Stochastic Multi-Armed Bandits with Unrestricted Delay Distributions

Tal Lancewicki [*1] **Shahar Segal** [*1] **Tomer Koren** [1 2] **Yishay Mansour** [1 2]

Abstract

We study the stochastic Multi-Armed Bandit (MAB) problem with random delays in the feedback received by the algorithm. We consider two settings: the *reward-dependent* delay setting, where realized delays may depend on the stochastic rewards, and the *reward-independent* delay setting. Our main contribution is algorithms that achieve near-optimal regret in each of the settings, with an additional additive dependence on the quantiles of the delay distribution. Our results do not make any assumptions on the delay distributions: in particular, we do not assume they come from any parametric family of distributions and allow for unbounded support and expectation; we further allow for infinite delays where the algorithm might occasionally not observe any feedback.

1 Introduction

Stochastic Multi-armed Bandit problem (MAB) is a theoretical framework for studying sequential decision making. Most of the literature on MAB assumes that the agent observes feedback immediately after taking an action. However, in many real world applications, the feedback might be available only after a period of time. For instance, in clinical trials, the observed effect of a medical treatment often comes in delay, that may vary between different treatments. Another example is in targeted advertising on the web: when a user clicks a display ad the feedback is immediate, but if a user decides *not* to click, then the algorithm will become aware to that only when the user left the website or enough time has elapsed.

In this paper, we study the stochastic MAB problem with randomized delays (Joulani et al., 2013). The reward of the chosen action at time t is sampled from some distribution, like in the classic stochastic MAB problem. However, the reward is *observed* only at time $t + d_t$, where d_t is a random variable denoting the delay at step t. This problem has been studied extensively in the literature (Joulani et al., 2013; Vernade et al., 2017; Pike-Burke et al., 2018; Gael et al., 2020) under an implicit assumption that the delays are *reward-independent*: namely, that d_t is sampled from an unknown delay distribution and may depend on the chosen arm, but *not* on the stochastic rewards on the same round. For example, Joulani et al. (2013); Pike-Burke et al. (2018) show a regret bound of the form $O(\mathcal{R}_T^{MAB} + K\mathbb{E}[D])$. Here $\mathcal{R}_T^{MAB}$ denotes the optimal instance-dependent T-round regret bound for standard (non-delayed) MAB: $\mathcal{R}_T^{MAB} = \sum_{\Delta_i>0} \log(T)/\Delta_i$, where Δ_i is the sub-optimality gap for arm i. In the second term, K is the number of arms and $\mathbb{E}[D]$ is the expected delay.

A significantly more challenging setting, that to the best of our knowledge was not explicitly addressed previously in the literature,[1] is that of *reward-dependent* delays. In this setting, the random delay at each round may also depend on the reward received on the same round (in other words, they are drawn together from a *joint distribution* over rewards and delays). This scenario is motivated by both of the examples mentioned earlier: e.g., in targeted advertisement the delay associated with a certain user is strongly correlated with the reward she generates (i.e., click or no click); and in clinical trials, the delay often depends on the effect of the applied treatment as some side-effects take longer than others to surface.

In contrast to the reward-independent case, with reward-dependent delays the observed feedback might give a biased impression of the true rewards. Namely, the expectation of the *observed* reward can be very different than the actual expected reward. For example, consider Bernoulli rewards. If the delays given reward 0 are shorter than the delays given reward 1, then the observed reward will be biased towards 0. Even worse, the direction of the bias can be opposite between different arms. Hence, as long as the fraction of *unobserved* feedback is significant, the expected *observed* reward of the optimal arm can be smaller than expected

[*]Equal contribution [1]Blavatnik School of Computer Science, Tel Aviv University, Israel [2]Google Research, Tel Aviv. Correspondence to: Tal Lancewicki <lancewicki@mail.tau.ac.il>, Shahar Segal <shaharsegal1@mail.tau.ac.il>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]Some of the results of Vernade et al. (2017); Gael et al. (2020) can be viewed as having a specific form of reward-dependent delays; we discuss this in more detail in the related work section.

observed reward of a sub-optimal arm, which makes the learning task substantially more challenging.

1.1 Our contributions

We consider both the reward-independent and reward-dependent versions of stochastic MAB with delays. In the reward-independent case we give new algorithms whose regret bounds significantly improve upon the state-of-the-art, and also give instance-dependent lower bounds demonstrating that our algorithms are nearly-optimal. In the reward-dependent setting, we give the first algorithm to handle such delay structure and the potential bias in the observed feedback that it induces. We provide both an upper bound on the regret and a nearly matching general lower bound.

Reward-independent delays: We first consider the easier *reward-independent* case. In this case, we provide an algorithm where the second term scales with a quantile of the delay distribution rather the expectation, and the regret is bounded by $O(\min_q\{\mathcal{R}_T^{MAB}/q + d(q)\})$, where $d(q)$ is the q-quantile of the delay distribution. Specifically, when choosing the median (i.e., $q = 1/2$), we obtain regret bound of $O(\mathcal{R}_T^{MAB} + d(1/2))$. We thus improve over the $O(\mathcal{R}_T^{MAB} + K\mathbb{E}[D])$ regret bound of Joulani et al. (2013); Pike-Burke et al. (2018), as the median is always smaller than the expected delay, up to factor of two. Moreover, the increase in regret due to delays in our bound does not scale with number of arms, so the improvement is significant even with fixed delays (Dudik et al., 2011; Joulani et al., 2013). Our bound is achieved using a remarkably simple algorithm, based on variant of Successive Elimination (Even-Dar et al., 2006). For this algorithm, we also prove a more delicate regret bound for arm-dependent delays that allows for choosing different quantiles q_i for different arms i (rather than a single quantile q for all arms simultaneously).

The intuition why the increase in regret due to delays should scale with a certain quantile is fairly straightforward: consider for instance the median of the delay, d_M. For simplicity, assume that the delay value is known when we take the action. One can simulate a black box algorithm for delays that are bounded by d_M on the rounds in which delay is smaller than d_M (approximately half of the rounds), and in the rest of the rounds, imitate the last action of the black-box algorithm. Since rewards are stochastic, and independent of time and the delay, the regret on rounds with delay larger than d_M is similar to the regret of the black-box algorithm on the rest of the rounds, resulting with total regret of twice the regret of the black-box algorithm. For example, when using the algorithm of (Joulani et al., 2013), this would give us $O(\mathcal{R}_T^{MAB} + Kd_M)$. We stress that unlike this reduction, our algorithm does not need to know the value of the delay at any time, nor the median or any other quantile. In addition, our bound is much stronger and does not depend on K on the second term.

