Online Learning under Delayed Feedback

Pooria Joulani POORIA@UALBERTA.CA
András György GYORGY@UALBERTA.CA
Csaba Szepesvàri SZEPESVA@UALBERTA.CA
Dept. of Computing Science, University of Alberta, Edmonton, AB, T6G 2E8 CANADA

Abstract

Online learning with delayed feedback has received increasing attention recently due to its several applications in distributed, web-based learning problems. In this paper we provide a systematic study of the topic, and analyze how the delay effects the regret of online learning algorithms. Somewhat surprisingly, it turns out that delay increases the regret in a multiplicative way in adversarial problems, and in an additive way in stochastic problems. We give meta-algorithms that transform, in a black-box fashion, algorithms developed for the non-delayed case into ones that can handle the presence of delays in the feedback loop. Modifications of the well-known UCB algorithm are also developed for the bandit problem with delayed feedback, with the advantage over the meta-algorithms that they can be implemented with much lower complexity.

1. Introduction

The problem of sequential learning received much attention in the recent years. While in its traditional formulation any information that is available on the effects of an action of a decision maker are immediately available after the action has been made, in many applications such information may come after some delay. This is the case, for example, in web advertisement, where the information whether a user has clicked on a certain ad may come back to the engine in a delayed fashion: after an ad is selected, while waiting for the information if the user clicks or not, the engine has to provide advertisements to other users. Also, the click information may enter to a very different module of the system than the one that decides about the ads, and propagating the information through the system also causes delays (Li et al., 2010; Dudik et al., 2011). Another example is parallel, distributed learning, where information about data processed at a node comes after large delay and aggregation to other nodes (Agarwal & Duchi, 2011).

While online learning has been proven successful in many different machine learning problems, and is applied in practice in situations where the feedback is delayed, the theoretical results for the non-delayed setup are not applicable when delays are present. While special algorithms and minimax analyses have been developed for different special online learning problems and delay models (mostly with constant delays), a comprehensive understanding of the effects of delays are missing. In this paper we provide a systematic study of online learning problems with delayed feedback. We consider the partial monitoring setting, which covers all settings previously considered in the literature, extending, unifying, and often improving upon existing results. In particular, we give general meta-algorithms that transform, in a black-box fashion, algorithms developed for the non-delayed case into algorithms that can handle delays efficiently. We analyze how the delay effects the regret of the algorithms. One interesting, perhaps somewhat surprising, result is that the delay inflates the regret in a multiplicative way in adversarial problems, while this effect is only additive in stochastic problems. While our general meta-algorithms are very useful, their time- and space-complexity may be unnecessarily large in certain practical situation. To resolve this problem, we work out modifications of variants of the UCB algorithm (Auer et al., 2002) for stochastic bandit problems with delayed feedback that have much smaller complexity than the black-box algorithms.

The rest of the paper is organized as follows. The problem of online learning with delayed feedback is defined in Section 2. The adversarial and stochastic problems

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

Parameters: Forecaster's action set $\mathcal{A}$, set of outcomes $\mathcal{B}$, side information set $\mathcal{X}$, reward function $r : \mathcal{X} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$, feedback function $h : \mathcal{X} \times \mathcal{A} \times \mathcal{B} \to \mathcal{H}$, time horizon n (optional).
At each time instant $t = 1, 2, \ldots, n$:

1. The environment chooses some side information $x_t \in \mathcal{X}$ and an outcome $b_t \in \mathcal{B}$.
2. The side information x_t is presented to the forecaster, who selects an action $a_t \in \mathcal{A}$, which results in the reward $r(x_t, a_t, b_t)$ (unknown to the forecaster).
3. The feedback $h_t = h(x_t, a_t, b_t)$ is scheduled to be revealed after τ_t time instants.
4. The agent observes all the feedback scheduled to be revealed at time step t, as well as their corresponding time steps, that is, it receives the set $H_t = \{(t', h_{t'}) : t' \le t, t' + \tau_{t'} = t\}$.

Figure 1: Online learning under delayed feedback.

are analyzed in Sections 3.1 and 3.2, while the modification of the UCB algorithm is given in Section 4.

2. Online learning with delayed feedback

We consider the following general model of online learning, the partial monitoring problem with side information: the forecaster (decision maker) has to make a sequence of predictions (actions), possibly based on some side information, and for each prediction it receives some reward and feedback. More formally, given a set of possible side information values $\mathcal{X}$, an action set $\mathcal{A}$, a set of reward functions $\mathcal{R} \subset \{r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}$, and a set of possible feedback values $\mathcal{H}$, at each time instant $t = 1, 2, \ldots$, the forecaster receives some side information $x_t \in \mathcal{X}$; then simultaneously, possibly based on the side information, the forecaster predicts some value $a_t \in \mathcal{A}$ while the environment chooses a reward function $r_t \in \mathcal{R}$; finally, the forecaster receives reward $r_t(x_t, a_t)$ and some feedback $h_t \in \mathcal{H}_t$.

Note that the forecaster might not have any information about the rewards it receives (i.e., the rewards may be hidden), the only information it receives comes from the feedback. In standard online learning models the feedback h_t is immediately available to the forecaster; in the delayed model we consider the feedback h_t corresponding to time instant t is received only after a delay of τ_t (τ_t is a natural number or ∞), that is, after the prediction in time instant $t + \tau_t$.

The goal of the forecaster is to maximize its cumulative reward $\sum_{t=1}^{n} r_t(x_t, a_t)$ in n time instants. More precisely we wish to minimize the loss in rewards relative to the best static strategy selected in hindsight, called the *regret*, defined as

$$R_n = \sup_{a:\mathcal{X}\to\mathcal{A}} \sum_{t=1}^{n} r_t(x_t, a(x_t)) - \sum_{t=1}^{n} t_t(a_t).$$

A forecaster is consistent if it achieves, asymptotically, the average reward of the best static strategy, that is $R_n/n \to 0$, and we are interested in how fast the average regret can be made to converge to 0.

The above general problem formulation includes most scenarios considered in online learning. In the full information case the feedback is the reward function itself, that is, $h_t = r_t$. In the bandit case the forecaster only learns the rewards of its own prediction, that is, $h_t(x_t, a_t)$. In the partial monitoring case there is a known reward function $r : \mathcal{X} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ and feedback function $h : \mathcal{X} \times \mathcal{A} \times \mathcal{B} \to \mathcal{H}$, where $\mathcal{B}$ is a set of choices (outcomes) of the environment, and for each time instant the environment picks an outcome $b_t \in \mathcal{B}$, and the reward and feedback functions become $r_t(x_t, a_t) = r(x_t, a_t, b_t)$, $h_t(x_t, a_t) = h(x_t, a_t, b_t)$, respectively. This interaction protocol is shown in Figure 1. Note that the bandit and full information problems can also be treated as special partial monitoring problems. Therefore, we will use this last formulation of the problem. When no stochastic assumption is made on how the sequence b_t is generated, we talk about the adversarial model. In the stochastic setting we will consider the case when b_t is a sequence of independent, identically distributed (i.i.d.) random variables. Side information may or may not be present in a real problem; in its absence $\mathcal{X}$ is a singleton set.

Finally, we may have different assumptions on how the delays τ_t are determined. Most often we will assume that τ_t is an i.i.d. sequence, and τ_t is also independent of the choice a_t. A more general assumption is that the delays can depend on the actions, i.e., that there is a set of random variables $\tau_{t,a}, t \in \mathbb{N}, a \in \mathcal{A}$, and $\tau_t = \tau_{t,a_t}$.

The delays may cause the feedback from two different time steps to be revealed at the same time; hence the forecaster may observe more than one piece of feedback information in a time instant, or not observe any feedback in others. Also, the delay may change the order of observing the feedbacks, with the feedback of a more recent action being observed before the feedback of an earlier one.

