Stochastic Bandit Models for Delayed Conversions

Claire Vernade
LTCI, Telecom ParisTech,
Université Paris-Saclay

Olivier Cappé
LIMSI, CNRS,
Université Paris-Saclay

Vianney Perchet
CMLA, ENS Paris-Saclay,
Université Paris-Saclay
& Criteo Research

Abstract

Online advertising and product recommendation are important domains of applications for multi-armed bandit methods. In these fields, the reward that is immediately available is most often only a proxy for the actual outcome of interest, which we refer to as a *conversion*. For instance, in web advertising, clicks can be observed within a few seconds after an ad display but the corresponding sale –if any– will take hours, if not days to happen. This paper proposes and investigates a new stochastic multi-armed bandit model in the framework proposed by Chapelle (2014) –based on empirical studies in the field of web advertising– in which each action may trigger a future reward that will then happen with a stochastic delay. We assume that the probability of conversion associated with each action is unknown while the distribution of the conversion delay is known, distinguishing between the (idealized) case where the conversion events may be observed whatever their delay and the more realistic setting in which late conversions are censored. We provide performance lower bounds as well as two simple but efficient algorithms based on the UCB and KLUCB frameworks. The latter algorithm, which is preferable when conversion rates are low, is based on a Poissonization argument, of independent interest in other settings where aggregation of Bernoulli observations with different success probabilities is required.

1 INTRODUCTION

Characterizing the relationship between marketing actions and users' decisions is of prime importance in advertising, product recommendation and customer relationship management. In online advertising a key aspect of the problem is that whereas marketing actions can be taken very fast –typically in less than a tenth of a second, if we think of an ad display–, user's buying decisions will happen at a much slower rate [7, 4, 18]. In the following, we refer to a user's decision of interest under the generic term of *conversion*. Chapelle, in [4], while analyzing data from the real-time bidding company Criteo, observed that, on average, only 35% of the conversions occurred within the first hour. Furthermore, about 13% of the conversions could be attributed to ad display that were more than two weeks old. Another interesting observation from this work is the fact that the delay distribution could be reasonably well fitted by an exponential distribution, particularly when conditioning on context variables that are available to the advertiser.

The present work addresses the problem of sequentially learning to select relevant items in the context where the feedback happens with long delays. By long we mean in particular that the feedback associated with a fraction of the actions taken by the learner will be practically unobserved because they will happen with an excessive delay. In the example cited above, if we were to run an online algorithm during two weeks, at least 13% of the actions would not receive an observable feedback because of delays. A related situation occurs if the online algorithm is run during, say, one month, but its memory is limited to a sliding window of two weeks. In Section 2 below we introduce models suitable for addressing these two related situations in the framework of multi-armed bandits.

Delayed feedback is a topic that has been considered before in the reinforcement learning literature and we defer the precise comparison between existing approaches and the proposed framework to Section 3. In a nutshell however, the distinctive features of our approach is to consider potentially infinite stochastic delays, resulting in some feedback being *censored* (ie. not observable). Existing works on delayed bandits focus on cases where

the feedback is observed after some delay, typically assumed to be finite. In contrast, we assume that delays are random with a distribution that may have an unbounded support – although we require that it has finite expectation. As a result, some conversion events cannot be observed within any finite horizon and the proposed learning algorithm must take this uncertainty into account.

In Section 2, we propose discrete-time stochastic multi-armed bandit models to address the problem of long delays with possibly censored feedback. We distinguish two situations that correspond to the cases mentioned informally above: In the *uncensored* model, conversions can be assumed to be eventually observed after some possibly arbitrarily long delay; In the *censored* model, it is assumed that the environment imposes that the conversions related to actions cannot be observed anymore after a finite window of m time steps.

Assuming that the delay distribution is known, we prove problem-dependent lower bounds on the regret of any uniformly efficient bandit algorithm for the censored and uncensored models in Section 4.

In Section 5, we describe efficient anytime policies relying on optimistic indices, based on the UCB [1] or KLUCB [5] algorithms. The latter uses a Poissonization argument that can be of independent interest in other bandit models. In typical scenarios where the conversion rates are less than one percent, using the KLUCB variant will ensure a much faster learning and provides near-optimal perfomance on the long run (see Theorem 11).

These algorithms are analyzed in Section 6, showing that they reach close to optimal asymptotic performance. In Section 7 we discuss the implementation of these methods, showing that it is further simplified in the case of geometrically distributed delays, and we illustrate their performance on simulated data.

2 A STOCHASTIC MODEL FOR THE DELAYS

We now describe our bandit setting for delayed conversion events, inspired by [4]. We first consider the setting in which delays may be potentially unbounded and then consider the case where censoring occurs.

2.1 GENERAL BANDIT MODEL UNDER DELAYED FEEDBACK

At each round $t \in \mathbb{N}^*$, the learner chooses an arm $A_t \in \{1, \dots, K\}$. This action simultaneously triggers two independent random variables:

- $C_t \in \{0, 1\}$, is the *conversion indicator* that is equal to 1 only if the action A_t will lead to a conversion;
- $D_t \in \mathbb{N}$, is *the delay* indicating the number of time steps needed before the conversion – if any – be disclosed to the learner.

At each round t, the agent then receives an integer-valued reward Y_t which corresponds to the number of observed conversions at time t:

$$Y_t = \sum_{s=1}^{t} C_s \mathbb{1}\{D_s = t - s\}.$$

In the following, we will use the short-hand notation $X_{s,t} = C_s \mathbb{1}\{D_s \leq t - s\}$, for $s \leq t$ to denote the possible contribution of the action taken at time s to the conversion(s) observed at a later time t. We emphasize that even if the actual reward of the learner is the sum of the conversions, we assume that the agent also observes all the individual contributions $(X_{s,t})_{1 \leq s \leq t}$ at time t triggered by actions taken before time t.

