. The forecaster strategy is defined as a follows:

Theorem 3. *(Auer et al. [11]) For any $0 < \delta < 1$ and for any $n \geq 4N \ln(N/\delta)$, if the forecaster for the multi-armed bandit problem (Algorithm 3) is run with parameters*

$$0 \leq \eta \leq \frac{\gamma}{2N} \text{ and } \sqrt{\frac{\ln(N/\delta)}{nN}} \leq \beta \leq 1$$

Algorithm 3 *(Multi-Armed Bandit) Fix* $\eta > 0$, $0 < \beta < 1$, $0 < \gamma < 1$.
Initialization*:* $w_{i,0} = 1$ *and* $p_{i,1} = 1/N$ *for* $i = 1, \ldots, N$.
For each round $t = 1, 2, \ldots$

(1) Select an action $I_t \in \{1, \ldots, N\}$ *according to the probability distribution* $\mathbf{p}_t = (p_{1,t}, \ldots, p_{N,t})$*;*
(2) Calculate the estimated gains

$$g'_{i,t} = \widetilde{g}_{i,t} + \frac{\beta}{p_{i,t}} = \begin{cases} \frac{g_{i,t}+\beta}{p_{i,t}}, & \text{if } I_t = i, \\ \frac{\beta}{p_{i,t}}, & \text{otherwise;} \end{cases}$$

(3) Update the weights $w_{i,t} = w_{i,t-1} e^{\eta g'_{i,t}}$*;*
(4) Calculate the updated probability distribution

$$p_{i,t+1} = (1-\gamma)\frac{w_{i,t}}{\sum_{j=1}^{N} w_{j,t}} + \frac{\gamma}{N}, \qquad i = 1, \ldots, N.$$

Fig. 3. Exponentially weighted average forecaster for multi-armed bandit problem

then, with probability at least $1 - \delta$,

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq n\left(\gamma + \eta(1+\beta)N\right) + \frac{\ln N}{\eta} + 2\beta n N.$$

In particular, choosing $\beta = \sqrt{\ln(N/\delta)/(nN)}$, $\gamma = \beta N$ *and* $\eta = \gamma/(2N)$,

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq 5\sqrt{nN\ln(N/\delta)}.$$

3 The Combination of the Label Efficient and the Multi-Armed Bandit Problem

In this section we introduce the combination on label efficient and the multi-armed bandit problem. Recall that S_t is a i.i.d. Bernoulli random variable such that $\mathbb{P}\{S_t = 1\} = \epsilon$. In that case the forecaster, after choosing an expert, learns the own loss $\ell_{I_t,t}$ if and only if it queries the "label", i.e., $S_t = 1$.

The modified gain function of the algorithm is:

$$\widetilde{g}_{i,t} = \begin{cases} \frac{g_{i,t}}{p_{i,t}\epsilon}, \text{ if } I_t = i \text{ and } S_t = 1, \\ 0, \quad \text{otherwise.} \end{cases}$$

which is also an unbiased estimation of the observed loss $g_{i,t}$

$$\mathbb{E}\left[\widetilde{g}_{i,t} | S_1^{t-1}, I_1^{t-1}\right] = g_{i,t}.$$

The biased version of the gain is

$$g'_{i,t} = \widetilde{g}_{i,t} + S_t \frac{\beta}{p_{i,t}\epsilon}.$$

Algorithm 4 *Fix $\eta > 0$, $0 < \beta < 1$, $0 < \gamma < 1$ and $0 < \epsilon \leq 1$.*
Initialization*: $w_{i,0} = 1$ and $p_{i,1} = 1/N$ for $i = 1, \ldots, N$.*
For each round $t = 1, 2, \ldots$