2.1. Related work

The effect of delayed feedback has been studied in the recent years under different online learning scenarios and different assumptions on the delay. A concise summary, together with the contributions of this paper, is given in Table 1.

To the best of our knowledge, Weinberger & Ordentlich (2002) were the first to analyze the delayed feedback problem; they considered the adversarial full information setting with a fixed and known delay τ_{const}. They showed that the minimax optimal solution is to run $\tau_{const}+1$ independent optimal predictors on the subsampled reward sequences: that is, $\tau_{const}+1$ prediction strategies are used such that i^{th} predictor is used at time instants t with $t \mod (\tau_{const}+1)+1 = i$. This approach forms the basis of our method devised for the adversarial case (see Section 3.1). Langford et al. (2009) showed that under the usual conditions, a sufficiently slowed-down version of the mirror descent algorithm achieves optimal decay rate of the average regret. Mesterharm (2005; 2007) considered another variant of the full information setting, using an adversarial model on the delays in the label prediction setting, where the forecaster has to predict the label corresponding to a side information vector x_t. While in the full information online prediction problem Weinberger & Ordentlich (2002) showed that the regret increases by a multiplicative factor of τ_{const}, in the work of Mesterharm (2005; 2007) the important quantity becomes the maximum/average gap defined as the length of the largest time interval the forecaster does not receive feedback. Mesterharm (2005; 2007) also shows that the minimax regret in the adversarial case increases multiplicatively by the average gap, while it increases only in an additive fashion in the stochastic case, by the maximum gap. Agarwal & Duchi (2011) also considered the problem of online stochastic optimization and showed that, for i.i.d. random delays, the regret increases with an additive factor of order $\mathbb{E}\left[\tau^2\right]$.

Qualitatively similar results were obtained in the bandit setting. Considering a fixed and known delay τ_{const}, Dudik et al. (2011) showed an additive $O(\tau_{const} \log n)$ penalty in the regret for the stochastic setting (with side information), while (Neu et al., 2010) showed a multiplicative regret for the adversarial bandit case. The problem of delayed feedback has also been studied for Gaussian process bandit optimization (Desautels et al., 2012), resulting in a multiplicative increase in the regret that is independent of the delay and an additive term depending on the maximum delay.

In the rest of the paper we generalize the above results to the partial monitoring setting, extending, unifying, and often improving existing results.

3. Black-Box Algorithms for Delayed Feedback

In this section we provide black-box algorithms for the delayed feedback problem. We assume that there exists a base algorithm Base solving the prediction problem without delay. We often do not specify the assumptions underlying the regret bounds of these algorithms, and assume that the problem we consider only differs from the original problem because of the delays. For example, in the adversarial setting, Base may build on the assumption that the reward functions are selected in an oblivious or non-oblivious way (i.e., independently or not of the predictions of the forecaster). First we consider the adversarial case in Section 3.1. Then in Section 3.2, we provide tighter bounds for the stochastic case.

3.1. Adversarial setting

The algorithm of Weinberger & Ordentlich (2002) for the adversarial full information setting subsamples the reward sequence by the constant delay $\tau_{const}+1$, and runs a base algorithm Base on each of the τ_{const} subsampled sequences. We say that an algorithm Base *enjoys a regret or expected regret bound* $f : [0,\infty) \to \mathbb{R}$ under the given assumptions in the non-delayed setting if (i) f is nondecreasing, concave, $f(0) = 0$; and (ii) $\sup_{b_1,\dots,b_n \in \mathcal{B}} R_n \le f(n)$ or, respectively, $\sup_{b_1,\dots,b_n \in \mathcal{B}} \mathbb{E}\left[R_n\right] \le f(n)$ for all n. Weinberger & Ordentlich (2002) also showed that if Base enjoys a regret bound f then their algorithm in the fixed delay case enjoys a regret bound $\tau_{const} f(n/\tau_{const})$. Furthermore, when Base is minimax optimal in the non-delayed setting, the subsampling algorithm is also minimax optimal in the (full information) delayed setting, as can be seen by constructing a reward sequence that changes only in every $\tau_{const}+1$ times. Note that Weinberger & Ordentlich (2002) does not need condition (i) of f. However, these conditions are satisfied by all regret bounds we are aware of, and imply that $yf(x/y)$ is a nondecreasing and concave function of y for any fixed x, a fact which will turn out to be useful in the analysis later.

In this section we extend the algorithm of Weinberger & Ordentlich (2002) to the case when the delays are not constant, and to the partial monitoring setting. The idea is that we run several instances of a non-delayed algorithm Base as needed: One step of an algorithm is completed if it receives the feedback cor-

		Stochastic Feedback	General (Adversarial) Feedback
Full Info	No Side Info	$R(n) \leq R'(n) + O(\mathbb{E}\left[\tau_t^2\right])$ (Agarwal & Duchi, 2011)	L $R(n) \leq O(\tau_{const}) \times R'(n/\tau_{const})$ (Weinberger & Ordentlich, 2002) (Langford et al., 2009) (Agarwal & Duchi, 2011)
	Side Info	L $R(n) \leq R'(n) + O(D^*)$ (Mesterharm, 2007)	L $R(n) \leq O(\bar{D}) \times R'(n/\bar{D})$ (Mesterharm, 2007)
Bandit Feedback	No Side Info	$R(n) \leq C_1 R'(n) + C_2 \tau_{\max} \log(\tau_{\max})$ (Desautels et al., 2012)	$R(n) \leq O(\tau_{const}) \times R(n/\tau_{const})$ (Neu et al., 2010)
	Side Info	$R(n) \leq R'(n) + O(\tau_{const}\sqrt{\log n})$ (Dudik et al., 2011)	
Partial Monitoring	No Side Info	$\mathbf{R_n \leq R'(n) + O(G_n^*)}$	$\mathbf{R_n \leq (1 + \mathbb{E}\left[G_n^*\right]) \times R'\left(\frac{n}{1 + \mathbb{E}\left[G_n^*\right]}\right)}$
	Side Info	N/A	$\mathbf{R_n \leq (1 + \mathbb{E}\left[G_n^*\right]) \times R'\left(\frac{n}{1 + \mathbb{E}\left[G_n^*\right]}\right)}$

Table 1. Summary of work on online learning under delayed feedback. $R(n)$ shows the (expected) regret in the delayed setting, while $R'(n)$ shows the (upper bound on) the (expected) regret in the non-delayed setting. L denotes a matching lower bound. D^* and $\bar{D}$ indicate the maximum and average *gap*, respectively, where a gap is a number of consecutive time steps the agent does not get any feedback (in the adversarial delay formulation used by Mesterharm (2005; 2007)). The term τ_{const} indicates that the results are for constant delays only. For the work of (Desautels et al., 2012), C_1 and C_2 are positive constants, with $C_1 > 1$, and $\tau_{\max}$ denotes the maximum delay. Results presented in this paper are boldface, where $\mathbf{G_n^* = \tau_{max}}$ when the delays have an upper bound τ_{max}, and $\mathbf{G_n^* = O\left(\mathbb{E}\left[\tau_t\right]\sqrt{\log n} + \log n\right)}$ when the delays τ_t are i.i.d. The new bounds for the partial monitoring problem are automatically applicable in the other, spacial, cases, and give improved results in most cases.

responding to its previous prediction. When we need to make a prediction, we use one of our algorithm instances that have already completed their previous step. If no such instance exists, we create a new one, and use that one. This results in the following algorithm, which we call Black-Box Online Learning under Delayed Feedback (BOLD) (note that when the delays are constant, BOLD reduces to the algorithm of Weinberger & Ordentlich (2002)):

Algorithm 1 Black-box Online Learning under Delayed-feedback (**BOLD**)

for each time instant $t = 1, 2, \ldots, n$ **do**
  Prediction:
    Pick a free instances of BASE, or create a new instance if none of the available instances are free. Feed this instance with x_t and use its prediction.
  Update:
  for each $(s, h_s) \in H_t$ **do**
    Update the instance used at time instant s with the feedback h_s.
  end for
end for

Clearly, the performance of BOLD depends on how many instances of BASE we need to create, and how many times each instance is used. Let M_t denote the number of BASE instances created by BOLD up to and including time t. That is, $M_1 = 1$, and we create a new instance at the beginning of any time instant when all instances are waiting for their feedback. Let $G_t = \sum_{s=1}^{t-1} \mathbb{I}\left\{s + \tau_s \geq t\right\}$ be the total number of outstanding (missing) feedbacks when the forecaster is making a decision at time instant t. Then we have G_t algorithms waiting for their feedback, and so $M_t \geq G_t + 1$. Since we only introduce new instances when it is necessary (and each time instant at most one new instance is created), it is easy to see that

$$M_t = G_t^* + 1 \tag{1}$$

for any t, where $G_t^* = \max_{1 \leq s \leq t} G_t$.