Table 1. Regret bounds comparison of this and previous works. The bounds in this table omit constant and $\log(K)$ factors.

	Previous work	This paper
General, Reward-indep.	$\mathcal{R}_T^{MAB} + \sum_i \mathbb{E}[G^*_{T,i}]$ $\mathcal{R}_T^{MAB} + K\mathbb{E}[D]$ (Joulani et al., 2013)	$\min_q\{\frac{1}{q}\mathcal{R}_T^{MAB} + d(q)\}$
Fixed delay d	$\mathcal{R}_T^{MAB} + Kd$ (Joulani et al., 2013) $\sqrt{TK} + \sqrt{K}d$ (Dudik et al., 2011)	$\mathcal{R}_T^{MAB} + d$
α-Pareto	$\mathcal{R}_T^{MAB} + \sum_i \left(\frac{8}{\Delta_i}\right)^{\frac{1-\alpha}{\alpha}}$ (Gael et al., 2020)	$\mathcal{R}_T^{MAB} + 2^{1/\alpha}$
Packet loss	$(1-p)T$ (Joulani et al., 2013)	$\frac{1}{p}\mathcal{R}_T^{MAB}$
General, Reward-dep.	—	$\mathcal{R}_T^{MAB} + d(1-\Delta_{min})$

Reward-dependent delays: We then proceed to consider the more challenging *reward-dependent* setting. In this setting, the feedback reveals much less information on the true rewards due to the selection bias in the *observed* rewards (in other words, the distributions of the observed feedback and the unobserved feedback might be very different). In order to deal with this uncertainty, we present another algorithm, also inspired by Successive Elimination. The algorithm widens the confidence bounds in order to handle the potential bias. We achieve a regret bound of the form $O(\mathcal{R}_T^{MAB} + \log(K)d(1-\Delta_{min}/4))$, where Δ_{min} is the minimal sub-optimality gap, and $d(\cdot)$ is the quantile function of the marginal delay distribution. We show that this bound is optimal, by presenting a matching lower bound, up to a factor of Δ in the second term (and $\log(K)$ factors).

Summary and comparison of bounds: Our main results, along with a concise comparison to previous work, are presented in Table 1. $G^*_{T,i}$ denotes the maximal number of unobserved feedback from arm i. The results show that our algorithm works well even under heavy-tailed distributions and some distributions with infinite expected value. For example, the arm-dependent delay distributions used by Gael et al. (2020) are all bounded by an α-pareto distribution (in terms of the delay distributions CDFs). Hence, their median is bounded by $2^{1/\alpha}$. Our algorithm suffer at most an additional $O(2^{1/\alpha})$ to the classical regret for MAB without delays (see bounds for the α-Pareto case in Table 1). In the "packet loss" setting, the delay is 0 with probability p, and ∞ (or T) otherwise. If p is a constant (e.g., $> 1/4$), our regret bound scales as the optimal regret bound for MAB without delays, up to constant factors. Previous work Joulani et al. (2013) show a regret bound which scales with the number of missing samples, and thus is linear. A Pareto distribution that will bound such delay would require a very small parameter α which also result in linear regret bound by the result of Gael et al. (2020).

1.2 Related work

To the best of our knowledge, Dudik et al. (2011) were the first to consider delays in stochastic MAB. They examine contextual bandit with fixed delay d, and obtain regret bound of $O(\sqrt{K\log(NT)}(d+\sqrt{T}))$, where N is number of possible policies. Joulani et al. (2013) use a reduction to non-delayed MAB. For their explicit bound they assume that expected value of the delay is bounded (see Table 1 for their *implicit* bound). Pike-Burke et al. (2018) consider a more challenging setting in which the learner observe the *sum* of rewards that arrive at the same round. They assume that the expected delay is known, and obtain similar bound as Joulani et al. (2013).

Vernade et al. (2017) study *partially observed feedback* where the learner cannot distinguish between reward of 0 and a feedback that have not returned yet, which is a special form of reward-dependent delay. However, they assume bounded expected delay and full knowledge on the delay distribution. Gael et al. (2020) also consider partially observed feedback, and aim to relax the bounded expected delay assumption. They consider delay distributions that their CDF are bounded from below by the CDF of an α-Pareto distribution, which might have infinite expected delay for $\alpha \leq 1$. However, this assumption still limits the distribution, e.g., the commonly examined fixed delay falls outside their setting. Moreover, they assume that the parameter α is known to the learner. Other extensions include Gaussian Process Bandit Optimization (Desautels et al., 2014) and linear contextual bandits (Zhou et al., 2019). As opposed to most of these works, we place no assumptions on the delay distribution, and the learner has no prior knowledge on it.

Delays were also studied in the context of the non-stochastic MAB problem (Auer et al., 2002b). Generally, when reward are chosen in an adversarial fashion, the regret increases by a multiplicative factor of the delay. Under full information, Weinberger & Ordentlich (2002) show regret bound of $O(\sqrt{dT})$, with fixed delay d. This was extended to bandit feedback by (Cesa-Bianchi et al., 2019), with near-optimal regret bound of $O(\sqrt{T(K+d)})$. Several works have studied the effect of adversarial delays, in which the regret scales with $O(\sqrt{T}+\sqrt{D})$, where D is the sum of delays (Thune et al., 2019; Bistritz et al., 2019; Zimmert & Seldin, 2020; György & Joulani, 2020). For last, Cesa-Bianchi et al. (2018) consider a similar setting to Pike-Burke et al. (2018), in which the learner observe only the sum of rewards. The increase in the regret is by a multiplicative factor of $\sqrt{d}$.