The above mechanism implies that if $C_t = 1$, the learner will observe D_t at time $t + D_t$, whereas if $C_t = 0$, the delay will not be directly observable. In particular, if at time t, $X_{s,u} = 0$, for all $s \leq u \leq t$, it is impossible to decide whether $C_s = 0$ or if $C_s = 1$ but $D_s > t - s$. Formally, the history of the algorithm is the σ-field generated by $\mathcal{H}_t := (X_{s,u})_{1 \leq u \leq t, 1 \leq s \leq u}$.

We consider the stochastic model under the following basic assumptions:

$$C_t | \mathcal{H}_{t-1} \sim \text{Bernoulli}(\theta_{A_t}),$$
$$D_t | \mathcal{H}_{t-1} \sim \text{distribution with CDF } \tau,$$

and C_t, D_t are conditionally independent given $\mathcal{H}_{t-1}$.

Lemma 1. *Denote by $a^* \in \{1, \dots, K\}$ an index such that $\theta_a^* \geq \theta_k$, for $k \in \{1, \dots, K\}$, and define by $r(T) = \sum_{t=1}^{T} Y_t$ the cumulated reward of the learner and by $r^*(T)$ the cumulated reward obtained by an oracle playing $A_t = a^*$ at each round. The expected regret of the learner is given by*

$$L(T) = \mathbb{E}\left[r^*(T) - r(T)\right] = \sum_{s=1}^{T} \mathbb{E}\left[\theta_{a^*} - \theta_{A_s}\right] \tau_{T-s} \tag{1}$$

where by definition $\tau_{T-s} = \mathbb{P}(D_s \leq T - s)$. Denoting $\mathbb{E}[N_k(T)] := \sum_{s=1}^{T-1} \mathbb{1}\{A_s = k\}$, it holds that

$$L(T) \leq \sum_{k=1}^{K} (\theta_{a^\star} - \theta_k) N_k(t)$$

and if $\mu = \mathbb{E}[D_s] < \infty$,

$$\sum_{k=1}^{K} (\theta_{a^\star} - \theta_k) N_k(t) - L(T) \leq \mu \sum_{k=1}^{K} (\theta_{a^\star} - \theta_k). \tag{2}$$

Proof The cumulated reward at time T satisfies

$$r(T) = \sum_{t=1}^{T} Y_t = \sum_{t=1}^{T}\sum_{s=1}^{t} C_s \mathbf{1}\{D_s = t-s\}$$
$$= \sum_{s=1}^{T} C_s \mathbf{1}\{D_s \leq T-s\} = \sum_{s=1}^{T} X_{s,T},$$

where the index T stands for the time at which all past conversions are observed while s is the index at which the action has been taken. Hence Eq. (1) is obtained by

$$\mathbb{E}\left[\sum_{t=1}^{T} r(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$$
$$= \sum_{s=1}^{T} \mathbb{E}\left[X_{s,T}\right] = \sum_{s=1}^{T} \theta_{A_s}\tau_{T-s}.$$

Obviously the fact that $\tau_{T-s} \leq 1$ implies that $L(T)$ is upper bounded by $\sum_{k=1}^{K}(\theta_{a^\star} - \theta_k)N_k(t)$, which corresponds to the usual regret formula in the bandit model with explicit immediate feedback. To upper bound the difference, note that

$$\sum_{k=1}^{K}(\theta_{a^\star} - \theta_k)N_k(t) - L(T)$$
$$= \sum_{k=1}^{K}(\theta_{a^\star} - \theta_k)\sum_{s=1}^{T}\mathbf{1}\{A_s = k\}(1-\tau_{T-s})$$
$$\leq \sum_{k=1}^{K}(\theta_{a^\star} - \theta_k)\sum_{n=0}^{\infty}(1-\tau_n) = \mu\sum_{k=1}^{K}(\theta_{a^\star} - \theta_k).$$

■

2.2 THRESHOLDED DELAYS: CENSORED OBSERVATIONS

The model with m-thresholded delays takes into account the fact that a conversion can only be observed within m steps after the action occurred. This basically changes the expression of the expected instantaneous reward Y_t which becomes,

$$Y_t = \sum_{s=t-m}^{t} C_s \mathbf{1}\{D_s = t-s\}$$

and the future contributions of each action are capped to the next m time steps: $(X_{s,t})_{t-m\leq s\leq t}$. The history of the algorithm only consists of $\mathcal{H}_t = \sigma((X_{s,u})_{1\leq u\leq t, u-m\leq s\leq u})$ and the regret expression of Lemma 1 can be split into two terms corresponding to old pulls and the m most recent pulls:

$$\sum_{s=1}^{T-m}(\theta_{a^\star} - \mathbb{E}[\theta_{A_s}])\tau_m + \sum_{s=T-m+1}^{T}(\theta_{a^\star} - \mathbb{E}[\theta_{A_s}])\tau_{T-s} \tag{3}$$

In the remaining, for $(p,q) \in [0,1]^2$, we will denote by $d(p,q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$ the binary entropy between p and q, that is the Kullback-Leibler divergence between Bernoulli distributions with parameters p and q. Moreover, without loss of generality, we will assume that $a^* = 1$ is the unique optimal arm of the considered bandit problems and denote by $\theta^* = \theta_1$ the optimal conversion rate.