(1) Select an action $I_t \in \{1, \ldots, N\}$ according to the probability distribution $\mathbf{p}_t = (p_{1,t}, \ldots, p_{N,t})$;
(2) Draw a Bernoulli random variable S_t such that $\mathbb{P}\{S_t = 1\} = \epsilon$;
(3) If $S_t = 1$ then obtain $g_{I_t,t}$ and compute the estimated gains ($g'_{i,t}$)

$$g'_{i,t} = \widetilde{g}_{i,t} + S_t \frac{\beta}{p_{i,t}\epsilon} = \begin{cases} \frac{g_{i,t}+\beta}{p_{i,t}\epsilon}, & \text{if } I_t = i,\ S_t = 1, \\ \frac{\beta}{p_{i,t}\epsilon}, & \text{if } I_t \neq i,\ S_t = 1, \\ 0 & \text{otherwise;} \end{cases}$$

(4) Update the weights $w_{i,t} = w_{i,t-1} e^{\eta g'_{i,t}}$;
(5) Calculate the updated probability distribution

$$p_{i,t+1} = (1-\gamma)\frac{w_{i,t}}{\sum_{j=1}^{N} w_{j,t}} + \frac{\gamma}{N}, \qquad i = 1, \ldots, N.$$

Fig. 4. Combination of the label efficient and the multi-armed bandit exponentially weighted average forecaster

The next theorem is a joint extension of Corollary 1 and Theorem 3.

Theorem 4. *For any $0 < \delta < 1$, $0 < \epsilon \leq 1$ and for any $n \geq \frac{4N}{\epsilon}\ln(2N/\delta)$ and parameters*

$$\sqrt{\frac{\ln(2N/\delta)}{nN\epsilon}} \leq \beta \leq \frac{1}{2N}\ , \quad \beta N \leq \gamma \leq \frac{1}{2} \quad \text{and} \quad 0 < \eta \leq \frac{\gamma\epsilon}{2N},$$

the performance of the Algorithm 4 can be bounded with probability at least $1-\delta$ as

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq n\left(\gamma + \frac{\eta(1+\beta)N}{\epsilon}\right) + 5\beta nN + \frac{\ln N}{\eta} + 3\sqrt{\frac{n\ln(2/\delta)}{2\epsilon}} + \frac{\ln(2/\delta)}{\epsilon}.$$

In particular, choosing $\beta = \sqrt{\frac{ln(2N/\delta)}{nN\epsilon}}$, $\gamma = \beta N$ and $\eta = \frac{\gamma\epsilon}{2N}$,

$$\widehat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq 9\sqrt{\frac{nN\ln(2N/\delta)}{\epsilon}} + 3\sqrt{\frac{n\ln(2/\delta)}{2\epsilon}} + \frac{\ln(2/\delta)}{\epsilon}.$$

Introduce the notations

$$G'_{i,n} = \sum_{t=1}^{n} g'_{i,t}, \qquad G_{i,n} = \sum_{t=1}^{n} g_{i,t} \qquad \text{and} \qquad \widehat{G}_n = \sum_{t=1}^{n} g_{I_t,t}.$$

The proof of the theorem depends on following lemma, which can be proved exactly same way as in Auer *et al.* [11].

Lemma 1. *For any* $0 < \delta < 1$, $0 < \epsilon \leq 1$ *and* $\sqrt{\frac{\ln(2N/\delta)}{\epsilon n N}} \leq \beta \leq 1$ *we have*

$$\mathbb{P}\left\{G_{i,n} > G'_{i,n} + 4\beta n N\right\} \leq \frac{\delta}{2N}, \qquad i \in \{1, \ldots, N\}.$$

Proof. For any $u > 0$ and $c > 0$ the Chernoff bounding technique (see, e.g., [12]) implies

$$\mathbb{P}\left\{G_{i,n} > G'_{i,n} + u\right\} \leq e^{-cu}\mathbb{E}e^{c(G_{i,n} - G'_{i,n})}. \tag{1}$$