2 Problem Setup and Background

We consider a variant of the classical stochastic Multi-armed Bandit (MAB) problem. In each round $t = 1, 2, \ldots, T$, an agent chooses an arm $a_t \in [K]$ and gets reward $r_t(a_t)$, where $r_t(\cdot) \in [0,1]^K$ is a random vector. Unlike the standard MAB setting, the agent does not immediately observe $r_t(a_t)$ at the end of round t; rather, only after $d_t(a_t)$ rounds (namely, at the end of round $t+d_t(a_t)$) the tuple $(a_t, r_t(a_t))$ is received as feedback. We stress that neither the delay $d_t(a_t)$ nor the round number t are observed as part of the feedback (so that the delay cannot be deduced directly from the feedback). The delay is supported in $\mathbb{N} \cup \{\infty\}$. In particular, we allow $d_t(a_t)$ to be infinite, in which case the associated reward is never observed. The pairs of vectors $\{(r_t(\cdot), d_t(\cdot))\}_{t=1}^T$ are sampled i.i.d from a *joint* distribution. Throughout the paper we sometimes abuse notation and denote $r_t(a_t)$ and $d_t(a_t)$ simply by r_t and d_t, respectively. This protocol is summarized in Protocol 1.

Protocol 1 MAB with stochastic delays

for $t \in [T]$ **do**
- Agent picks an action $a_t \in [K]$.
- Environment samples a pair, $(r_t(\cdot), d_t(\cdot))$, from a joint distribution.
- Agent get a reward $r_t(a_t)$ and observes feedback $\{(a_s, r_s(a_s)) : t = s + d_s(a_s)\}$.

We discuss two forms of stochastic delays: *(i)* reward-independent delays, where the vectors $r_t(\cdot)$ and $d_t(\cdot)$ are *independent* from each other, and *(ii)* reward-dependent delays, where there is no restriction on the joint distribution.

The performance of the agent is measured as usual by the the difference between the algorithm's cumulative expected reward and the best possible total expected reward of any fixed arm. This is known as the *expected pseudo regret*, formally defined by

$$\begin{aligned}\mathcal{R}_T &= \max_i \mathbb{E}\left[\sum_{t=1}^T r_t(i)\right] - \mathbb{E}\left[\sum_{t=1}^T r_t(a_t)\right] \\ &= T\mu_{i^*} - \mathbb{E}\left[\sum_{t=1}^T \mu_{a_t}\right] = \mathbb{E}\left[\sum_{t=1}^T \Delta_{a_t}\right],\end{aligned}$$

where μ_i is the mean reward of arm i, i^* denotes the optimal arm and $\Delta_i = \mu_{i^*} - \mu_i$ for all $i \in [K]$.

For a fixed algorithm for the agent (the relevant algorithm will always be clear from the context), we denote by $m_t(i)$ the number of times it choose arm i by the end of round $t-1$. Similarly $n_t(i)$ denotes the number of *observed* feedback from arm i, by the end of round $t-1$. The two might differ as some of the feedback is delayed. Let $\hat{\mu}_t(i)$ be the *observed* empirical average of arm i, defined as:

$$\hat{\mu}_t(i) = \frac{1}{n_t(i) \vee 1} \sum_{s: s+d_s<t} \mathbb{I}\{a_s = i\} r_s,$$

where $a \vee b = \max\{a, b\}$ and $\mathbb{I}\{\pi\}$ is the indicator function of predicate π.

We denote $d_i(q)$ to be the quantile function for arm i's delay distribution; formally, if D_i is the delay of arm i then the quantile function is defined as

$$d_i(q) = \min\left\{\gamma \in \mathbb{N} \mid \Pr[D_i \leq \gamma] \geq q\right\}.$$

3 Reward-independent Delays

We first consider the case where delays are independent of the realized stochastic rewards. We begin with an analysis of two classic algorithms: UCB (Auer et al., 2002a) and Successive Elimination (SE) (Even-Dar et al., 2006), adjusted to handle delayed feedback in a straightforward naive way (see Procedure 2).

3.1 Suboptimality of UCB with delays

UCB (Auer et al., 2002a) is based on the basic principle of "optimism under uncertainty." It maintains for each arm an upper confidence bound (UCB): a value that upper bounds the true mean with high probability. In each round it simply pulls the arm with the highest UCB. The exact description appears in Algorithm 3.

Procedure 2 Update-Parameters

for $i \in [K]$ **do**
  # number of observed feedback
  $n_t(i) \leftarrow \sum_{s:s+d_s<t} \mathbb{I}\{a_s = i\}$
  # observed empirical mean
  $\hat{\mu}_t(i) \leftarrow \frac{1}{n_t(i)\vee 1} \sum_{s:s+d_s<t} \mathbb{I}\{a_s = i\} r_s$
  $LCB_t(i) \leftarrow \hat{\mu}_t(i) - \sqrt{\frac{2\log T}{n_t(i)\vee 1}}$
  $UCB_t(i) \leftarrow \hat{\mu}_t(i) + \sqrt{\frac{2\log T}{n_t(i)\vee 1}}$

Algorithm 3 UCB with Delays

Input: number of rounds T, number of arms K
Initialization: $t \leftarrow 1$
begin with sampling each arm once
Pull each arm $i \in [K]$ once
Observe any incoming feedback
Set $t \leftarrow t + K$
while $t < T$ **do**
  Call *Update-Parameters* (Procedure 2)
  Pull arm $a_t \in \arg\max_i UCB_t(i)$
  # With deterministic tie breaking rule i.e. by index
  Observe feedback $\{(a_s, r_s) : s + d_s = t\}$
  Set $t \leftarrow t + 1$

In the standard non-delayed setting, UCB is known to be optimal. However, with delays this is no longer the case. Consider the simpler case where all arms suffers from a constant fixed delay d. Joulani et al. (2013) show that the regret of UCB with delay is bounded by $O(\mathcal{R}_T^{MAB} + Kd)$. We show that the increase in the regret is necessary for UCB, and the additional regret due to the delay can in general scale as $\Omega(Kd)$. The reason is due to the nature of UCB: it always samples the currently most promising arm, and it might take as much as d rounds to update the latter. This is formalized in the following theorem (proof is deferred to the full version of the paper (Lancewicki et al., 2021).)

Theorem 1. *Under fixed delay $d \geq K$, there exist a problem instance such that UCB suffers regret of $\Omega(Kd)$.*

3.2 Successive Elimination with delays

Successive Elimination (SE) maintains a set of active arms, where initially all arms are active. It pulls all arms equally and whenever there is a high-confidence that an arm is suboptimal, it eliminates it from the set of active arms. The exact description appears in Algorithm 4.