We can use the result above to transfer the regret guarantee of the non-delayed base algorithm BASE to a guarantee on the regret of BOLD. We will assume that the expected regret of the non-delayed base algorithm is bounded by a concave function f. This assumption holds for almost every online learning contexts (e.g.,

multi-armed bandits, contextual bandits, partial monitoring, etc.), which all have a regret upper bound of the form $\widetilde{O}(n^\alpha)$ for some $0 \le \alpha \le 1$, with, typically, $\alpha = 1/2$ or $2/3$.[1]

Theorem 1. *Suppose that the non-delayed algorithm* BASE *used in BOLD enjoys an (expected) regret bound* f_{BASE}. *Assume, furthermore, that the delays* τ_t *are independent of the forecaster's prediction* a_t. *Then the expected regret of BOLD after* n *time steps satisfies*

$$\begin{aligned}\mathbb{E}[R_n] &\le \mathbb{E}\left[(G_n^*+1) f_{\text{BASE}}\left(\frac{n}{G_n^*+1}\right)\right] \\ &\le (\mathbb{E}[G_n^*]+1) f_{\text{BASE}}\left(\frac{n}{\mathbb{E}[G_n^*]+1}\right). \end{aligned} \quad (2)$$

Proof. We prove the first inequality of the statement for deterministic delays $\tau_1, \dots, \tau_n$. The second inequality follows from the concavity of $y f_{\text{BASE}}(x/y)$ for any x. The extension to random delays follows by first conditioning on the delay sequence; note that here we implicitly use the fact that the delays can be assumed to be the same for each prediction since the forecaster makes one prediction in each time instant, and the delays are assumed to be independent of the chosen prediction.

Since the delays are assumed to be deterministic, the sequence of the BASE algorithm instances generated and the order they are used is determined. Therefore, each BASE instance plays in a non-delayed online learning problem. For any $1 \le j \le M_n$, let L_j denote the list of time instants in which BOLD has used the prediction chosen by instance j, and let $n_j = |L_j|$ be the number of time instants this happens. Furthermore, let $R_{n_j}^j$ denote the regret of instance j in its non-delayed problem, that is,

$$R_{n_j}^j = \sup_{a:\mathcal{X}\to\mathcal{A}} \sum_{t\in L_j} r_t(a(x_t)) - \sum_{t\in L_j} r_t(a_t),$$

where a_t is the action chosen by BOLD (and instance j) at time instant t. Then

$$\begin{aligned} R_n &= \sup_{a:\mathcal{X}\to\mathcal{A}} \sum_{t=1}^n r_t(a(x_t)) - \sum_{t=1}^n r_t(a_t) \\ &= \sup_{a:\mathcal{X}\to\mathcal{A}} \sum_{j=1}^{M_n} \sum_{t\in L_j} r_t(a(x_t)) - \sum_{j=1}^{M_n} \sum_{t\in L_j} r_t(a_t) \\ &\le \sum_{j=1}^{M_n} \left(\sup_{a:\mathcal{X}\to\mathcal{A}} \sum_{t\in L_j} r_t(a(x_t)) - \sum_{t\in L_j} r_t(a(a_t)) \right) \\ &= \sum_{j=1}^{M_n} R_{n_j}^j. \end{aligned} \quad (3)$$

Now using the fact that f_{BASE} is an (expected) regret bound, we obtain

$$\begin{aligned} \mathbb{E}[R_n] &\le \sum_{j=1}^{M_n} \mathbb{E}\left[R_{n_j}^j\right] \le \sum_{j=1}^{M_n} f_{\text{BASE}}(n_j) \\ &= M_n \sum_{j=1}^{M_n} \frac{1}{M_n} f_{\text{BASE}}(n_j) \\ &\le M_n f_{\text{BASE}}\left(\sum_{j=1}^{M_n} \frac{1}{M_n} n_j\right) = M_n f_{\text{BASE}}\left(\frac{n}{M_n}\right), \end{aligned} \quad (4)$$

where the last inequality follows from Jensen's inequality and the concavity of f_{BASE}. Substituting M_n from (1) concludes the proof. □

Now we need to bound G_n^* to make the theorem meaningful. When all delays are the same constants, for $n > \tau_{const}$ we get $G_n^* = \tau_t = \tau_{const}$, and we get back the regret bound

$$\mathbb{E}[R_n] \le (\tau_{const}+1) f_{\text{BASE}}\left(\frac{n}{\tau_{const}+1}\right)$$

of Weinberger & Ordentlich (2002), but in the much more general partial monitoring case. However, we do not know whether this bound is tight when BASE is minimax optimal, as the argument for the lower bound does not work in the partial information setting (the forecaster can gain extra information in each block with the same reward functions).

Assuming the delays are i.i.d., we can give an interesting bound on G_n^*. The result is based on the fact that although the G_t can be as large as t, but both its expectation and variance are upper bounded by $\mathbb{E}[\tau_1]$.

Lemma 2. *Assume* $\tau_1, \dots, \tau_n$ *is a sequence of i.i.d. random variables with finite expected value, and let* $B(n,t) = t + 2\log n + \sqrt{4t\log n}$. *Then*

$$\mathbb{E}[G_n^*] \le B(n, \mathbb{E}[\tau_1]) + 1.$$

[1] $u_n = \widetilde{O}(v_n)$ means that there is a $\beta \ge 0$ such that $\lim_{n\to\infty} u_n/(v_n \log^\beta n) = 0$.

Proof. First consider the expectation and the variance of G_t. For any t,

$$\mathbb{E}[G_t] = \mathbb{E}\left[\sum_{s=1}^{t-1} \mathbb{I}\{s+\tau_s \geq t\}\right] = \sum_{s=1}^{t-1} \mathbb{P}\{s+\tau_s \geq t\} = \sum_{s=0}^{t-2} \mathbb{P}\{\tau_1 > s\} \leq \mathbb{E}[\tau_1]$$

and, similarly

$$\sigma^2[G_t] = \sum_{s=1}^{t-1} \sigma^2[\mathbb{I}\{s+\tau_s \geq t\}] \leq \sum_{s=1}^{t-1} \mathbb{P}\{s+\tau_s \geq t\} \leq \mathbb{E}[\tau_1].$$

By Bernstein's inequality (Cesa-Bianchi & Lugosi, 2006, Corollary A.3), for any $0 < \delta < 1$ and any t we have, with probability at least $1-\delta$,

$$G_t - \mathbb{E}[G_t] \leq \log\tfrac{1}{\delta} + \sqrt{2\sigma^2[G_t]\log\tfrac{1}{\delta}}.$$

Applying the union bound for $\delta = 1/n^2$, and our previous bounds on the variance and expectation of G_t, we obtain that with probability at least $1-1/n$,

$$\max_{1\leq t\leq n} G_t \leq \mathbb{E}[\tau_1] + 2\log n + \sqrt{4\mathbb{E}[\tau_1]\log n}.$$

Taking into account that $\max_{1\leq t\leq n} G_t \leq n$, we get the statement of the lemma. □

Corollary 3. *Under the conditions of Theorem 1, if the sequence of delays is i.i.d, then*

$$\mathbb{E}[R_n] \leq (B(n,\mathbb{E}[\tau_1])+2) f_{\text{BASE}}\left(\frac{n}{B(n,\mathbb{E}[\tau_1])+2}\right).$$

Note that although the delays can be arbitrarily large, whenever the expected value is finite, the bound only increases by a $\log n$ factor.