3 RELATED WORK ON DELAYED BANDITS

Delayed feedback recently received increasing attention in the bandit and online learning literature due to its various applications ranging from online advertising [4] to distributed optimization [10, 3]. Indeed, delayed feedback have been extensively considered in the context of Markov Decision Processes (MDPs) [12, 19]. However, the present work focuses on unbounded delays and the models considered therein would result in an infinite space MDP for which even the planning problem would be challenging. In contrast, the lack of memory in bandits makes it possible to propose relatively simple algorithms even in the case where the delays may be very long. For a review of previous works in online learning in the stochastic and non-stochastic settings, see [8] and references therein. The latter work tackles the more general problem of partial monitoring under delayed feedback, with Sections 3.2 and 4 of the paper focusing on the stochastic delayed bandit problem. A key insight from this work is that, in minimax analysis, delay increases the regret in a multiplicative way in adversarial problems, and in an additive way in stochastic problems.

The algorithm of [9] relies on a queuing principle termed Q-PMD that uses an optimistic bandit referred to as "BASE" to perform exploration; in [9] UCB is chosen as BASE strategy while the follow-up work [8] also considers the use of KLUCB. The idea is to store all the observations that arrive at the same time t in a FIFO buffer and to feed BASE with the information related to an arm k only when this arm is about to be chosen. It means that the number of draws of an arm as well as the cumulated sum of the subsequent rewards are only updated whenever the observation arrives to the learner. Meanwhile, the algorithm acts as if nothing happened.

However, in the setting considered in the present work,

updating counts only after the observations are eventually received cannot lead to a practical algorithm: Whenever a click is received, the associated reward is 1 by definition, otherwise the ambiguity between non-received and negative feedback remains. Thus, the empirical average of the rewards for each arm computed by the updating mechanism of Q-PMD sticks to 1 and does not allow to compare the arms. As a consequence, the Q-PMD policy cannot be used for the models described in Section 2, except in the specific case of the uncensored delay model with bounded delays: Then there is no censoring anymore as one only needs to wait long enough (longer than the maximal possible delay) to reveal with certainty the exact value of the feedback.

Also, [16] notices that the empirical performances of this queuing-based heuristic are not fully satisfying because of the lack of variability in the decisions made by the policy while waiting for feedback. Their suggestion is to use random policies instead of deterministic ones in order to improve the overall exploration. Note that even though we stick to deterministic, history-based, policies, this problem is taken care of by our algorithm thanks to the use of the CDF of the delays that allow us to correct the confidence intervals continuously after a pull has been made.

Another possible way to handle bounded delays would be to plan ahead the sequence of pulls by batches, following the principles of Explore Then Commit, see [17]. With finite delays, a new un-necessary batch of exploration pulls might be started before the algorithm enters the exploitation (or commitment) phase. The extra cost would therefore be the maximal observable delay. Although these techniques are random and not deterministic, they have the same drawbacks as the other ones: The policy is not updated while waiting for feedback and, as a consequence, cannot handle arbitrarily large delays.

An obvious limitation of our work is that we assume that the delay distribution is known. We believe that it is a realistic assumption however as the delay distribution can be identified from historical data as reported in [4]. In addition, as we assume that the same delay distribution is shared by all actions, it is natural to expect that estimating the delay distribution on-line can be done at no additional cost in terms of performance. Perhaps more interestingly, it is possible to extend the model so as to include cases where the context of each action is available to the learner and determines the distribution of the corresponding delay, using for instance the generalized linear modeling of [4]. In particular, the same algorithms can be used in this case, by considering the proper CDFs corresponding to different instances. Of course the analysis to be described below would need to be extended to cover also this contextual case.

4 LOWER BOUND ON THE REGRET

The purpose of this section is to provide lower bounds on the regret of *uniformly efficient* algorithms in the two different settings of the Stochastic Delayed Bandit problem that we consider. This class of policies, introduced by [15], refers to algorithms such that for any bandit model ν, and any $\alpha \in (0,1)$, $\mathbb{E}[R(T)]/T^\alpha \to 0$ when $T \to \infty$.

Our results rely on changes of measure argument that are encapsulated in Lemma 1 of [13], or more recently, and more generally, in Inequality (F) of [6]. Those results can actually be reformulated as a lower bound on the expected log-likelihood ratio of the observations under the originally considered bandit model θ and the alternative one θ'

$$\mathbb{E}[\ell_T] = \mathbb{E}_\theta\left[\frac{p_\theta((X_{s,t})_{1\le t\le T, 1\le s\le t})}{p_{\theta'}((X_{s,t})_{1\le t\le T, 1\le s\le t})}\right].$$

The following inequality is obtained using proof techniques from Appendix B of [13] that are detailed in Appendix B.

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\ell_T]}{\log(T)} \ge 1. \tag{4}$$

To obtain explicit regret lower bounds for the models introduced in Section 2, we compute below the expected log-likelihood ratio corresponding to these two models.

Lemma 2. *In the censored delayed feedback setting, the expected log-likelihood ratio is given by*

$$\mathbb{E}_\theta[\ell_T] = \sum_{s=1}^{T-m} d(\theta_{A_s}\tau_m, \theta'_{A_s}\tau_m) + \sum_{s=T-m}^{T} d(\theta_{A_s}\tau_{T-s}, \theta'_{A_s}\tau_{T-s}).$$

In the uncensored setting, the above sum is not split and we have

$$\mathbb{E}_\theta[\ell_T] = \sum_{s=1}^{T} d(\theta_{A_s}\tau_{T-s}, \theta'_{A_s}\tau_{T-s}).$$

Proof Given $\mathcal{H}_{s-1}$, $(X_{s,s}, \dots, X_{s,T})$ can be equal to

- $(0,\dots,0)$, with proba. $(1-\theta_{A_s}) + \theta_{A_s}(1-\tau_{T-s})$,
- $(0,\dots,0,1,1,\dots,1)$ with proba. $\theta_{A_s}\delta_{u-s}$, for $u = s,\dots,T$ (u denotes the position of 1 in the vector), where $\delta_k = \mathbb{P}(D_s \le k)$.