Algorithm 4 Successive Elimination with Delays

Input: number of rounds T, number of arms K
Initialization: $S \leftarrow [K]$, $t \leftarrow 1$
while $t < T$ **do**
  Pull each arm $i \in S$
  Observe any incoming feedback
  Set $t \leftarrow t + |S|$
  Call *Update-Parameters* (Procedure 2)
  # Elimination Step
  Remove from S all arms i such that exists j with $UCB_t(i) < LCB_t(j)$

Unlike UCB, SE continues to sample all arms equally, and not just the most promising arm. In fact, the number of rounds that SE runs before it observes m samples for K arms is approximately $Km+d$, whereas UCB might require $K(m+d)$ rounds in certain cases. More generally, we prove:

Theorem 2. *For reward-independent delay distributions, the expected pseudo-regret of Algorithm 4 is bounded by*

$$\mathcal{R}_T \leq \min_{\vec{q}\in(0,1]^K} \sum_{i\neq i^*} \frac{40\log T}{\Delta_i}\left(\frac{1}{q_i} + \frac{1}{q_{i^*}}\right) + \log(K) \max_{i\neq i^*}\left\{(d_i(q_i) + d_{i^*}(q_{i^*}))\Delta_i\right\}. \tag{1}$$

Additionally, if instead we minimize over a single quantile $q \in (0,1]$, the expected pseudo-regret becomes

$$\mathcal{R}_T \leq \min_{q\in(0,1]} \sum_{i\neq i^*} \frac{325\log(T)}{q\Delta_i} + 4\max_{i\in[K]} d_i(q). \tag{2}$$

Particularly, Theorem 2 implies that for fixed delay d, we have $\mathcal{R}_T = O(\mathcal{R}_T^{MAB} + d)$. Note that the bounds in Eqs. (1) and (2) are incomparable: Eq. (1) allows choosing a different quantile for each arm, while Eq. (2) gives a slightly better dependence on K.

We now turn to show the main ideas of the proof of Theorem 2, deferring the full proof to the full version of the paper (Lancewicki et al., 2021).

Proof of Theorem 2 (sketch). Here we sketch the proof of Eq. (1); proving Eq. (2) is similar, but requires a more delicate argument in order to eliminate the K dependency in the second term.

Fix some vector $\vec{q} \in (0,1]^K$ and let $d_{max} = \max_{i \neq i^*} d_i(q_i)$. First, with high probability all the true means of the reward remain within the confidence interval (i.e., $\forall t, i : \mu_i \in [LCB_t(i), UCB_t(i)]$). Under this condition, the optimal arm is never eliminated. If a sub-optimal arm i was not eliminated by time t then, $LCB_t(i^*) \leq UCB_t(i)$. which implies with high probability,

$$\Delta_i = \mu_{i^*} - \mu_i \leq 2\sqrt{\frac{2\log(T)}{n_t(i)}} + 2\sqrt{\frac{2\log(T)}{n_t(i^*)}}.$$

Now, using a concentration bound, we show that the amount of *observed* feedback from arm j at time t, is approximately a fraction q_j of the number of pulls at time $t - d_j(q_j)$. We use that to bound $n_t(i)$ and $n_t(i^*)$ from below and obtain,

$$m_{t-d_{max}}(i) = O\left(\frac{\log T}{\Delta_i^2}\left(\frac{1}{q_i} + \frac{1}{q_{i^*}}\right)\right).$$

Now, if t is the last time we pulled arm i, then we can write the total regret from arm i as,

$$\begin{aligned} m_t(i)\Delta_i &= m_{t-d_{max}}(i)\Delta_i + (m_t(i) - m_{t-d_{max}}(i))\Delta_i \\ &\leq O\left(\frac{\log T}{\Delta_i}\left(\frac{1}{q_i} + \frac{1}{q_{i^*}}\right)\right) + m_t(i) - m_{t-d_{max}}(i). \end{aligned}$$

The difference $m_t(i) - m_{t-d_{max}}(i)$ is number of times we pull i between time $t - d_{max}$ and t. This is trivially bounded by d_{max}, but since we round-robin over active arms, we can divide it by the number of active arms. At the first elimination there are K active arms, in second there $K-1$ active arms, and so on. When summing the regret of all arms we get,

$$\begin{aligned} \mathcal{R}_T &= O\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\left(\frac{1}{q_i} + \frac{1}{q_{i^*}}\right)\right) + \log(K) d_{max} \\ &= O\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\left(\frac{1}{q_i} + \frac{1}{q_{i^*}}\right)\right) + \log(K) \max_i d_i(q_i), \end{aligned}$$

where we have used the fact that $1/K + 1/(K-1) + ... + 1/2 \leq \log(K)$. This proves the bound in Eq. (1). □

3.3 Phased Successive Elimination

Next, we introduce a phased version of successive elimination, we call Phased Successive Elimination (PSE). Inspired by phased versions of the commonly used algorithms (Auer & Ortner, 2010), the algorithm works in phases. Unlike SE, it does not round-robin naively, instead it attempts to maintains a balanced number of *observed* feedback at the end of each phase. As a result, PSE does not depend on the delay of the optimal arm. Surprisingly, the dependence on the delay of the sub-optimal arms remain similar, up to log-factors.

On each phase ℓ of PSE, we sample arms that were not eliminated in previous phase in a round-robin fashion. When we observe at least $16\log(T)/2^{-2\ell}$ samples for an active arm, we stop sampling it, but keep sampling the rest of active arms. Once we reach enough samples from all active arms, we perform elimination the same way we do on SE, and advance to the next phase $\ell + 1$. The full description of the algorithm is found in Algorithm 5.

Algorithm 5 Phased Successive Elimination (PSE)

Input: number of rounds T, number of arms K
Initialization: $S \leftarrow [K]$, $\ell \leftarrow 0$, $t \leftarrow 1$
while $t < T$ **do**
  Set $\ell \leftarrow \ell + 1$ (phase counter)
  Set $S_\ell \leftarrow S$
  while $S_\ell \neq \emptyset$ **do**
    Pull each arm $i \in S_\ell$, observe incoming feedback
    Set $t \leftarrow t + |S_\ell|$
    Call *Update-Parameters* (Procedure 2)
    Remove from S_ℓ all arms that where observed at least $16\log(T)/2^{-2\ell}$ times.
  Remove from S all arms i such that exists j with $UCB_t(i) < LCB_t(j)$

Theorem 3. *For reward-independent delay distributions, the expected pseudo-regret of Algorithm 5 (PSE) satisfies*

$$\begin{aligned} \mathcal{R}_T \leq \min_{\vec{q} \in (0,1]^K} \sum_{i \neq i^*} \frac{290\log T}{q_i \Delta_i} \\ + \log(T)\log(K)\max_{i \neq i^*} d_i(q_i)\Delta_i. \end{aligned} \tag{3}$$