3.2. Stochastic setting with finite action spaces

In this section we consider the case when the action set $\mathcal{A}$ of the forecaster is finite; without loss of generality we assume $\mathcal{A} = \{1,2,\ldots,K\}$. We also assume that there is no side information (that is, x_t is a constant for all t, and, hence, will be omitted; the results can be extended easily to the case of a finite side information set, where we can repeat the procedures described below for each value of the side information separately). We also assume that the outcomes b_t form an i.i.d. sequence, which is also independent of the choices of the forecaster. When $\mathcal{B}$ is finite, this leads to the standard i.i.d. partial monitoring (IPM) setting, while the conventional multi-armed bandit (MAB) setting is recovered when the feedback is the reward of the last choice, that is, $h_t = r_t(a_t, b_t)$. As in the previous section, we will assume that the feedback delays are independent of the choices of the forecaster.

By the independence assumption on the outcomes, the sequences of potential rewards $r_t(i) = r_t(i, b_t)$ and feedbacks $h_t(i) = h_t(i, b_t)$ are i.i.d., respectively, for the same action $i \in \mathcal{A}$. Let $\mu_i = \mathbb{E}[r_t(i)]$ denote the expected reward of choosing action i, $\mu^* = \max_{i\in\mathcal{A}} \mu_i$ the optimal reward and i^* with $\mu_{i^*} = \mu^*$ the optimal action. Moreover, let $T_i(n) = \sum_{t=1}^{n} \mathbb{I}\{A_t = i\}$ denote the number of times action i is chosen by the end of time instant n. Then, defining the gaps $\Delta_i = \mu^* - \mu_i$ for all $i \in \mathcal{A}$, the expected regret of the forecaster becomes

$$E[R_n] = \sum_{t=1}^{n} \mu^* - \mu_{A_t} = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]. \qquad (5)$$

In this section we study the IPM setting when the feedback is delayed. Under these assumptions, we can show a smaller penalty compared to the adversarial case: we show that the penalty in the regret grows in an additive fashion, as opposed to the multiplicative penalty in the adversarial case. Similarly to the adversarial setting, we build on a base algorithm BASE for the non-delayed case. The advantage in the IPM setting (and that we consider expected regret) is that here BASE can consider a permuted order of rewards and feedbacks, and so we do not have to wait for the actual feedback, but it is enough to receive a feedback for the same action.

This idea is in the core of our algorithm, Queued Partial Monitoring with Delayed Feedback (QPM-D):

Algorithm 2 Queued Partial Monitoring Algorithm with Delays (QPM-D)

```
Create an empty FIFO buffer Q[i] for each action i.
Let I be the first prediction of BASE.
for each time instant t = 1, 2, ..., n do
  Predict:
  while Q[I] is not empty do
    Update BASE with a feedback from Q[I].
    Let I be the next prediction of BASE.
  end while
  There are no buffered feedbacks for action I, so
  predict a_t = I at time instant t to get a feedback.
  Update:
  for each (s, h_s) ∈ H_t do
    Add the feedback h_s to the buffer Q[a_s].
  end for
end for
```

Here we have a BASE partial monitoring algorithm

for the non-delayed case, which is run inside the algorithm. The feedback information coming from the environment is stored in separate queues for each action. The outer algorithm constantly queries BASE: While feedbacks for the chosen actions are available in the queues, only the inner algorithm BASE runs (that is, this happens within a single time instant in the real prediction problem). When no feedback is available, the outer algorithm keeps sending the same action to the real environment until a feedback for the selected action arrives. In this way BASE is run in a simulated non-delayed environment.

The next lemma implies that the inner algorithm BASE actually runs in a non-delayed version of the problem, as it experiences the same distributions:

Lemma 4. *Consider a delayed stochastic IPM problem, and assume that the delays are independent of the outcomes of the environment. For any action i, for any $s \in \mathbb{N}$ let $h'_{i,s}$ denote the s^{th} feedback the algorithm receives for choosing action i. Then the sequence $\{h'_{i,s}\}_{s\in\mathbb{N}}$ is an i.i.d. sequence with the same distribution as the sequence of feedbacks $\{h_{t,i}\}_{t\in\mathbb{N}}$.*

To relate the non-delayed performance of BASE and the regret of QPM-D, we need a few definitions. For any t, let $S_i(t)$ denote the number of feedbacks received by the end of time instant t. Then the number of missing feedbacks for the choice of action i when making a decision at time instant t is $G_{i,t} = T_i(t-1) - S_i(t-1)$, and let $G^*_{i,n} = \max_{1\le t\le n} G_{i,t}$. For each $i \in \mathcal{A}$, let $T'_i(t')$ be the number of times algorithm BASE has chosen action i after being queried t' times.

Let n' denote the number of steps the inner algorithm BASE makes in n steps of the real IPM problem. Next we relate n and n', as well as the number of times QPM-D and BASE (in its simulated environment) choose an action.

Lemma 5. *Suppose QPM-D is run for $n \ge 1$ time instants, and has queried the base algorithm for n' times. Then $n' \le n$ and*

$$0 \le T_i(n) - T'_i(n') \le G^*_{i,n}. \tag{6}$$

Proof. Since it can take at most one step for each feedback that arrives, and the real algorithm has to make at least one step for each arriving feedback, $n' \le n$.

If BASE, and hence, QPM-D, has not chosen an action i by time instant n, (6) trivially holds. Otherwise, let $t_{n,i}$ denote the last time instant (up to time n) when QPM-D chooses action i. Then $T_i(n) = T_i(t_{n,i}) = T_i(t_{n,i}-1)+1$. Suppose BASE has been queried $n'' \le t_{n,i}$ times by, and including, time instant $t_{n,i}$. At this time instant, the buffer $Q[i]$ must be empty and BASE must be picking i, otherwise QPM-D would not choose action i. This means that all the $S_i(t_{n,i}-1)$ feedbacks that have arrived before this time instant have been fed to the base algorithm, which has also made an extra step, that is, $T'_i(n') \ge T'_i(n'') \ge S_i(t_{n,i}-1)+1$, and so

$$\begin{aligned} T_i(n) - T'_i(n') &\le T_i(t_{n,i}-1) + 1 - (S_i(t_{n,i}-1)+1) \\ &\le G_i(t_n) \le G^*_{i,n}. \end{aligned}$$

□

We can now give an upper bound on the expected regret of Algorithm 2

Theorem 6. *Suppose the non-delayed* BASE *algorithm is used in QPM-D in a delayed stochastic IPM environment. Then the expected regret of QPM-D is upper-bounded by*

$$\mathbb{E}[R_n] \le \mathbb{E}\left[R_n^{\text{BASE}}\right] + \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[G^*_{i,n}\right], \tag{7}$$

where $\mathbb{E}\left[R_n^{\text{BASE}}\right]$ is the expected regret of BASE *when run in the same environment without delays.*

When the delay τ_t is bounded by τ_{max} for all t, we also have $G^*_{i,n} \le \tau_{max}$, and $\mathbb{E}[R_n] \le \mathbb{E}\left[R_n^{\text{BASE}}\right] + O(\tau_{max})$. When the sequence of delays is i.i.d. with a finite expected value but unbounded support, we can use Lemma 2 to bound $G^*_{i,n}$ (which is clearly at most the maximum number of all missing rewards), and obtain a bound $\mathbb{E}\left[R_n^{\text{BASE}}\right] + O(\mathbb{E}[\tau_1]\sqrt{\log n} + \log n)$.