Letting $u = 4\beta n N$ and $c = \beta\epsilon/4$, therefore from (1):

$$\begin{aligned} e^{-cu}\mathbb{E}e^{c(G_{i,n} - G'_{i,n})} = e^{\beta^2 n N \epsilon}\mathbb{E}e^{c(G_{i,n} - G'_{i,n})} &\leq e^{-\ln(2N/\delta)}\mathbb{E}e^{c(G_{i,n} - G'_{i,n})} \\ &= \frac{\delta}{2N}\mathbb{E}e^{c(G_{i,n} - G'_{i,n})}, \end{aligned}$$

where the inequality comes from $\sqrt{\frac{\ln(2N/\delta)}{\epsilon n N}} \leq \beta$. Thus it suffices to prove that

$$\mathbb{E}e^{c(G_{i,n} - G'_{i,n})} \leq 1.$$

For $t = 1, \ldots, n$, introducing, a random variable

$$Z_t = e^{c(G_{i,t} - G'_{i,t})},$$

we clearly have

$$Z_t = e^{c\left(g_{i,t} - g'_{i,t}\right)} Z_{t-1}.$$

Next, for $t = 2, \ldots, n$, we bound $\mathbb{E}[Z_t | I_1, S_1, \ldots, I_{t-1}, S_{t-1}] = \mathbb{E}\left[Z_t | I_1^{t-1}, S_1^{t-1}\right]$ as follows. Note that $c(g_{i,t} - g'_{i,t}) < 1$ because

$$c\left(g_{i,t} - \widetilde{g}_{i,t} - S_t\frac{\beta}{p_{i,t}\epsilon}\right) \leq c\left(1 - \widetilde{g}_{i,t} - S_t\frac{\beta}{p_{i,t}}\right) < 1 - c\widetilde{g}_{i,t} - cS_t\frac{\beta}{p_{i,t}} \leq 1,$$

where the second inequality comes from $c = \beta\epsilon/4 \leq \epsilon/4 < 1$. Moreover, for $x \leq 1$

$$e^x \leq 1 + x + x^2, \tag{2}$$

therefore

$$\begin{aligned} &\mathbb{E}[Z_t | I_1, S_1, \ldots, I_{t-1}, S_{t-1}] \\ &\quad = Z_{t-1}\mathbb{E}\left[e^{c\left(g_{i,t} - \widetilde{g}_{i,t} - S_t\frac{\beta}{p_{i,t}\epsilon}\right)} \middle| I_1^{t-1}, S_1^{t-1}\right] \\ &\quad \leq Z_{t-1}\mathbb{E}\left[1 + c\left(g_{i,t} - \widetilde{g}_{i,t} - S_t\frac{\beta}{p_{i,t}\epsilon}\right) + c^2\left(g_{i,t} - \widetilde{g}_{i,t} - S_t\frac{\beta}{p_{i,t}\epsilon}\right)^2 \middle| I_1^{t-1}, S_1^{t-1}\right]. \end{aligned} \tag{3}$$

Since

$$\mathbb{E}\big[g_{i,t} - \widetilde{g}_{i,t} | I_1^{t-1}, S_1^{t-1}\big] = 0$$

and

$$\mathbb{E}\big[(g_{i,t} - \widetilde{g}_{i,t})^2 | I_1^{t-1}, S_1^{t-1}\big] \leq \mathbb{E}\big[\widetilde{g}_{i,t}^2 | I_1^{t-1}, S_1^{t-1}\big] \leq \frac{1}{p_{i,t}\epsilon},$$

we get from (3) that

$$\begin{aligned}
&\mathbb{E}[Z_t | I_1, S_1, \ldots, I_{t-1}, S_{t-1}] \\
&\leq Z_{t-1}\mathbb{E}\left[1 - \frac{c\beta}{p_{i,t}} + c^2\left(g_{i,t} - \widetilde{g}_{i,t}\right)^2 + c^2 S_t \frac{\beta}{p_{i,t}\epsilon}\left(2\widetilde{g}_{i,t} - 2g_{i,t} + \frac{\beta}{p_{i,t}\epsilon}\right) \middle| I_1^{t-1}, S_1^{t-1}\right] \\
&\leq Z_{t-1}\mathbb{E}\left[1 - \frac{c\beta}{p_{i,t}} + \frac{c^2}{p_{i,t}\epsilon} + c^2 S_t \frac{\beta}{p_{i,t}\epsilon}\left(2\widetilde{g}_{i,t} - 2g_{i,t} + \frac{\beta}{p_{i,t}\epsilon}\right) \middle| I_1^{t-1}, S_1^{t-1}\right] \\
&\leq Z_{t-1}\left[1 + \frac{c}{p_{i,t}}\left(-\beta + \frac{c}{\epsilon} + c\beta\left(\frac{2}{\epsilon} - 2 + \frac{\beta}{p_{i,t}\epsilon}\right)\right)\right], \qquad (4)
\end{aligned}$$