Hence,

$$\mathbb{E}_\theta\left[\log\frac{p_\theta(X_{s,s},\dots,X_{s,T})}{p_{\theta'}(X_{s,s},\dots,X_{s,T}}\middle|\mathcal{H}_{s-1}\right]$$
$$=\log\frac{1-\theta_{A_s}\tau_{T-s}}{1-\theta'_{A_s}\tau_{T-s}}(1-\theta_{A_s}\tau_{T-s})$$
$$+\sum_{u=s}^{T}\log\frac{\theta_{A_s}\delta_{u-s}}{\theta'_{A_s}\delta_{u-s}}\theta_{A_s}\delta_{u-s}$$
$$=\log\frac{1-\theta_{A_s}\tau_{T-s}}{1-\theta'_{A_s}\tau_{T-s}}(1-\theta_{A_s}\tau_{T-s})+\log\frac{\theta_{A_s}}{\theta'_{A_s}}\theta_{A_s}\tau_{T-s}$$
$$=d(\theta_{A_s}\tau_{T-s},\theta'_{A_s}\tau_{T-s}).$$

The equivalent expression for the censored case is easily deduced from the same calculations. ■

4.1 CENSORED SETTING

Using our notations, the following theorem provides a problem-dependent lower bound on the regret.

Theorem 3. *The regret of any uniformly efficient algorithm is bounded from below by*

$$\liminf_{T\to\infty}\frac{R(T)}{\log(T)}\geq\sum_{k\neq k^*}\frac{\tau_m(\theta^*-\theta_k)}{d(\tau_m\theta_k,\tau_m\theta^*)}.$$

Proof The details of the proof can be found in Appendix B but we provide here a sketch of the main argument. The log-likelihood ratio is given by Lemma 2:

$$\mathbb{E}_\theta\left[\ell_T\right]=\sum_{s=1}^{T-m}d(\theta_{A_s}\tau_m,\theta'_{A_s}\tau_m)$$
$$+\sum_{s=T-m}^{T}d(\theta_{A_s}\tau_{T-s},\theta'_{A_s}\tau_{T-s}),$$

which is bounded from below by Eq.(4). However, obtaining a lower bound on the regret requires to decompose this quantity into $(K-1)$ terms depending on the suboptimal arms. For a fixed arm $k\neq 1$, we consider $\theta'=(\theta_1,\dots,\theta_{k-1},\theta_1+\epsilon,\dots,\theta_K)$ for which the expected log-likelihood ratio is

$$\mathbb{E}[N_k(T)]d(\tau_m\theta_k,\tau_m(\theta_1+\epsilon))$$
$$+\sum_{s=T-m}^{T}d(\theta_k\tau_{T-s},(\theta_1+\epsilon)\tau_{T-s})\geq\mathbb{E}_\theta\left[\ell_T\right].$$

Divide by $\log(T)$ and let T to infinity, to get the result for $\epsilon\to 0$, as the second term in the l.h.s. is bounded. ■

This lower bound implies that the delayed bandits problem with trespassing probability τ_m is as hard as solving the scaled bandit problem with expected rewards $(\tau_m\theta_1,\dots,\tau_m\theta_K)$. In the long run, one cannot learn faster than the heuristic approach discarding the last m observations and considerimg the fictitious bandit model with parameters $(\tau_m\theta_1,\dots,\tau_m\theta_K)$. However, on horizons of the order of m time-steps, we will show empirically in Section 7 that taking delay distributions into account allows for much faster learning. Note also that the convexity of the function $\tau\to d(\tau p,\tau q)$ proved in Lemma 15 implies that the regret lower bound is a monotonically increasing function of τ_m. Hence, either reduced values of m or longer values of the expected delay μ actually make the problem harder.

4.2 UNCENSORED SETTING

In the uncensored model, the same argument shows that the lower bound does not differ from the classical Lai & Robbins Lower Bound [15].

Theorem 4. *The regret of any uniformly efficient algorithm in the Uncensored Delays Setting is bounded from below by*

$$\liminf_{T\to\infty}\frac{R(T)}{\log(T)}\geq\sum_{k\neq k^*}\frac{(\theta^*-\theta_k)}{d(\theta_k,\theta^*)}.$$

The full proof of this result is similar to the proof of Theorem 3 and can be found in Appendix B.

5 DELAY-CORRECTED ESTIMATORS AND CONFIDENCE INTERVALS

In this section, for a fixed arm $k\in\{1,\dots,K\}$, we define a conditionally unbiased estimator for the conversion rate θ_k. Then, based on suitable concentration results we derive optimistic indices: a delay-corrected UCB as in [1] as well as a delay-corrected KLUCB as in [5].

5.1 PARAMETER ESTIMATOR

Define the sum of rewards up to time t as

$$S_k(t)=\sum_{s=1}^{t}\sum_{u=1}^{s}X_{u,s}\mathbb{1}\{A_u=k\}.$$

We recall that we defined the exact number of pulls of arm k up to time t as $N_k(t):=\sum_{s=1}^{t-1}\mathbb{1}\{A_s=k\}$. However, defining an estimator of θ_k that is unbiased – when conditioning on the selections of arms – requires to consider a delay-corrected count $\tilde{N}(t)$ taking into account the probability of having eventually observed the reward

associated with each previous pull of k. We distinguish the expression of $\tilde{N}(t)$ according to whether feedback is censored or not.