Proof. Assume that the QPM-D is run longer so that the base algorithm BASE is queried for n times (i.e., it is queried $n-n'$ more times). Then, since $n' \le n$, the number of times action i is chosen by the base algorithm, namely $T'_i(n)$, can only increase, that is, $T'_i(n') \le T'_i(n)$. Combining this with the expectation of (6) gives

$$\mathbb{E}[T_i(n)] \le \mathbb{E}[T'_i(n)] + \mathbb{E}\left[G^*_{i,n}\right].$$

Multiplying both sides by Δ_i and summing over all the actions, we get

$$\begin{aligned} &\sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)] \\ &\qquad \le \sum_{i=1}^{K} \Delta_i \mathbb{E}[T'_i(n)] + \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[G^*_{i,n}\right]. \end{aligned} \tag{8}$$

As shown in Lemma 4, the reordered feedbacks and rewards $h'_{i,1}, h'_{i,2}, \ldots, h'_{i,T'_i(n')}, \ldots h'_{i,T_i(n)}$ are i.i.d. with the same distribution as the original feedback sequence

$\{h_{t,i}\}_{t\in\mathbb{N}}$. The base algorithm BASE has worked on the first $T'_i(n)$ of these rewards (in its extended run). Therefore, BASE has operated for n steps in a simulated environment with the same reward and feedback distributions, but without delay. Therefore, the first summation in the right hand side of (8) is in fact $\mathbb{E}\left[R_n^{\text{Base}}\right]$, the expected regret of the base algorithm in a non-delayed environment. This concludes the proof. □

4. UCB for the Multi-Armed Bandit Problem with Delayed Feedback

While the algorithms in the previous section provide an easy way to convert algorithms devised for the non-delayed case to ones that can handle delays in the feedback, improvements can be achieved if one makes modification inside the existing non-delayed algorithms while retaining their theoretical guarantees. This can be viewed as a "white-box" approach to extending online learning algorithms to the delayed setting, and enables us to escape the high memory requirements of black-box algorithms that arises for both of our methods in the previous section when the delays are large. We consider the stochastic multi-armed bandit problem, and extend the UCB-family of algorithms (Auer et al., 2002; Garivier & Cappé, 2011) to the delayed setting. The modification proposed is quite natural, and the common characteristics of UCB-type algorithms enable a unified way of extending their performance guarantees to the delayed setting (up to an additive penalty due to delays).

Recall that in the stochastic MAB setting, which is a special case of the stochastic IPM problem of Section 3.2, the feedback at time instant t is $h_t = r(a_t, b_t)$, and there is a distribution ν_i from which the rewards of each action i are drawn in an i.i.d. manner. Here we assume that the rewards of different actions are independent of each other. We use the same notation as in Section 3.2.

Several algorithms devised for the non-delayed stochastic MAB problem are based on upper confidence bounds (UCBs), which are optimistic estimates of the expected reward of different actions. Different UCB-type algorithms use different upper confidence bounds, and choose, at each time instant, an action with the largest UCB.

Let $B_{i,s,t}$ denote the UCB for action i at time instant t, where s is the number of reward samples used in computing the estimate. In a non-delayed setting, the choice of a UCB-type algorithm at time instant t is given by $a_t = \text{argmax}_{i\in\mathcal{A}}\ B_{i,T_i(t-1),t}$. In the presence of delays, one can simply use the same upper confidence bounds using the rewards that have been observed, and choose action

$$a_t = \text{argmax}_{i\in\mathcal{A}}\ B_{i,S_i(t-1),t} \tag{9}$$

at time instant t (recall that $S_i(t-1)$ is the number of rewards that can be observed for action i before time instant t). Note that if the delays are zero, this algorithm reduces to the corresponding non-delayed version of the algorithm.

The algorithms defined by (9) can easily be shown to enjoy the same regret guarantees compared to their non-delayed versions, up to an additive penalty depending on the delays. This is because the analyses of the regrets of UCB algorithms follow the same pattern of upper bounding the number of trials of a suboptimal action using concentration inequalities suitable for the specific form of UCBs they use.

As an example, the UCB1 algorithm of Auer et al. (2002) uses UCBs of the form $B_{i,s,t} = \hat{\mu}_{i,s} + \sqrt{2\log(t)/s}$, where $\hat{\mu}_{i,s} = \frac{1}{s}\sum_{t=1}^{s} h'_{i,t}$ is the average over the first s observed rewards. Using this upper confidence bound in our decision rule (9), we can bound the regret of the resulting algorithm in the delayed setting:

Theorem 7. *For any $n \geq 1$, the expected regret of the Delayed-UCB1 algorithm is bounded by*

$$\mathbb{E}[R_n] \leq 8\left[\sum_{i:\mu_i<\mu_{i^*}} \frac{\log n}{\Delta_i}\right] + 3.5\sum_{i=1}^{K}\Delta_i + \left(\sum_{i=1}^{K}\Delta_i \mathbb{E}\left[G^*_{i,n}\right]\right). \tag{10}$$

Note that the last term in the bound is the additive penalty, and, under different assumptions, it can be bounded in the same way as after Theorem 6. A similar analysis is also shown in the appendix for the KL-UCB algorithm of Garivier & Cappé (2011).

5. Conclusion and future work

We analyzed the effect of feedback delays in online learning problems. We examined the partial monitoring case (which also covers the full information and the bandit settings), and provided general algorithms that transform algorithms devised for the non-delayed case into ones that can tolerate delays in the feedback. It turns out that the price of delay is a multiplicative increase in the regret in adversarial problems, and only an additive increase in stochastic problems. While we

believe that these findings are qualitatively correct, we do not have lower bounds to prove such a claim (matching lower bounds are available for the full information case only).

It also turns out that the most important quantity that determines the performance of our algorithms is G_n^*, the maximum number of missing rewards. It is interesting to note that G_n^* is the maximum number of servers used in a multi-server queuing system with infinitely many servers, deterministic arrival times and independent service times. It is also the maximum deviation of a certain type of Markov chains. While we have not found any immediately applicable results in these fields, we think that applying techniques from these areas could lead to an improved understanding of G_n^*, and hence an improved analysis of online learning problems with delayed feedback.

References

Agarwal, Alekh and Duchi, John C. Distributed delayed stochastic optimization. In Shawe-Taylor, J., Zemel, R.S., Bartlett, P., Pereira, F., and Weinberger, K.Q. (eds.), *Advances in Neural Information Processing Systems 24*, 2011.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006.

Desautels, Thomas, Krause, Andreas, and Burdick, Joel. Parallelizing exploration-exploitation tradeoffs with gaussian process bandit optimization. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

Doob, J.L. *Stochastic processes*. Wiley, 1953.

Dudik, Miroslav, Hsu, Daniel, Kale, Satyen, Karampatziakis, Nikos, Langford, John, Reyzin, Lev, and Zhang, Tong. Efficient optimal learning for contextual bandits. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 2011.

Garivier, Aurélien and Cappé, Olivier. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011.

Hoeffding, Wassily. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):pp. 13–30, 1963. ISSN 01621459.

Langford, John, Smola, Alexander, and Zinkevich, Martin. Slow learners are fast. *arXiv:0911.0491*, November 2009.

Li, Lihong, Chu, Wei, Langford, John, and Schapire, Robert E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pp. 661–670, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772758.

Mesterharm, Chris J. On-line learning with delayed label feedback. In Jain, Sanjay, Simon, Hans, and Tomita, Etsuji (eds.), *Algorithmic Learning Theory*, volume 3734 of *Lecture Notes in Computer Science*, pp. 399–413. Springer Berlin / Heidelberg, 2005. ISBN 978-3-540-29242-5.