where the last step we used that

$$\mathbb{E}\left[\frac{S_t}{\epsilon}(\widetilde{g}_{i,t} - g_{i,t}) \middle| I_1^{t-1}, S_1^{t-1}\right] = \mathbb{E}\left[g_{i,t}\left(\frac{\mathbb{I}_{\{I_t = i\}} S_t}{p_{i,t}\epsilon^2} - \frac{S_t}{\epsilon}\right) \middle| I_1^{t-1}, S_1^{t-1}\right] \leq \frac{1}{\epsilon} - 1.$$

Since $c = \beta\epsilon/4$ we obtain from (4):

$$\begin{aligned}
-\beta + \frac{c}{\epsilon} + c\beta\left(\frac{2}{\epsilon} - 2 + \frac{\beta}{p_{i,t}\epsilon}\right) &= -\beta + \frac{\beta}{4} + \frac{\beta^2\epsilon}{4}\left(\frac{2}{\epsilon} - 2 + \frac{\beta}{p_{i,t}\epsilon}\right) \\
&\leq -\frac{3\beta}{4} + \frac{\beta^2}{2} + \frac{\beta^3}{4p_{i,t}} \\
&\leq -\frac{\beta}{4} + \frac{\beta^3}{4p_{i,t}} \\
&\leq -\frac{\beta}{4} + \frac{\beta^3 N}{4\gamma},
\end{aligned}$$

where the last inequality comes from $\gamma/N \leq p_{i,t}$. $\gamma \geq \beta N \geq \beta^2 N$, therefore

$$-\beta + \frac{c}{\epsilon} + c\beta\left(\frac{2}{\epsilon} - 2 + \frac{\beta}{p_{i,t}\epsilon}\right) \leq -\frac{\beta}{4} + \frac{\beta}{4} = 0. \qquad (5)$$

Combining (4) and (5) we get that

$$\mathbb{E}[Z_t | I_1, S_1, \ldots, I_{t-1}, S_{t-1}] \leq Z_{t-1}.$$

Then taking expectations on both sides of the inequality we get $\mathbb{E}[Z_t] \leq \mathbb{E}[Z_{t-1}]$ and since $\mathbb{E}[Z_1] \leq 1$, we obtain $\mathbb{E}[Z_n] \leq 1$.

□

Proof of Theorem 4. For the proof of theorem the quantity of $\ln \frac{W_n}{W_0}$ is bounded, where

$$W_t = \sum_{i=1}^{N} w_{i,t}, \qquad t \geq 1$$

and

$$W_0 = N.$$

The lower bound is

$$\begin{aligned} \ln \frac{W_n}{W_0} &= \ln \left(\sum_{i=1}^{N} e^{\eta G'_{i,n}} \right) - \ln N \\ &\geq \ln \left(\max_{i=1,\ldots,N} e^{\eta G'_{i,n}} \right) - \ln N \\ &= \eta \max_{i=1,\ldots,N} G'_{i,n} - \ln N. \end{aligned} \tag{6}$$

For the upper bound note first that the conditions $\beta \leq 1$ and $\eta \leq \frac{\gamma\epsilon}{2N}$ imply that $\eta g'_{i,t} \leq 1$ for all i and t, therefore