Censored model. When rewards cannot be disclosed after m rounds following the action, the current available information on the pulls is split into two main groups: The 'oldest' pulls, censored if not observed yet, and the most recent ones. Namely, we now define $\tilde{N}_k(t)$ as

$$\tilde{N}_k(t) = \sum_{s=1}^{t-m} \mathbb{1}\{A_s = k\}\tau_m + \sum_{s=t-m+1}^{t-1} \mathbb{1}\{A_s = k\}\tau_{t-s}.$$

Overall, the conversion rate estimator is defined as

$$\hat{\theta}_k(t) = \frac{S_k(t)}{\tilde{N}_k(t)}. \tag{5}$$

Remark 5. *In the uncensored case, defining* $\tilde{N}_k(t) := \sum_{s=1}^{t} \mathbb{1}\{A_s = k\}\tau_{t-s}$. *leads to an equivalent definition of* $\hat{\theta}_k(t)$ *as a conditionally unbiased estimator.*

5.2 UCB INDEX

We first define a delay-corrected UCB-index for bounded rewards.

Concentration bound. Using the self-normalized concentration inequality of Proposition 8 of [14], yields the following result, that we recall here for completeness.

Proposition 6. *Let* k *be an arm in* $\{1, ..., K\}$*, then for any* $\beta > 0$ *and for all* $t > 0$*,*

$$\mathbb{P}\left(\theta_k > \hat{\theta}_k(t) + \sqrt{\frac{N_k(t)}{\tilde{N}_k(t)}}\sqrt{\frac{\beta}{2\tilde{N}_k(t)}}\right) < \beta e \log(t) e^{-\beta}.$$

Upper-confidence Bound. Thus, an UCB index for $\hat{\theta}_k(t)$ may be defined as

$$U_k^{\mathrm{UCB}}(t) = \hat{\theta}_k(t) + \sqrt{\frac{N_k(t)}{\tilde{N}_k(t)}}\sqrt{\frac{\beta_\epsilon(t)}{2\tilde{N}_k(t)}},$$

where $\beta\epsilon(t)$ is a suitable slowly growing exploration function (see below). This upper confidence interval is scaled by $N_k(t)/\tilde{N}_k(t)$ when compared to the classical UCB index. This ratio gets bigger when the (τ_d)'s are small for large delays d, that is when the median delay is large: The longer we need to wait for observations to come, the largest our uncertainty about our current cumulated reward.

5.3 KLUCB INDEX

Concentration bound. We first state a concentration inequality that controls the underestimation probability based on an alternative Chernoff bound for a sum of independent binary random variables (Lemma 13 proved in Appendix A).

This lemma only holds for a sequence of pulls fixed before-hand, independently of realizations, i.e., the values of A_t do not depend on the sequence of X_s. Although with a restrictive scope, it provides intuition on the construction of the algorithm.

Lemma 7. *Assume that the sequence of pulls is fixed beforehand and let* k *be an arm in* $\{1, ..., K\}$*. Then for any* $\delta > 0$ *and for all* $t > 0$*,*

$$\mathbb{P}\left(\left\{\hat{\theta}_k(t) < \theta_k\right\} \cap \left\{\tilde{N}_k(t) d_{\mathrm{Pois}}(\hat{\theta}_k(t), \theta_k) > \delta\right\}\right) < e^{-\delta}.$$

where $d_{\mathrm{Pois}}(p, q) = p \log p/q + q - p$ *denotes the Poisson Kullback-Leibler divergence.*

To get upper confidence bounds for θ_k from Lemma 7, we follow [5] and define the KL-UCB index by

$$U_k^{\mathrm{KL}}(t) = \max\Big\{q \in [\hat{\theta}_k(t), 1] : \tilde{N}_k(t) d_{\mathrm{Pois}}(\hat{\theta}_k(t), q) \le \beta_\epsilon(t)\Big\}.$$

Using $\beta_\epsilon(t) = \beta$, this KL-UCB index satisfies a result analogous to Proposition 6 (see Proposition 14 in Appendix A.1):

$$\mathbb{P}\left(\theta_k > U_k^{\mathrm{KL}}(t)\right) \le e\lceil \beta \log(t) \rceil e^{-\beta}.$$

Even though the Kullback-Leibler divergence does not have the same expression for Bernoulli and Poisson random variables, the following lemma (proved in Appendix A.1) shows that for a certain range or parameters they are actually very close.

Lemma 8. *For* $0 < p < q < 1$*,*

$$(1 - q) d(p, q) \le d_{\mathrm{Pois}}(p, q) \le d(p, q).$$

6 ALGORITHMS

Algorithm 1 present the scheme common to both the censored and uncensored cases, which differ only by the definition of the parameter estimator. In both cases, one may also consider either of the two UCB or KL-UCB index defined in the previous section, resulting in the DelayedUCB and DelayedKLUCB algorithms. We provide a finite-time analysis of the regret of these algorithms, when using an exploration function of the form $\beta_\epsilon(t) = (1 + \epsilon) \log(t)$, for some positive ϵ.

Algorithm 1 – DelayedUCB and DelayedKLUCB.