The proof of Theorem 3 appears in the full version of the paper (Lancewicki et al., 2021). Similarly to the proof Theorem 2, both SE and PSE eliminate arm i approximately whenever

$$\sqrt{\frac{\log T}{n_t(i)}} + \sqrt{\frac{\log T}{n_t(i^*)}} \approx \Delta_i.$$

In a sense, PSE aims to shrink both terms in the left-hand side at a similar rate, which avoid the dependence on $1/q_{i^*}$ in the first term of Eq. (3). The down side is in the second

term: SE keeps sampling all active arms at the same rate, which gives rise to the $\log(K)$ dependence in the second term. Under PSE this is no longer the case: naively, one could show a linear dependence on K, but a more careful analysis that uses round-robin sampling within phases gives a $\log(T)\log(K)$ dependence in the second term of Eq. (3).

One important example in which PSE dominates SE is the arm-dependent packet loss setting, where we get the feedback of arm i immediately (i.e., zero delay) with probability p_i, and infinite delay otherwise. The regret of SE in this setting is $O(\sum_{i\neq i^*} \log(T)/\Delta_i \cdot (1/p_i + 1/p_{i^*}))$. On the other hand, PSE's regret is bounded by $O(\sum_{i\neq i^*} \log T/(\Delta_i p_i))$. The difference in the regret is substantial when p_{i^*} is very small. In fact, small amount of feedbacks from the optimal arm only benefits PSE, as it would keep sampling it until it gets enough feedbacks.

3.4 Lower Bound

We conclude this section with showing an instance-dependent lower bound (an instance is defined by the set of sub-optimality gaps Δ_i).

Theorem 4. *Let ALG^{delay} be an algorithm that guarantees a regret bound of T^α over any instance. For any sub-optimality gaps set $S_\Delta = \{\Delta_i : \Delta_i \in [0, \frac{1}{4}]\}$ of cardinality K, a quantile $q \in (0,1]$, and $\tilde{d} \leq T$, there exists an instance with an order on S_Δ, and delay distributions with $d_i(q) = \tilde{d}$ for any i, such that ALG^{delay}'s regret on that instance is*

$$\mathcal{R}_T \geq \frac{1}{128} \sum_{i:S_\Delta \ni \Delta_i > 0} \frac{(1-\alpha)\log T}{q\Delta_i} + \frac{1}{2}\overline{\Delta} \max_{i\in[K]} d_i(q) \tag{4}$$

for sufficiently large T, where $\overline{\Delta} = \frac{1}{K}\sum_{i\in[K]} \Delta_i$.

The lower bound is proved using delay distribution which is homogeneous across all arms: at time t, the delay is $\tilde{d}$ with probability q and ∞ otherwise. The upper bound of SE and PSE involves a minimization over q_i. In this case, it is solved by $q_i = q$ for all i. Therefore, the best comparison is to Eq. (2) in Theorem 2, where a single quantile is chosen. Theorem 4 shows that SE is near optimal in this case. The first term in Eq. (2) is aligned with Eq. (4), up to constant factors. The difference between the two is on the second term, where there is a $\overline{\Delta}$ factor in the lower bound.

The second term in Eq. (4) is due to the fact that the algorithm does not get any feedback for the first $\tilde{d} = d_i(q)$ rounds. Thus, any order on Δ is statistically indistinguishable from the others for the first rounds. Therefore, the learner suffers $\overline{\Delta}$ regret on average, over the first $\tilde{d}$ rounds, under at least one of the instances. The first term is achieved using a reduction from instance-depended lower bound for MAB without delays (Kleinberg et al., 2010; see also Lattimore & Szepesvári, 2020). The regret is bounded from below by this term, even if the instance I is known to the learner (the regret guarantee over the other instances ensures that the algorithm does not specialized particularly for that instance). A more detailed lower bound and its full proof is provided in the full version of the paper (Lancewicki et al., 2021).

4 Reward-dependent Delays

We next consider the more challenging case where we let the reward and the delays to be probabilistically dependent. Namely, there is no restriction on the reward-delay joint distribution.

The main challenge in this setting is that the *observed* empirical mean is no longer an unbiased estimator of the expected reward; e.g., if the delay given a reward of 0 is shorter than the delay given that the reward is 1, then the observed empirical mean would be biased towards 0. Therefore, the analysis from the previous section does not hold anymore. To tackle the problem, we present a new variant of successive elimination, Optimistic-Pessimistic Successive Elimination (OPSE), described in Algorithm 6. When calculating UCB the agent is optimistic regarding the unobserved samples, by assuming all missing samples have the maximal reward (one). When calculating LCB the agent assumes all missing samples have the minimal reward (zero). We emphasize that unlike the previous section, here the estimators take into account all samples, including the unobserved ones. The above implies that the confidence interval computed by OPSE contains the confidence interval computed by non-delayed SE.

Algorithm 6 Optimistic-Pessimistic Successive Elimination

Input: number of rounds T, number of arms K
Initialization: $S \leftarrow [K]$, $t \leftarrow 1$
while $t < T$ **do**
  Pull each arm $i \in S$
  Observe any incoming feedback
  Set $t \leftarrow t + |S|$
  for $i \in S$ **do**
    # the number of pulls and observations
    $m_t(i) \leftarrow \sum_{s<t} \mathbb{I}\{a_s = i\}$
    $n_t(i) \leftarrow \sum_{s:s+d_s<t} \mathbb{I}\{a_s = i\}$
    # pessimistic and optimistic estimators for μ_i
    $\hat{\mu}_t^-(i) \leftarrow \frac{1}{m_t(i)} \sum_{s:s+d_s<t} \mathbb{I}\{a_s = i\} r_s$
    $\hat{\mu}_t^+(i) \leftarrow \frac{m_t(i) - n_t(i)}{m_t(i)} + \hat{\mu}_t^-(i)$
    $LCB_t(i) \leftarrow \hat{\mu}_t^-(i) - \sqrt{\frac{2\log T}{m_t(i)}}$
    $UCB_t(i) \leftarrow \hat{\mu}_t^+(i) + \sqrt{\frac{2\log T}{m_t(i)}}$
  Remove from S all arms i such that exists j with $UCB_t(i) < LCB_t(j)$

For OPSE we prove the following regret guarantee.