$$\begin{aligned} \ln \frac{W_t}{W_{t-1}} &= \ln \sum_{i=1}^{N} \frac{w_{i,t-1}}{\sum_{j=1}^{N} w_{j,t-1}} e^{\eta g'_{i,t}} \\ &= \ln \sum_{i=1}^{N} \frac{p_{i,t} - \gamma/N}{1-\gamma} e^{\eta g'_{i,t}} \\ &\leq \ln \sum_{i=1}^{N} \frac{p_{i,t} - \gamma/N}{1-\gamma} (1 + \eta g'_{i,t} + \eta^2 g'^2_{i,t}) \qquad (7) \\ &\leq \ln \left(1 + \frac{\eta}{1-\gamma} \sum_{i=1}^{N} p_{i,t} g'_{i,t} + \frac{\eta^2}{1-\gamma} \sum_{i=1}^{N} p_{i,t} g'^2_{i,t} \right) \qquad (8) \\ &\leq \frac{\eta}{1-\gamma} \sum_{i=1}^{N} p_{i,t} g'_{i,t} + \frac{\eta^2}{1-\gamma} \sum_{i=1}^{N} p_{i,t} g'^2_{i,t} \qquad (9) \end{aligned}$$

where (7) holds because of (2), (8) follows from the definition of $p_{i,t}$ and (9) holds by the inequality $\ln(1+x) \leq x$ for all $x > -1$.

Next we bound the sums in (9). On the one hand,

$$\sum_{i=1}^{N} p_{i,t} g'_{i,t} = \sum_{i=1}^{N} p_{i,t} \left(\mathbb{I}_{\{I_t = i\}} S_t \frac{g_{i,t}}{p_{i,t}\epsilon} + S_t \frac{\beta}{p_{i,t}\epsilon} \right) = \frac{S_t}{\epsilon} \left(g_{I_t,t} + N\beta \right).$$

On the other hand,

$$\begin{aligned}\sum_{i=1}^{N} p_{i,t} g_{i,t}'^2 &= \sum_{i=1}^{N} p_{i,t} g_{i,t}' \left(\mathbb{I}_{\{I_t=i\}} S_t \frac{g_{i,t}}{p_{i,t}\epsilon} + S_t \frac{\beta}{p_{i,t}\epsilon} \right) \\ &= \frac{g_{I_t,t}' g_{I_t,t}}{\epsilon} S_t + \sum_{i=1}^{N} S_t \frac{\beta}{\epsilon} g_{i,t}' \\ &\leq \frac{S_t}{\epsilon}(1+\beta) \sum_{i=1}^{N} g_{i,t}'.\end{aligned}$$

Therefore, we get that

$$\ln \frac{W_t}{W_{t-1}} \leq \frac{S_t}{\epsilon} \left(\frac{\eta}{1-\gamma} \left(g_{I_t,t} + N\beta \right) + \frac{\eta^2(1+\beta)}{1-\gamma} \sum_{i=1}^{N} g_{i,t}' \right).$$

Summing over $t = 1, \ldots, n$, we have that

$$\begin{aligned}\ln \frac{W_n}{W_0} &\leq \frac{\eta}{1-\gamma} \sum_{t=1}^{n} \frac{S_t}{\epsilon} \left(g_{I_t,t} + N\beta \right) + \frac{\eta^2(1+\beta)}{(1-\gamma)\epsilon} \sum_{i=1}^{N} \sum_{t=1}^{n} S_t g_{i,t}' \\ &\leq \frac{\eta}{1-\gamma} \sum_{t=1}^{n} \frac{S_t}{\epsilon} \left(g_{I_t,t} + N\beta \right) + \frac{\eta^2(1+\beta)}{(1-\gamma)\epsilon} \sum_{i=1}^{N} G_{i,t}' \\ &\leq \frac{\eta}{1-\gamma} \sum_{t=1}^{n} \frac{S_t}{\epsilon} \left(g_{I_t,t} + N\beta \right) + \frac{\eta^2(1+\beta)}{(1-\gamma)\epsilon} N \max_{i=1,\ldots,N} G_{i,n}'.\end{aligned}$$

Combining the upper and the lower bounds for $\ln \frac{W_n}{W_0}$ and rearranging we get

$$\begin{aligned}\sum_{t=1}^{n} \frac{S_t}{\epsilon} \left(g_{I_t,t} + N\beta \right) &\geq \left(1 - \gamma - \frac{\eta(1+\beta)N}{\epsilon} \right) \max_{i=1,\ldots,N} G_{i,n}' - (1-\gamma) \frac{\ln N}{\eta} \\ &\geq \left(1 - \gamma - \frac{\eta(1+\beta)N}{\epsilon} \right) \max_{i=1,\ldots,N} G_{i,n}' - \frac{\ln N}{\eta}. \qquad (10)\end{aligned}$$