Require: K, CDF parameters $(\tau_d)_{d\geq 0}$, threshold $m > 0$ if feedback is censored.
Initialization: First K rounds, play each arm once.
for $t > K$ **do**
Compute $S_k(t)$ and $\tilde{N}_k(t)$ for all k according to the assumed feedback model (censored or not),
Compute $\hat{\theta}_k(t)$ for al k,
For a given choice of algorithm $\mathcal{A} \in \{\text{KLUCB}, \text{UCB}\}$,
$A_t \leftarrow \arg\max_k U_k^{\mathcal{A}}(t)$.
Observe reward Y_t and all individual feedback $(X_s, t)_{s\leq t}$
end for

Finite-time Analysis of DelayedUCB.

Theorem 9. *In the censored setting, the regret of* D*elayedUCB is bounded from above by*

$$L_{\text{UCB}}(T) \leq (1+\epsilon)\log(T)\sum_{k\neq *}\frac{1}{2\tau_m\Delta_k} + o_{\epsilon,m}(\log(T)).$$

Proof Outline of the proof (cf Appendix C.1):

1. First upper-bound the regret using Lemma 1 in the uncensored case:
$$R(T) \leq \sum_{k>1}\Delta_k\mathbb{E}[N_k(T)],$$
and bounding the first m losses by 1 in the censored case:
$$R(T) \leq m + \sum_{k>1}\tau_m\Delta_k\mathbb{E}\left[\sum_{t>m}^{T}\mathbb{1}\{A_t = k\}\right].$$
2. Then, decompose the event $\mathbb{1}\{A_t = k\}$ as in [1]
$$\sum_{t>m}^{T}\mathbb{1}\{A_t = k\} \leq \sum_{t>m}^{T}\mathbb{1}\left\{U_1^{\text{UCB}}(t) < \theta_1\right\} + \sum_{t>m}^{T}\mathbb{1}\left\{A_{t+1} = k, U_k^{\text{UCB}}(t) \geq \theta_1\right\}.$$
3. Remark that the first sum is handled by Proposition 6 so it suffices to control the second sum.
$$\mathbb{E}\left[\mathbb{1}\left\{A_{t+1} = k, U_k^{\text{UCB}}(t) \geq \theta_1\right\}\right] \leq \frac{(1+\epsilon)\log(T)}{2\tau_m^2\Delta_i^2} + \sum_{s>\frac{(1+\epsilon)\log(T)}{2\Delta_i^2}}\mathbb{P}\left(U_k^{\text{UCB}}(t) \geq \theta_i + \Delta_i\right).$$

The last term is actually $O(\sqrt{\log(T)})$, giving the desired result. Details, as well as explicit constants and dependencies can be found in Appendix C.1. ■

Corollary 10. *In the uncensored setting, we also assume that there exists $c > 0$ such that $1 - \tau_m \leq \frac{c}{m}$ for all $m \geq 1$. Then, is bounded from above by*

$$L_{\text{UCB}}(T) \leq \frac{1+\epsilon}{1-\epsilon}\log(T)\sum_{k>1}\frac{1}{2\Delta_k} + o_{\epsilon,m}(\log(T)).$$

Proof The analysis of DelayedUCB given in Appendix C.1 (in the censored setting) shows that the performances of DelayedUCB in the uncensored setting can be upper-bounded by its performances in the censored setting, where the threshold m can be arbitrarily fixed to some value. Choosing m will only have an impact on the analysis of the algorithm. The specific choice of m satisfying $\tau_m \geq 1 - \epsilon$ gives the claimed result. As indicated in Appendix C.1, the dependency of $o_{\epsilon,m}(\log(T))$ is actually only linear in m. As a consequence, along with the assumption on the decay of $1 - \tau_m$, this yields that the overall dependency in the parameter m is reduced to $1/\epsilon$. ■

We emphasize that the assumption that $1 - \tau_m \leq 1/m$, is actually rather natural. Indeed, if $1 - \tau_m \leq c/m^\gamma$, for some constants $c, \gamma > 0$, then the finiteness requirement on the expected delay is satisfied if and only if $\gamma > 1$.

Finite-time Analysis of DelayedKLUCB.

Theorem 11. *For any $\eta > 0$, the regret of* D*elayedKLUCB is bounded in the censored setting as*

$$L_{\text{KLUCB}}(T) \leq (1+\eta)\frac{\beta_\epsilon(t)}{1-\theta_1}\sum_{k>1}\frac{\tau_m\Delta_k}{d(\tau_m\theta_k, \tau_m\theta_1)} + o_{\epsilon,m,\eta}(\log(T)).$$

Proof Outline of the proof (cf. AppendixC.2):

1. We start by decomposing the regret according to the different types of unfavorable events. Note that thanks to the upper bound on the regret provided by Lemma 1, we need to control on the number of suboptimal pulls $\mathbb{E}[N_k(T)]$ for arms $k > 1$.
$$\mathbb{E}[N_k(T)] \leq m + \mathbb{E}\left[\sum_{t=m+1}^{T}\mathbb{1}\{U_1(t) < \theta_1\}\right] + \mathbb{E}\left[\sum_{t=m+1}^{T}\mathbb{1}\{A(t) = k, U_k(t) \geq \theta_1\}\right].$$
2. The first sum is handled by Theorem 14 in Appendix A which shows that it is $o(\log(T))$. For the second term, we bound the indices using the fact

that $\tilde{N}_k(t) \geq \tau_m N_k(t-m)$ to obtain

$$U_k^{\mathrm{KL}}(t) \leq U_k^{\mathrm{KL}+}(t)$$
$$:= \underset{q \in [\hat{\theta}_k, 1]}{\arg\max} \big\{ q | \tau_m d_{\mathrm{Pois}}(\hat{\theta}_k(t), q) \leq \frac{\beta_\epsilon(t)}{N_k(t-m)} \big\}.$$

Notice that the $U_k^{\mathrm{KL}+}(t)$ indices are well defined for $t > m$.