Mesterharm, Chris J. *Improving on-line learning*. PhD thesis, Rutgers University, New Brunswick, NJ, 2007.

Neu, G., György, A., Antos, A., and Szepesvári, Cs. Online markov decision processes under bandit feedback. Extended version of a paper submitted to NIPS-2010, 2010.

Titchmarsh, E.C. *The theory of the Riemann zeta-function*. Oxford University Press, USA, 1987.

Weinberger, M.J. and Ordentlich, E. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, July 2002. doi: 10.1109/ISIT.2002.1023420.

A. Proof of Lemma 4

Here we prove Lemma 4, used in Section 3.2. First, we will need the following two technical lemmas. The first lemma shows that the i.i.d. property is preserved when the sequence of feedbacks is reordered.

Lemma 8. *Let $\{X_t\}_{t\in\mathbb{N}}$, be a sequence of independent, identically distributed random variables. If we reorder this sequence according to an independent random permutation, then the resulting sequence is i.i.d. with the same distribution as $\{X_t\}_{t\in\mathbb{N}}$.*

Proof. Let the reordered sequence be denoted by $\{Z_t\}_{t\in\mathbb{N}}$. We need to show that for all $n \in \mathbb{N}$, for all $y_1, y_2, \ldots, y_n$, we have

$$\mathbb{P}\{Z_1 \le y_1, Z_2 \le y_2, \ldots, Z_n \le y_n\} = \mathbb{P}\{X_1 \le y_1, X_2 \le y_2, \ldots, X_n \le y_n\}.$$

Since $\{X_t\}_{t\in\mathbb{N}}$ is i.i.d., for any fixed permutation the equation above holds as both sides are equal to $\Pi_{t=1}^{n}\mathbb{P}\{X_t \le y_t\}$. Since the permutations are independent of the sequence $\{X_t\}_{t\in\mathbb{N}}$, using the law of total probability this extends to the general case as well. □

We also need the following lemma (Doob, 1953, Page 145, Chapter III, Theorem 5.2).

Lemma 9. *Let $\{X_t\}_{t\in\mathbb{N}}$ be a sequence of i.i.d. random variables, and $\{X'_t\}_{t\in\mathbb{N}}$ be its subsequence such that the decision whether to include X_t in the subsequence is independent of future values in the sequence, i.e., X_s for $s \ge t$. Then the sequence $\{X'_t\}_{t\in\mathbb{N}}$ is an i.i.d. sequence with the same distribution as $\{X_t\}_{t\in\mathbb{N}}$.*

We can now proceed to the proof of Lemma 4.

Proof (Lemma 4). Define $\{Z_{i,t}\}_{t\in\mathbb{N}}$ to be the sequence resulting from sorting the variables $h_{i,t}$ by their *possible* observation times $t + \tau_{i,t}$ (that is, $Z_{i,1}$ is the earliest reward that can be obtained if action i is chosen at the appropriate time, and so on). Since delays are independent of the outcomes, they define an independent reordering on the sequence of feedbacks. Hence, by Lemma 8, $\{Z_{i,t}\}_{t\in\mathbb{N}}$ is an i.i.d. sequence with the same distribution as $\{h_{i,t}\}_{t\in\mathbb{N}}$. Note that $\{h'_{i,s}\}_{s\in\mathbb{N}}$, the sequence of feedbacks of the choices of action i by the agent sorted by their observation times, is a subsequence of $\{Z_{i,t}\}_{t\in\mathbb{N}}$ where the decision whether to include each $Z_{i,t}$ in the subsequence cannot depend on future possible observations $Z_{i,s}, s \ge t$. Also, the feedbacks of other actions that are used in this decision are independent of $Z_{i,t}$. Hence, by Lemma 9, $\{h'_{i,t}\}_{t\in\mathbb{N}}$ is an i.i.d. sequence with the same distribution as $Z_{i,t}$, which in turn has the same distribution as $h_{i,t}$. □

B. UCB for the Multi-Armed Bandit Problem with Delayed Feedback

Here we give a detailed analysis of our UCB-based algorithms defined in Section 4. The regret of a UCB algorithm is usually analyzed by upper bounding the number of times a suboptimal action is chosen, and then using Equation 5. Consider a UCB algorithm with upper confidence bounds $B_{i,s,t}$, and fix a suboptimal action i. The typical analysis (e.g., in (Auer et al., 2002)) guesses a logarithmic upper bound $\ell > 1$ on the expected number of trials of suboptimal action i, and uses concentration inequalities suitable for the specific form of the upper-confidence bound to show that it is unlikely to choose a suboptimal action after having seen ℓ samples of its reward. Examples of such concentration inequalities include Hoeffding's inequality (Hoeffding, 1963) and Theorem 10 of (Garivier & Cappé, 2011), which are used for UCB1 and KL-UCB, respectively.

The general analysis works as follows: for $\ell > 1$, $T_i(n) \le \ell + \sum_{t=1}^{n}\mathbb{I}\{a_t = i,\ T_i(t) > \ell\}$, where the sum on the right hand side captures how much larger than ℓ the number of trials $T_i(n)$ is. Whenever action i is chosen, its UCB, $B_{i,T_i(t-1),t}$, must have been greater than that of an optimal action, $B_{i^*,\ T_{i^*}(t-1),t}$, which implies

$$T_i(n) \le \ell + \sum_{t=1}^{n}\mathbb{I}\{B_{i,T_i(t-1),t} \ge B_{i^*,T_{i^*}(t-1),t}, a_t = i,\ T_i(t-1) \ge \ell\}. \tag{11}$$

In the delayed setting, in the same way as above we can write

$$T_i(n) \le \ell + \sum_{t=1}^{n} \mathbb{I}\{B_{i,S_i(t-1),t} \ge B_{i^*,S_{i^*}(t-1),t},\ a_t = i,\ T_i(t-1) \ge \ell\}.$$

Since $T_i(t-1) = G_{i,t} + S_i(t-1)$, with $\ell' = \ell - G^*_{i,n}$ we get:

$$T_i(n) \le \ell' + G^*_{i,n} + \sum_{t=1}^{n} \mathbb{I}\{B_{i,S_i(t-1),t} \ge B_{i^*,S_{i^*}(t-1),t}, a_t = i,\ S_i(t-1) \ge \ell'\}. \tag{12}$$

The same concentration inequalities used to bound (11) in the analysis for the non-delayed setting can now be used to upper bound the sum in (12). Thus, putting this into (5), we see that one can use the same upper confidence bound in the delayed setting (with only the observed rewards) and get a performance similar to the non-delayed setting. The following two sections illustrate the use of this method on two UCB-type algorithms.

B.1. UCB1 under delayed feedback: Proof of Theorem 7

Below comes the proof of Theorem 7.