Introduce the notation

$$X_t = \frac{S_t}{\epsilon} \left(g_{I_t,t} + N\beta \right) - \left(g_{I_t,t} + N\beta \right),$$

$t = 1, \ldots, n$. $\{X_t\}$ is a martingale difference sequence with respect to I_1^{t-1} and S_1^{t-1}, i.e.

$$\mathbb{E}\left[X_t | I_1^{t-1}, S_1^{t-1} \right] = 0.$$

Now bound, for all $t = 1, \ldots, n$

$$\begin{aligned}
\mathbb{E}\big[X_t^2 | I_1^{t-1}, S_1^{t-1}\big] &= \mathbb{E}\left[\frac{S_t}{\epsilon^2}(g_{I_t,t} + N\beta)^2 + (g_{I_t,t} + N\beta)^2 - 2\frac{S_t}{\epsilon}(g_{I_t,t} + N\beta)^2 \middle| I_1^{s-1}, S_1^{s-1}\right] \\
&\leq \mathbb{E}\left[\frac{S_t}{\epsilon^2}(g_{I_t,t} + N\beta)^2 \middle| I_1^{s-1}, S_1^{s-1}\right] \\
&\leq \frac{(1+N\beta)^2}{\epsilon} \\
&\leq \frac{9}{4\epsilon} \stackrel{\text{def}}{=} \sigma^2 \qquad (11)
\end{aligned}$$

where (11) holds by the assumption of the theorem that $\beta N \leq \gamma \leq \frac{1}{2}$.

We know that

$$X_t \in \left[-\frac{3}{2}, \left(\frac{1}{\epsilon} - 1\right)\frac{3}{2}\right]$$

for all t. Now apply the Bernstein's inequality for martingale differences (Lemma 2)

$$\mathbb{P}\left\{\sum_{t=1}^{n} X_t > u\right\} \leq \frac{\delta}{2}, \qquad (12)$$

where

$$u = \sqrt{2n\frac{9}{4\epsilon}\ln\left(2\delta^{-1}\right)} + \frac{1}{\epsilon}\ln\left(2\delta^{-1}\right).$$

We get from (12)

$$\mathbb{P}\left\{\sum_{t=1}^{n}\frac{S_t}{\epsilon}\left(g_{I_t,t} + N\beta\right) \leq \widehat{G}_n + \beta nN + u\right\} \geq 1 - \frac{\delta}{2}. \qquad (13)$$

By Lemma 1 and the union bound, we obtain

$$\mathbb{P}\left\{\max_{i=1,\ldots,N} G'_{i,n} \geq \max_{i=1,\ldots,N} G_{i,n} - 4\beta nN\right\} \geq 1 - \frac{\delta}{2}. \qquad (14)$$

If A and B are events then by the union bound,

$$\mathbb{P}\left\{A \cap B\right\} = 1 - \mathbb{P}\left\{A^c \cup B^c\right\} \geq 1 - \left(\mathbb{P}\left\{A^c\right\} + \mathbb{P}\left\{B^c\right\}\right),$$

therefore from (13) and (14) combining (10) at least $1 - \delta$

$$\begin{aligned}
\widehat{G}_n \geq &\left(1 - \gamma - \frac{\eta(1+\beta)N}{\epsilon}\right)\max_{i=1,\ldots,N} G_{i,n} - \left(1 - \gamma - \frac{\eta(1+\beta)N}{\epsilon}\right)4\beta nN, \\
&- \frac{\ln N}{\eta} - \beta nN - u.
\end{aligned}$$

because of the coefficient of the $G_{i,n}$ is greater than zero, i.e.,

$$\begin{aligned} 1-\gamma-\frac{\eta(1+\beta)N}{\epsilon} &\geq 1-\gamma-\frac{\gamma\epsilon}{2N}\frac{(1+\beta)N}{\epsilon} \\ &\geq 1-2\gamma \\ &\geq 0 \end{aligned}$$

by the assumption of the theorem.