Theorem 5. *For reward-dependent delay distributions, the expected pseudo-regret of Algorithm 6 is bounded by*

$$\mathcal{R}_T \leq \sum_{i \neq i^*} \frac{1166 \log T}{\Delta_i} + 4 \log(K) \left(\max_{i \neq i^*} d_i(q_i) + d_{i^*}(q_{i^*}) \right), \tag{5}$$

where $q_{i^*} = 1 - \min_{i \neq i^*} \Delta_i/4$ *and* $q_i = 1 - \Delta_i/4$ *for* $i \neq i^*$.

Theorem 5 is analogous to Theorem 2 in the reward-independent setting. We show a variant of SE, rather than PSE, because the algorithm relies on the entire feedback, rather than just the observed feedback. In addition, the dependence in $1/q_i$ was the main motivation to introduce PSE in the previous section, here it is bounded by a constant. In the reward-dependent setting we have much less information on the unobserved feedback, thus it would be unrealistic to expect similar regret bounds. The main difference between the two bounds is that here we are restricted to specific choice of quantiles q_i and q_{i^*}, while the bound in Theorem 2 hold for any vector $\vec{q}$. A second difference between the theorems is in the additive penalty due to the delay, here it is not multiplied by the sub-optimality gap, Δ_i. This factor Δ_i also appears in the lower bound in Theorem 6, which we discuss later on.

Proof of Theorem 5 (sketch). Consider time t in which arm i is still active. Define $\lambda_t(i) = \sqrt{2 \log(T)/m_t(i)}$. Let $\tilde{\mu}_t(i)$ be the empirical mean of arm i that is based on all $m_t(i)$ samples. Formally, $\tilde{\mu}_t(i) = \frac{1}{m_t(i)} \sum_{s<t} \mathbb{I}\{a_s = i\} r_s$.

This is the estimator that we would use to compute the confidence interval in non-delayed setting, but since not all observations are available at time t, we cannot compute it directly. Note that by definition,

$$\forall t, i: \quad \hat{\mu}_t^-(i) \leq \tilde{\mu}_t(i) \leq \hat{\mu}_t^+(i). \tag{6}$$

With high probability, using concentration bound on $\tilde{\mu}_t$ and Eq. (6) we can show that,

$$\begin{aligned} \Delta_i &= \mu_{i^*} - \mu_i \\ &\leq 4\lambda_t(i) + \hat{\mu}_t^+(i) - \hat{\mu}_t^-(i) + \hat{\mu}_t^+(i^*) - \hat{\mu}_t^-(i^*) \\ &= 4\lambda_t(i) + \frac{m_t(i) - n_t(i)}{m_t(i)} + \frac{m_t(i^*) - n_t(i^*)}{m_t(i^*)}. \end{aligned} \tag{7}$$

Let $d_{\max} = \max_{i \neq i^*} d_i(1 - \Delta_i/4)$. Using Hoeffding's inequality, with high probability, we have that,

$$n_t(i) \geq (1 - \Delta_i/4) m_{t - d_{\max}}(i) - \lambda_t(i) m_t(i).$$

Hence,

$$\begin{aligned} &\frac{m_t(i) - n_t(i)}{m_t(i)} \\ &= \frac{m_t(i) - m_{t-d_{\max}}(i)}{m_t(i)} + \frac{m_{t-d_{\max}}(i) - n_t(i)}{m_t(i)} \\ &\leq \frac{m_t(i) - m_{t-d_{\max}}(i)}{m_t(i)} + \Delta_i/4 + \lambda_t(i). \end{aligned}$$

The third term on the right hand side in Eq. (7) is bounded in a similar fashion, which gives us the following bound:

$$\Delta_i = O\left(\frac{2m_t(i) - m_{t-d_{\max}}(i) - m_{t-d^*_{\max}}(i)}{m_t(i)} + \sqrt{\frac{\log T}{m_t(i)}} \right),$$

where $d^*_{\max} = \max_{i \neq i^*} d_{i^*}(1 - \Delta_i)$. Either the last term on the right hand side is larger than the first two, or vice versa. By considering both cases and solving them, we yield the following result:

$$m_t(i)\Delta_i = O\left(\frac{\log T}{\Delta_i} + m_t(i) - m_{t-d_{\max}}(i) + m_t(i) - m_{t-d^*_{\max}}(i) \right).$$

The above holds for the last time we pull arm i, τ_i. Summing over the sub-optimal arms gives us a bound on regret. Similar to the setting of Section 3, $\sum_i m_{\tau_i}(i) - m_{\tau_i - d}(i) \leq \log(K) d$. Here, we set d to d_{max} or $d^*_{\max}$ accordingly, which gives us the desired regret bound. □

Optimistic-UCB. The dependency on the delay of the optimal arm comes from the bias of $\hat{\mu}_t^-$. A similar proof would hold for a variant of UCB that uses $\hat{\mu}_t^+$. In that case, one can obtain a regret bound of

$$O\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i} + \sum_{i \neq i^*} d_i(1 - \Delta_i/4) \right). \tag{8}$$

In most cases, this is a weaker bound than the bound of Theorem 5, as the second term scales linearly with the number of arms. The advantage of Optimistic-UCB is that it does not depend on the delay of the optimal arm. It still remains an open question whether we can enjoy the benefits of both bounds, and achieve a regret bound that depends only on $\max_{i \neq i^*} d_i(1 - \Delta_i)$.

On the other hand, in Theorem 6 we show that the dependence in $\max_{i \neq i^*} d_i(1 - \Delta_i)$ cannot be avoided, which establishes that our bound is not far from being optimal.

Theorem 6. *Let* $K = 2$. *For any* $\tilde{d} \leq T$ *and* $\Delta \in [0, 1/2]$, *there exist reward distributions with sub-optimality gap* Δ *and reward-dependent delay distributions with* $d_i(1 - 2\Delta) = \tilde{d}$, *such that,*

$$\mathcal{R}_T \geq \frac{1}{2} \Delta \cdot d_i(1 - 2\Delta). \tag{9}$$

Moreover, for any algorithm ALG^{delay} that guarantees a regret bound of T^{α} over any instance, the regret is at least,

$$\mathcal{R}_T = \Omega\left(\frac{(1-\alpha)\log(T)}{\Delta} + \Delta \cdot d_i(1-2\Delta)\right),$$

for sufficiently large T.