Proof of Theorem 7. Following the outline of the previous section, we can bound the summation in (12) using the same analysis as in the original UCB1 paper (Auer et al., 2002). In particular, for any action i we can write

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{I}\left\{B_{i,S_i(t-1),t} \ge B_{i^*,S_{i^*}(t-1),t},\ S_i(t-1) \ge \ell'\right\} \\
&\le \sum_{t=1}^{n} \mathbb{I}\left\{B_{i^*,S_{i^*}(t-1),t} \le \mu_{i^*},\ S_i(t-1) \ge \ell'\right\} + \sum_{t=1}^{n} \mathbb{I}\left\{B_{i,S_i(t-1),t} \ge \mu_{i^*},\ S_i(t-1) \ge \ell'\right\}.
\end{aligned} \tag{13}$$

The event in the first summation implies that either $\hat{\mu}_{i^*,S_{i^*}(t-1)} + \sqrt{\frac{2\log(t)}{S_{i^*}(t-1)}} \le \mu_{i^*}$ or $\hat{\mu}_{i,S_i(t-1)} - \sqrt{\frac{2\log(t)}{S_i(t-1)}} \ge \mu_i$. Hence,

$$\begin{aligned}
(13) \le\ & \sum_{t=1}^{n} \mathbb{I}\left\{\hat{\mu}_{i^*,S_{i^*}(t-1)} + \sqrt{\frac{2\log(t)}{S_{i^*}(t-1)}} \le \mu_{i^*}\right\} + \\
& \sum_{t=1}^{n} \mathbb{I}\left\{\hat{\mu}_{i,S_i(t-1)} - \sqrt{\frac{2\log(t)}{S_i(t-1)}} \ge \mu_i\right\} + \\
& \sum_{t=1}^{n} \mathbb{I}\left\{\mu_i + 2\sqrt{\frac{2\log(t)}{S_i(t-1)}} > \mu_{i^*},\ S_i(t-1) \ge \ell'\right\}.
\end{aligned} \tag{14}$$

Choosing $\ell' = \left\lceil \frac{8\log(n)}{\Delta_i^2} \right\rceil$ makes the events in the last summation above impossible, because $S_i(t-1) \ge \ell' \ge \frac{8\log(n)}{\Delta_i^2}$ which implies $2\sqrt{\frac{2\log(t)}{S_i(t-1)}} \le 2\sqrt{\frac{2\log(n)}{\ell'}} \le \Delta_i$. Therefore, combining with (12), we can write

$$T_i(n) \ \le\ \left\lceil \frac{8\log(n)}{\Delta_i^2} \right\rceil + G^*_{i,n} + \sum_{t=1}^{n} \sum_{s=1}^{t} \left(\mathbb{I}\left\{\hat{\mu}_{i^*,s} \ + \ \sqrt{\frac{2\log(t)}{s}} \le \mu_{i^*}\right\} + \mathbb{I}\left\{\hat{\mu}_{i,s} - \sqrt{\frac{2\log(t)}{s}} \ge \mu_i\right\} \right).$$

Taking expectation gives

$$\mathbb{E}\left[T_i(n)\right] \ \le\ \left\lceil \frac{8\log(t)}{\Delta_i^2} \right\rceil + \mathbb{E}\left[G^*_{i,n}\right] + \sum_{t=1}^{n} \sum_{s=1}^{t} \left(\mathbb{P}\left\{\hat{\mu}_{i^*,s} \ + \ \sqrt{\frac{2\log(t)}{s}} \le \mu_{i^*}\right\} + \mathbb{P}\left\{\hat{\mu}_{i,s} - \sqrt{\frac{2\log(t)}{s}} \ge \mu_i\right\} \right).$$

We can use the concentration inequality used in the original analysis, namely Hoeffding's inequality, to bound each of the probabilities in the summation:

$$\mathbb{P}\left\{\hat{\mu}_{i^*,s} \; + \; \sqrt{\frac{2\log(t)}{s}} \leq \mu_{i^*}\right\} \leq e^{-4\log(t)} = t^{-4},$$
$$\mathbb{P}\left\{\hat{\mu}_{i,s} - \sqrt{\frac{2\log(t)}{s}} \geq \mu_i\right\} \leq e^{-4\log(t)} = t^{-4}.$$

Therefore, we have

$$\mathbb{E}\left[T_i(n)\right] \; \leq \; \left\lceil \frac{8\log(t)}{\Delta_i^2} \right\rceil + \mathbb{E}\left[G^*_{i,n}\right] \; + \; \sum_{t=1}^{\infty} 2t^{-3} \leq \frac{8\log(n)}{\Delta_i^2} + 1 + \mathbb{E}\left[G^*_{i,n}\right] \; + 2\zeta(3),$$

where $\zeta(3) < 1.21$ is the Riemann Zeta function.[2] Combining with (5) proves the theorem. □

B.2. KL-UCB under delayed feedback

The KL-UCB algorithm was introduced by Garivier & Cappé (2011). The upper confidence bound used by KL-UCB for action i at time t, is $B_{i,T_i(t-1),t}$, where $B_{i,s,t}$ is

$$\max\left\{q \in [\hat{\mu}_{i,s}, 1] : sd(\hat{\mu}_{i,s}, q) \leq \log t + 3\log(\log t)\right\},$$

with $d(p,q) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$ being the KL-divergence of two Bernoulli random variables with parameters p and q. The original authors show (Garivier & Cappé, 2011, Theorem 2) that there exists a constant $C_1 \leq 7$, as well as functions $C_2(\epsilon) = O(\epsilon^{-2})$ and $0 \leq \beta(\epsilon) = O(\epsilon^2)$, such that for any $\epsilon > 0$, the expected regret of the KL-UCB algorithm satisfies

$$\mathbb{E}\left[R_n\right] \leq \sum_{\Delta_i > 0} \Delta_i \left[\frac{\log(n)}{d(\mu_i, \mu_{i^*})}(1+\epsilon) + C_1\log(\log n) + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}}\right]. \tag{15}$$

.

Using this upper confidence bound with (9), we arrive at the Delayed-KL-UCB algorithm. We can prove the following theorem using the general scheme of Section 4 and the same techniques as in (Garivier & Cappé, 2011), again having an additive penalty compared to non-delayed setting.

Theorem 10. *For any $\epsilon > 0$, the expected regret of the Delayed-KL-UCB algorithm after n time instants is bounded by*

$$\mathbb{E}\left[R_n\right] \leq \sum_{i:\Delta_i>0} \Delta_i \left(\frac{\log(n)}{d(\mu_i,\mu_{i^*})}(1+\epsilon)C_1\log(\log(n))\right) + \sum_{i=1}^{K} \Delta_i \left(\frac{C_2(\epsilon)}{n^{\beta(\epsilon)}}\mathbb{E}\left[G^*_{i,n}\right] + \mathbb{E}\left[G^*_{i,n}\right] + 1\right),$$

where C_1, C_2, and β are the same as in (15).

In this case, working out the proof and reusing the analysis is a bit more complicated than in the case of UCB1. In particular, we will need the following lemma which is an adaptation of Lemma 7 of Garivier & Cappé (2011).

Lemma 11. *Let $d^+(x,y) = d(x,y)\mathbb{I}\{x < y\}$. Then for any $n \geq 1$,*

$$\sum_{t=1}^{n} \mathbb{I}\left\{a_t = i, \mu_{i^*} \leq B_{i^*,S_{i^*}(t-1),t}, \; S_i(t-1) \geq \ell'\right\} \leq G^*_{i,n} \sum_{s=\ell'}^{n} \mathbb{I}\left\{sd^+(\hat{\mu}_{i,s}, \mu_{i^*}) < \log(n) + 3\log(\log(n))\right\}.$$

Proof of Lemma 11. We start in the same way as the original proof. Note that $d^+(p,q)$ is non-decreasing in its second parameter, and that $a_t = i$ and $\mu_{i^*} \leq B_{i^*,S_{i^*}(t-1),t}$ together imply $B_{i,S_i(t-1),t} \geq B_{i^*,S_{i^*}(t-1),t} \geq \mu_{i^*}$, which in turn gives

$$d^+(\hat{\mu}_{i,S_i(t-1)}, \mu_{i^*}) \leq d(\hat{\mu}_{i,S_i(t-1)}, B_{i,S_i(t-1),t}) \leq \frac{\log(t) + 3\log(\log(t))}{S_i(t-1)}.$$

[2] For properties and theory of the Riemann Zeta function, see the book of (Titchmarsh, 1987)