3. Then, we proceed as in the proof of Theorem 10 in Appendix B.2 of [9]. For any $\eta > 0$, we define the characteristic number of pulls

$$K_k(T) = \frac{(1+\eta)\beta_\epsilon(t)}{d_{\mathrm{Pois}}(\tau_m \theta_k, \tau_m \theta_1)},$$

and we prove

$$\sum_{s \geq K_k(T)} \mathbb{P}\left(\tau_m s d_{\mathrm{Pois}}(\hat{\theta}_{k,s}, \theta_1) \leq \beta_\epsilon(t) \right) = o_{\epsilon,m,\eta}(\log T)$$

using Fact 2 of [2] for exponential families.

■

Corollary 12. *In the uncensored setting, under the same hypothesis than in Corollary 10, namely that there exists a constant c such that $1 - \tau_m \leq \frac{c}{m}$ for all $m \leq 1$. Then, the regret of* D*elayed**KLUCB is bounded from above as*

$$L_{\mathrm{KLUCB}}(T) \leq \frac{\beta_\epsilon(t)}{1-\theta_1} \sum_{k>1} \frac{(1+\eta)(1-\epsilon)\Delta_k}{d((1-\epsilon)\theta_k, (1-\epsilon)\theta_1)} + o_{\eta,\epsilon}(\log(T)).$$

Proof As for the proof of Corollary 10, the performance of DelayedKLUCB in the uncensored case can be bounded as in the censored case by a specific choice of $m(\epsilon)$ such that $\tau_m \geq 1 - \epsilon$, namely $m(\epsilon) \geq c/\epsilon$. As shown in the proof of Theorem 11 in the censored case, the dependency in m of the term of rest is linear, reducing to $1/\epsilon$. ■

Naive benchmark: The DISCARDING policy. An obvious benchmark algorithm in the censored setting is to use the regular UCB and KLUCB policies only using the first $t - m$ pulls and observed rewards at each time t. In that case the empirical average considered is simply $\hat{\theta}_k^m(t) = S_k(t-m)/\tau_m N_k(t-m)$ and the corresponding optimistic indices are

$$U^m(t) = \hat{\theta}_k^m(t) + \sqrt{\beta_\epsilon(t)/2\tau_m N_k(t-m)},$$
$$U_k^{m|\mathrm{KL}}(t) = \max_{q \in [\hat{\theta}_k^m, 1]} \Big\{ q | \, \tau_m d_{\mathrm{Pois}}(\hat{\theta}_k^m, q) \leq \frac{\beta_\epsilon(t)}{N(t-m)} \Big\}.$$

These indices can only be computed after at least m rounds. The proof technique used for the analysis of our algorithms in the censored case actually shows that the DISCARDINGUCB and DISCARDINGKLUCB policies are asymptotically optimal. Nonetheless, in practice it is very undesirable to have an arbitrarily long linear regret phase at the beginning of the learning until the threshold m is reached. This is especially true if the threshold m is large as compared to the horizon T. In that case, we empirically show in Section 7 that our algorithms achieve drastically improved short-horizon performance.

7 EXPERIMENTS

In this section we perform simulations in our two delayed feedback frameworks. The algorithms described in the previous section will be denoted D-UCB and D-KLUCB in the censored setting, and UD-UCB and UD-KLUCB in the uncensored setting.

As a matter of fact, the bottleneck of such policies is to compute $\tilde{N}(t)$ which is theoretically a weighted sum over all past actions and, without any assumption on the weights $(\tau_s)_{s \geq 0}$, it requires to store all previous rewards and recompute $\tilde{N}(t)$ at each iteration.

Following the conclusions of [4], we assume all along this section that the delays follow a geometric distribution with parameter $\lambda := 1/\mu$. This assumption allows us to implement our algorithms in a computationally, memory-efficient manner. Indeed, for each $s \geq 0$, we now have $(1 - \tau_{s+1}) = \lambda(1 - \tau_s)$ and this remark provides a sequential updating scheme of the quantity $\tilde{N}_k(t)$ for $k \in [K]$. In the uncensored setting, we have:

$$\tilde{N}_k(t) = \sum_{s=1}^{t} (1 - \lambda^{t-s+1}) \mathbb{1}\{A_s = k\} = N_k(t) - O_k(t),$$

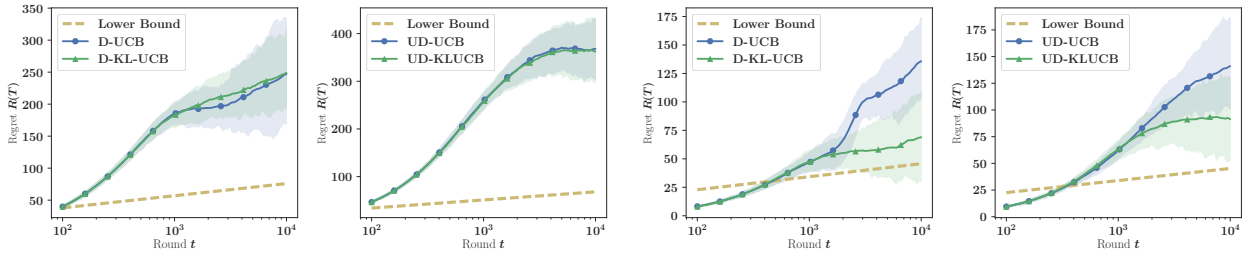
where $O_k(t)$ is updated after each round as follows

$$O_k(t+1) \leftarrow \lambda O_k(t) + \mathbb{1}\{A_t = k\}. \tag{6}$$

In the censored setting, however, one must still keep track of some of the previous pulls in order to compute

$$\tilde{N}_k(t) = N_k(t-m)\tau_m + \sum_{s=t-m+1}^{t-1} \mathbb{1}\{A_s = k\} \tau_{t-s}.$$

In practice this can be done by maintaining a buffer of size m containing the last m pulls that are multiplied by the probability of observing a reward with the delay corresponding to their current position in the buffer. In addition to this buffer, we add old pulls in a separate count $N_k(t-m)$ for which the weight will stay τ_m.