Note that $d_i(1-2\Delta) \leq d_i(1-\Delta/4)$, which complies with our upper bound. It seems necessary to have the Δ factor in Eq. (9), and we conjecture that it should also appear in the upper bound.

The proof for Theorem 6 is built upon two instances which are indistinguishable until time $\tilde{d}$. The reward distributions are Bernoulli and the index of the optimal arm alternates in the two instances. The idea is that when arm 2 is optimal, samples with reward 1 are delayed more often than samples with reward 0. When arm 2 is sub-optimal, the opposite occurs. The delay distribution is tailored such that under both instances, *(i)* the probability to observe feedback immediately is exactly $1-2\Delta$; and *(ii)* the probability for reward 1 given that the delay is 0, is identical for both arms under both instances. These two properties guarantee that the learner cannot distinguish between the two instances until time $\tilde{d}$. After that, it is possible to distinguish between them whenever a sample with delay $\tilde{d}$ is observed. The full details of the proof appears in the full version of the paper (Lancewicki et al., 2021).

5 Experiments

We conducted a variety of synthetic experiments to support our theoretical findings. We provide additional experiments in the full version of the paper (Lancewicki et al., 2021).

Fixed delays. In Fig. 1 we show the effect of different fixed delays on UCB and SE. We ran both algorithms with a confidence radius $\lambda_t(i) = \sqrt{2/n_t(i)}$, for $K = 20$ arms, each with Bernoulli rewards with mean uniform in $[0.25, 0.75]$, under various fixed delays. Top plots show cumulative regret until $T = 2 \cdot 10^4$. Bottom plot shows regret over increasing delays for $T = 2 \cdot 10^5$. The results are averaged over 100 runs and intervals in both plots are 4 times the standard error.

As delay increases, the regret of UCB increases as well, while SE is quite robust to the delay, and around delay of 200 SE becomes superior. These empirical results coincides with our theoretical results: As in the proof Theorem 1, the regret UCB grows linearly in the first Kd rounds. On the other hand, SE created a pipeline of observations, so it keeps getting observations from all active arms. While it cannot avoid from sampling each sub-optimal arm for d/K times, as long as this does not exceed the minimal amount of observations required for SE to eliminate a sub-optimal arm, the effect on the regret is minor.

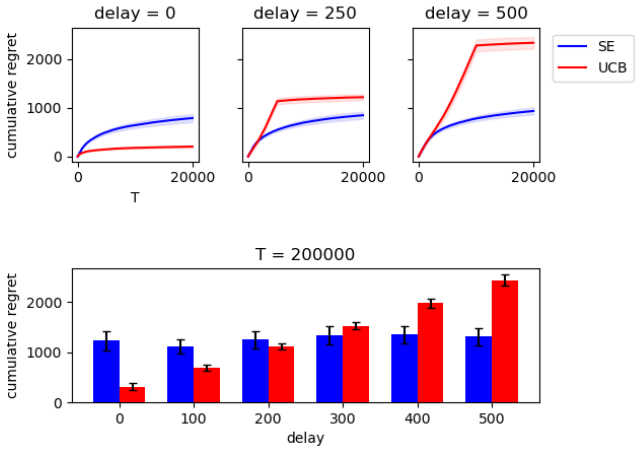

Figure 1. Regret of SE and UCB for fixed delays.

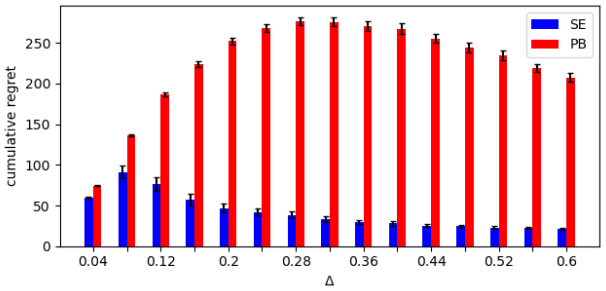

Figure 2. Regret of SE and PatientBandits (PB) for Pareto delays.

α-Pareto delays. We reproduce an experiment done by Gael et al. (2020) under our reward-independent setting, in Fig. 2. We compare their algorithm, PatientBandits (PB), with SE. For $T = 3000$ rounds and $K = 2$ arms, we ran sub-optimality gaps $\Delta \in [0.04, 0.6]$. The expected rewards are $\mu_1 = 0.4$ and $\mu_2 = 0.4 + \Delta$. The delay is sampled from Pareto distribution with $\alpha_1 = 1$ for arm 1 and $\alpha_2 = 0.2$ for arm 2. The results are averaged over 300 runs.

PB is a UCB-based algorithm that uses a prior knowledge on distribution in order to tune confidence radius. Even though it is designed to work under Pareto distributions, SE's regret is strictly smaller for any value of Δ. For small values of Δ, the regret increases with Δ, as the algorithms are not able to distinct between the arms. When Δ becomes large enough the regret starts to decrease as Δ increases. This transition occurs much sooner under SE, which indicates that SE starts to distinguish between the arms at lower values of Δ. We note that PB is designed for partial observation setting, which is more challenging than the reward-independent setting. However, the work of (Gael et al., 2020) is the only previous work, as far as know, to present a regret bound for delay distributions that potentially have infinite expected value and arm-dependent delays, as in this experiment.

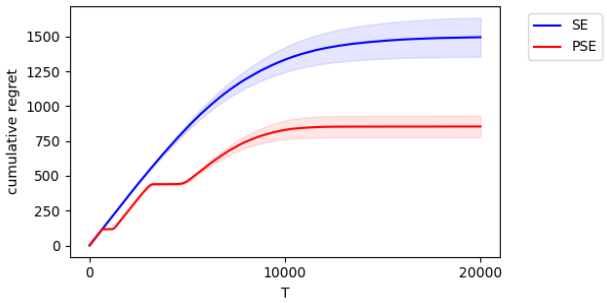

Figure 3. Regret of SE and PSE for packet loss delays.

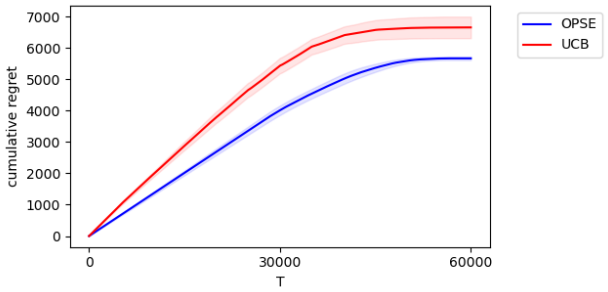

Figure 4. Reward-dependent setting. Regret of OPSE and UCB.