Since $\widehat{L}_n = n - \widehat{G}_n$ and $L_{i,n} = n - G_{i,n}$, we have

$$\begin{aligned} \widehat{L}_n &\leq \left(1-\gamma-\frac{\eta(1+\beta)N}{\epsilon}\right)\min_{i=1,\ldots,N} L_{i,n} + n\left(\gamma+\frac{\eta(1+\beta)N}{\epsilon}\right) \\ &\quad + \left(1-\gamma-\frac{\eta(1+\beta)N}{\epsilon}\right)4\beta nN + \beta nN + \frac{\ln N}{\eta} + u \\ &\leq \min_{i=1,\ldots,N} L_{i,n} + n\left(\gamma+\frac{\eta(1+\beta)N}{\epsilon}\right) + 5\beta nN + \frac{\ln N}{\eta} + u. \end{aligned}$$

□

4 Conclusion

In this paper, we prove worst-case loss bounds for online learning for forecasting in the extension of the label efficient and multi-armed bandit problems.

Type of the Algorithm	Amount of the information	Upper bound
Full Information	nN	$O\left(\sqrt{n \ln N}\right)$
Label Efficient	$nN\epsilon$	$O\left(\sqrt{n \ln N/\epsilon}\right)$
Multi-Armed Bandit	n	$O\left(\sqrt{nN \ln N}\right)$
Label Efficient and Multi-Armed Bandit	$n\epsilon$	$O\left(\sqrt{nN \ln N/\epsilon}\right)$

Table 1. The connection between the upper bounds of the different algorithms and the amount of information.

The Table 1 implies simple connection between the amount of information and the upper bound of the algorithms.

5 Appendix

We recall the Bernstein inequality for martingale differences (Berstein [13]).

Lemma 2. *Let $X_1, \ldots, X_n$ be a martingale differences such that $X_t \in [a, b]$ with probability one ($t = 1, \ldots, n$). Assume that, for all t,*

$$\mathbb{E}\big[X_t^2 | X_{t-1}, \ldots, X_1\big] \leq \sigma^2 \ a.s.$$

Then, for all $\epsilon > 0$,

$$\mathbb{P}\left\{\sum_{t=1}^{n} X_t > \epsilon\right\} \leq e^{\frac{-\epsilon^2}{2n\sigma^2 + 2\epsilon(b-a)/3}}$$

and therefore

$$\mathbb{P}\left\{\sum_{t=1}^{n} X_t > \sqrt{2n\sigma^2 \ln \delta^{-1}} + 2 \ln \delta^{-1}(b-a)/3\right\} \leq \delta.$$

Acknowledgements

The authors would like to thank László Györfi for useful comments.

References

1. D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.
2. J. Hannan. Approximation to bayes risk in repeated plays. In M. Dresher, A. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games*, volume 3, pages 97–139. Princeton University Press, 1957.
3. V. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 372–383, Rochester, NY, Aug. 1990. Morgan Kaufmann.
4. N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.
5. N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, D. Haussler, R. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.
6. N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, to appear, Cambridge, 2006.
7. E. Gelenbe, M. Gellman, R. Lent, P. Liu, and P. Su. Autonomous smart routing for network QoS. In *Proceedings of First International Conference on Autonomic Computing*, pages 232–239, New York, May 2004. IEEE Computer Society.
8. E. Gelenbe, R. Lent, and Z. Xhu. Measurement and performance of a cognitive packet network. *Journal of Computer Networks*, 37:691–701, 2001.
9. N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE Trans. Inform. Theory*, IT-51:2152–2162, June 2005.

10. H. Robbins. Some aspects of the sequential design of experiments. *Bullettin of the American Mathematical Society*, 55:527–535, 1952.
11. P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: the adversial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science, FOCS 1995*, pages 322–331, Washington, DC, USA, Oct. 1995. IEEE Computer Society Press, Los Alamitos, CA.
12. L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.
13. S. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.