Therefore, we have

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{I}\left\{a_t = i, \mu_{i^*} \le B_{i^*, S_{i^*}(t-1), t},\ t > S_i(t-1) \ge \ell'\right\} \\
&\le \sum_{t=\ell'}^{n} \mathbb{I}\left\{a_t = i, S_i(t-1) d^+(\hat{\mu}_{i,S_i(t-1)}, \mu_{i^*}) \le \log(t) + 3\log(\log(t)),\ S_i(t-1) \ge \ell'\right\} \\
&\le \sum_{t=\ell'}^{n} \mathbb{I}\left\{a_t = i, S_i(t-1) d^+(\hat{\mu}_{i,S_i(t-1)}, \mu_{i^*}) \le \log(n) + 3\log(\log(n)),\ S_i(t-1) \ge \ell'\right\} \\
&\le \sum_{t=\ell'}^{n} \sum_{s=\ell'}^{t} \mathbb{I}\left\{a_t = i, S_i(t-1) = s\right\} \times \mathbb{I}\left\{s d^+(\hat{\mu}_{i,s}, \mu_{i^*}) \le \log(n) + 3\log(\log(n))\right\} \\
&= \sum_{s=\ell'}^{n} \sum_{t=s}^{n} \mathbb{I}\left\{a_t = i, S_i(t-1) = s\right\} \times \mathbb{I}\left\{s d^+(\hat{\mu}_{i,s}, \mu_{i^*}) \le \log(n) + 3\log(\log(n))\right\} \\
&= \sum_{s=\ell'}^{n} \mathbb{I}\left\{s d^+(\hat{\mu}_{i,s}, \mu_{i^*}) \le \log(n) + 3\log(\log(n))\right\} \times \left(\sum_{t=s}^{n} \mathbb{I}\left\{a_t = i, S_i(t-1) = s\right\}\right).
\end{aligned}$$

But note that the second summation is bounded by $G^*_{i,n}$, because for each s, there cannot be more than $G^*_{i,n}$ time instants at which action i is played while $S_i(t) = s$ stays constant, otherwise we would have $T_i(t'-1) - S_i(t'-1) = G_{i,t'} > G^*_{i,n}$ for some $t' \in \{s, \dots, n\}$, which is not possible. Substituting this bound in the last expression proves the lemma. □

We also need the following two results from the original paper by (Garivier & Cappé, 2011).

Theorem 12 (Theorem 10 of (Garivier & Cappé, 2011)). *Let $\{Y_t\}, t \ge 1$ be a sequence of independent random variables bounded in $[0,1]$, with common expectation $\mu = \mathbb{E}[Y_t]$. Consider a sequence $\{\epsilon_t\}, t \ge 1$ of Bernoulli variables such that for all $t > 0$, ϵ_t is a random function of $Y_1, \dots, Y_{t-1}$ [3], and is independent of $Y_s, s \ge t$. Let $\delta > 0$ and for every $1 \le t \le n$, let*

$$S_t = \sum_{s=1}^{t} \epsilon_s \quad \text{and} \quad \hat{\mu}_t = \frac{\sum_{s=1}^{t} \epsilon_s Y_s}{S_t},$$

with $\bar{Y}_t = 0$ when $S_t = 0$, and

$$B_n = \operatorname{argmax}\left\{q > \hat{\mu}_n : S_n d(\hat{\mu}_n, q) \le \delta\right\}.$$

Then

$$\mathbb{P}\left\{B_n < \mu\right\} \le e\lceil \delta \log(n) \rceil e^{-\delta}.$$

Lemma 13 (Lemma 8 of (Garivier & Cappé, 2011)). *For a suboptimal action i, for every $\epsilon > 0$, let $K_n = \left\lfloor \frac{1+\epsilon}{d(\mu_i, \mu_{i^*})} \left(\log(n) + 3\log(\log(n)) \right) \right\rfloor$. Then there exist $C_2(\epsilon) > 0$ and $\beta(\epsilon) > 0$ such that*

$$\sum_{s=K_n+1}^{\infty} \mathbb{P}\left\{ d^+(\hat{\mu}_{i,s}, \mu_{i^*}) < \frac{d(\mu_i, \mu_{i^*})}{1+\epsilon} \right\} \le \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}}.$$

Now, we are ready to prove Theorem 10.

[3] That is, a function of $Y_1, \dots, Y_{t-1}$ together with possibly an extra, independent randomization.

Proof of Theorem 10. For an action i, bounding the terms in (12) gives

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{I}\left\{a_t = i, B_{i,S_i(t-1),t} \geq B_{i^*,S_{i^*}(t-1),t},\ S_i(t-1) \geq \ell'\right\} \\
&\leq \sum_{t=1}^{n} \mathbb{I}\left\{B_{i^*,S_{i^*}(t-1),t} < \mu_{i^*}\right\} + \sum_{t=1}^{n} \mathbb{I}\left\{a_t = i,\ \mu_{i^*} \leq B_{i^*,S_{i^*}(t-1),t},\ S_i(t-1) \geq \ell'\right\} \\
&\leq \sum_{t=1}^{n} \mathbb{I}\left\{B_{i^*,S_{i^*}(t-1),t} < \mu_{i^*}\right\} + G^*_{i,n} \sum_{s=\ell'}^{n} \mathbb{I}\left\{s d^+(\hat{\mu}_{i,s}, \mu_{i^*}) < \log(n) + 3\log(\log(n))\right\}, \qquad (16)
\end{aligned}$$

where the last inequality follows from Lemma 11. Let

$$K_n = \left\lfloor \frac{1+\epsilon}{d(\mu_i, \mu_{i^*})} \left(\log(n) + 3\log(\log(n)) \right) \right\rfloor \qquad (17)$$

and

$$\ell' = 1 + K_n. \qquad (18)$$

Then we have:

$$\begin{aligned}
&\sum_{s=\ell'}^{n} \mathbb{I}\left\{s d^+(\hat{\mu}_{i,s}, \mu_{i^*}) \leq \log(n) + 3\log(\log(n))\right\} \\
&\qquad \leq \sum_{s=K_n+1}^{\infty} \mathbb{I}\left\{(K_n+1) d^+(\hat{\mu}_{i,s}, \mu_{i^*}) \leq \log(n) + 3\log(\log(n))\right\} \\
&\qquad \leq \sum_{s=K_n+1}^{\infty} \mathbb{I}\left\{ d^+(\hat{\mu}_{i,s}, \mu_{i^*}) < \frac{d(\mu_i, \mu_{i^*})}{1+\epsilon} \right\}. \qquad (19)
\end{aligned}$$

Putting inequalities (17) trough (19) back into inequality (16), from (12) we get:

$$\begin{aligned}
\mathbb{E}\left[T_i(n)\right] \leq\ & \frac{1+\epsilon}{d(\mu_i, \mu_{i^*})} \left(\log(n) + 3\log(\log(n)) \right) + \mathbb{E}\left[G^*_{i,n}\right] + 1 + \\
& \sum_{t=1}^{n} \mathbb{P}\left\{B_{i^*,S_{i^*}(t-1),t} < \mu_{i^*}\right\} + \mathbb{E}\left[G^*_{i,n}\right] \sum_{s=K_n+1}^{\infty} \mathbb{P}\left\{ d^+(\hat{\mu}_{i,s}, \mu_{i^*}) < \frac{d(\mu_i, \mu_{i^*})}{1+\epsilon} \right\}.
\end{aligned}$$

where the last term is a result of the delays being independent of the rewards. The first summation can be bounded using Theorem 12, for which it suffices to set $\epsilon_t = 1, 1 \leq t \leq n$, and use the sequence of observed rewards $\{h'_{i,t}\}$ for the arm under consideration as the sequence $\{Y_t\}$ in the theorem. The second summation can be bounded by Lemma 13. Therefore, the expectation of the number of trials of the suboptimal action is bounded by:

$$\mathbb{E}\left[T_i(n)\right] \leq \frac{1+\epsilon}{d(\mu_i, \mu_{i^*})} \left(\log(n) + 3\log(\log(n)) \right) + C_1 \log(\log(n)) + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}} \mathbb{E}\left[G^*_{i,n}\right] + \mathbb{E}\left[G^*_{i,n}\right] + 1.$$

Combining the above inequality with (5) finishes the proof. □