Packet-loss. We study the regret of SE and PSE in the packet loss setting. Specifically to evaluate the difference when amount of feedback from the best arm is significantly smaller than the other arms. We ran the algorithms for $T = 2 \cdot 10^4$ rounds and $K = 10$ arms with randomized values of sub-optimality gaps between $\Delta \in [0.15, 0.25]$. The probability to observe the best arm is 0.1, and 1 for the sub-optimal arms. The results are averaged over 300 runs. As seen in Fig. 3, the slope of PSE zeroes in some regions. This is the part of a phase in which the algorithm observed enough feedback from all sub-optimal arms and keeps sampling only the optimal arm. This happens due to the fact that the feedback of the optimal arm is unobserved 90% of the time. Meanwhile, SE samples each arm equally and receives less reward. The slope of PSE in other regions, is similar to the one of SE which indicates that the set of active arms is similar as well.

Reward-dependent case. We compare between OPSE (Algorithm 6) and UCB. We show that unlike in the reward-independent case, here an "off-the-shelf" solution doesn't perform very well, thus this case requires a modified algorithm. We set $T = 6 \cdot 10^4$ and $K = 3$ arms with random sub-optimality gaps of $\Delta \in [0.15, 0.25]$. The delay is *biased* with fixed delay of 5,000 rounds for reward 1 of the best arm and reward 0 of the sub-optimal arms. The results are averaged over 100 runs. In Fig. 4, OPSE outperforms UCB, mostly due to UCB's unawareness that the observed reward empirical means are biased. Thus, it favors the sub-optimal arms at the beginning and never recovers from that regret loss. We remark that in this settings, standard SE eliminates the best arm and suffers linear regret, so we omitted it.

6 Discussion

We presented algorithms for multi-arm bandits under two stochastic delayed feedback settings. In the reward-independent, which was studied previously, we present near-optimal regret bounds that scale with the delays quantiles. Those are significantly stronger, in many cases, then previous results. In addition we show a surprising gap between two classic algorithms: UCB and SE. While the former suffers a regret of $\Omega(\mathcal{R}_T^{MAB} + Kd)$ under fixed delays, the latter achieves $O(\mathcal{R}_T^{MAB} + d)$ for fixed delays and $O(\min_q \mathcal{R}_T^{MAB}/q + \max_i d_i(q))$ in the general setting. We further showed the PSE algorithm, which removes the dependency on the delay of the best arm. We then presented the reward-dependent delay setting, which is more challenging since the observed and the actual rewards distribute differently. Our novel OPSE algorithm achieves $O(\mathcal{R}_T^{MAB} + \log(K)d(1 - \Delta_{min}))$ by widening the gap of the confidence bounds to incorporate the potential observed biases. In both settings we provided almost matching lower bounds.

Our paper leaves some interesting future lines of research. The reward-dependent setting is mostly unaddressed in the literature and we believe there is more to uncover in this setting. One important question regards the gap between UCB and SE with fixed delays. In non-delayed multi-arm bandits, UCB and SE have similar regret bounds (and UCB even outperforms SE empirically when the delay is zero as evidence by Fig. 1). This raises the question: Can a variant of UCB or any other optimistic algorithm achieve similar regret bounds as a round-robin algorithm in the delayed settings? Lastly, another interesting direction is to tighten the regret bounds: In the reward independent case the gap between the lower and upper bound is either logarithmic in K (e.g., the bound in Eq. (1)) or missing a Δ factor on the delay term (e.g., Eq. (2)). In the reward dependent case it is still remains open question whether we can enjoy the benefits of both optimistic-SE and optimistic-UCB and obtain a regret bound that scales with $\max_{i \neq i^*} d_i(1 - \Delta_i)$.

Acknowledgments

The work of YM and TL has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grant number 993/17) and the Yandex Initiative for Machine Learning at Tel Aviv University. SS and TK were supported in part by the Israeli Science Foundation (ISF) grant no. 2549/19, by the Len Blavatnik and the Blavatnik Family foundation, and by the Yandex Initiative in Machine Learning.

References

Auer, P. and Ortner, R. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Bistritz, I., Zhou, Z., Chen, X., Bambos, N., and Blanchet, J. Online exp3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, pp. 11349–11358, 2019.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. Nonstochastic bandits with composite anonymous feedback. In *Conference On Learning Theory*, pp. 750–773, 2018.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research*, 20(1):613–650, 2019.

Csiszár, I. and Talata, Z. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Transactions on Information theory*, 52(3):1007–1016, 2006.

Desautels, T., Krause, A., and Burdick, J. W. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine Learning Research*, 15:3873–3923, 2014.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 169–178, 2011.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

Gael, M. A., Vernade, C., Carpentier, A., and Valko, M. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pp. 3348–3356. PMLR, 2020.

György, A. and Joulani, P. Adapting to delays and data in adversarial multi-armed bandits. *arXiv preprint arXiv:2010.06022*, 2020.

Joulani, P., Gyorgy, A., and Szepesvári, C. Online learning under delayed feedback. In *International Conference on Machine Learning*, pp. 1453–1461, 2013.

Kleinberg, R., Niculescu-Mizil, A., and Sharma, Y. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2-3):245–272, 2010.

Lancewicki, T., Segal, S., Koren, T., and Mansour, Y. Stochastic multi-armed bandits with unrestricted delay distributions. *arXiv preprint arXiv:2106.02436*, 2021.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Pike-Burke, C., Agrawal, S., Szepesvari, C., and Grunewalder, S. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pp. 4105–4113. PMLR, 2018.

Thune, T. S., Cesa-Bianchi, N., and Seldin, Y. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, pp. 6541–6550, 2019.

Vernade, C., Cappé, O., and Perchet, V. Stochastic bandit models for delayed conversions. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

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

Zhou, Z., Xu, R., and Blanchet, J. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, pp. 5197–5208, 2019.

Zimmert, J. and Seldin, Y. An optimal algorithm for adversarial bandits with arbitrary delays. In *International Conference on Artificial Intelligence and Statistics*, pp. 3285–3294. PMLR, 2020.