(a) $\theta_H = (0.5, 0.4, 0.3)$. (b) $\theta_L = (0.1, 0.05, 0.03)$.

Figure 1: Expected regret of D-UCB and D-KLUCB (censored setting), and UD-UCB and UD-KLUCB (uncensored setting) for two bandit problems: $\theta_H = (0.5, 0.4, 0.3)$, $\theta_L = (0.1, 0.05, 0.03)$. For all experiments, $T = 10000$, $\mu = 500$, $m = 1000$ (if censored) and the results are averaged over 100 runs.

Comparing DelayedUCB and DelayedKLUCB. We compare the regret of both delayed bandits policies in the censored and uncensored setting for $T = 10000$, $\mu = 500$ and $m = 1000$.

Simulations on Figure 1, for two problems, $\theta_H = (0.5, 0.4, 0.3)$ on the left, and $\theta_L = (0.1, 0.05, 0.03)$ on the right, display the classical pattern that while UCB-based algorithms perform satisfactorily for central values (close to 0.5) of the conversion rate, they are clearly sub-optimal with more realistic values for the conversion rate. The two right plots also confirm that, for the KLUCB-based algorithms, the loss with respect to the optimal regret growth rate due to the use of the Poisson divergence is – as expected from Theorem 11 – not significant for low values (here $\theta^* = 0.1$) of the conversion rates.

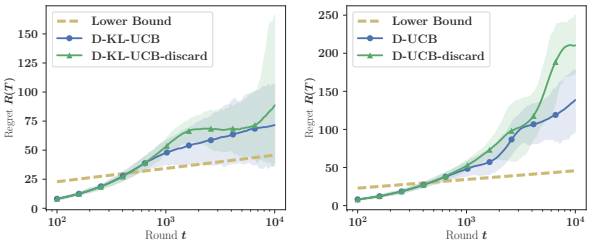


Figure 2: Expected regret of D-UCB and D-KLUCB in the censored setting vs the equivalent discarding policies when $\mu = 500$ and $m = 1000$. Results are averaged over 200 independent runs.

DelayedUCB and DelayedKLUCB vs. DISCARDING. In this section, we illustrate the good empirical initial performance of DelayedUCB and DelayedKLUCB, when compared to the heuristic DISCARDING approach presented in Section 6.

Figure 2 compares results for both DelayedUCB and DelayedKLUCB with $\theta = (0.1, 0.05, 0.03)$, $T = 10000$, $\mu = 500$ and $m = 1000$ in the censored setting. We observe that discarding policies incur a linear regret phase at the beginning of the learning and happen to catch up with the expected regret growth rate only after a large number of rounds. These figures reveal a non-negligible gap in performance between the naive DISCARDING approach and our delay-adapted quasi-optimal algorithms.

8 CONCLUSION

The stochastic delayed bandit setting introduced in this work addresses an important problem in many applications where the feedback associated to each action is delayed and censored, due to the ambiguity between conversions that will never happen and conversions that will occur at some later – perhaps unobservable– time. Under the hypothesis that the distribution of the delay is known, we provided a complete analysis of this model as well as simple and efficient algorithms. An interesting generalization of the present work would be to relax the model hypothesis and estimate the delay distribution on-the-go, possibly using context-dependent delay distributions.

References

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[2] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

[3] Nicolo Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *29th Annual Conference on Learning Theory*, pages 605–622, 2016.

[4] Olivier Chapelle. Modeling delayed feedback in

display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1097–1105. ACM, 2014.

[5] Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *COLT*, pages 359–376, 2011.

[6] Aurélien Garivier, Pierre Menard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems (*to appear*). *Mathematics of Operations Research*, 2017.

[7] Wendi Ji, Xiaoling Wang, and Dell Zhang. A probabilistic multi-touch attribution model for online advertising. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1373–1382. ACM, 2016.

[8] Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. *CoRR*, abs/1306.0686, 2013.

[9] Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1453–1461, 2013.

[10] Kwang-Sung Jun, Kevin Jamieson, Robert Nowak, and Xiaojin Zhu. Top arm identification in multi-armed bandits with batch arm pulls. In *The 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[11] Sumeet Katariya, Branislav Kveton, Csaba Szepesvari, Claire Vernade, and Zheng Wen. Stochastic rank-1 bandits. *arXiv preprint arXiv:1608.03023*, 2016.

[12] Konstantinos V Katsikopoulos and Sascha E Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE transactions on automatic control*, 48(4):568–574, 2003.

[13] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 2015.

[14] Paul Lagrée, Claire Vernade, and Olivier Cappé. Multiple-play bandits in the position-based model. In *Advances in Neural Information Processing Systems*, 2016.

[15] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[16] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[17] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Eric Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016.

[18] Rómer Rosales, Haibin Cheng, and Eren Manavoglu. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 293–302. ACM, 2012.

[19] Thomas J Walsh, Ali Nouri, Lihong Li, and Michael L Littman. Learning and planning in environments with delayed feedback. *Autonomous Agents and Multi-Agent Systems*, 18(1):83–105, 2009.