An Optimal Algorithm for Adversarial Bandits with Arbitrary Delays

Julian Zimmert **Yevgeny Seldin**
University of Copenhagen

Abstract

We propose a new algorithm for adversarial multi-armed bandits with unrestricted delays. The algorithm is based on a novel hybrid regularizer applied in the Follow the Regularized Leader (FTRL) framework. It achieves $\mathcal{O}(\sqrt{kn} + \sqrt{D\log(k)})$ regret guarantee, where k is the number of arms, n is the number of rounds, and D is the total delay. The result matches the lower bound within constants and requires no prior knowledge of n or D. Additionally, we propose a refined tuning of the algorithm, which achieves $\mathcal{O}(\sqrt{kn} + \min_S(|S| + \sqrt{D_{\bar{S}}\log(k)}))$ regret guarantee, where S is a set of rounds excluded from delay counting, $\bar{S} = [n] \setminus S$ are the counted rounds, and $D_{\bar{S}}$ is the total delay in the counted rounds. If the delays are highly unbalanced, the latter regret guarantee can be significantly tighter than the former. The result requires no advance knowledge of the delays and resolves an open problem of Thune et al. (2019). The new FTRL algorithm and its refined tuning are anytime and require no doubling, which resolves another open problem of Thune et al. (2019).

1 Introduction

Multi-armed bandits are a fundamental sequential decision making problem with an increasing number of industrial applications. In the multi-armed bandit setting, a learner repeatedly chooses an action from a finite set of actions and immediately observes a loss for that specific action. The action might be, for example, a choice of an advertisement layout out of a finite set of layouts. The loss could be the response of a user to the layout, for example, a lack of a click on the advertisement. In practice, it is often required to make decisions for new users before observing the feedback of all previous users, either due to response latency or parallel interaction with multiple users. This can be modeled by introducing a *delay* between the action and observation.

We focus on the oblivious adversarial (a.k.a. non-stochastic) bandit setting, meaning that the sequence of losses and the delays are fixed before the start of the game. The setting was first studied by Cesa-Bianchi et al. (2016) under the assumption of uniform delays, which are all equal to d. They proved a lower bound of $\Omega(\max\{\sqrt{kn}, \sqrt{dn\log(k)}\})$ for $d \leq n/\log(k)$ (they do not report the $\log(k)$ term) and an almost matching upper bound of $\mathcal{O}(\sqrt{kn\log(k)} + \sqrt{dn\log(k)})$. By translating individual delays into the total delay $D = dn$ the lower bound for uniform delays is $\Omega(\max\{\sqrt{kn}, \sqrt{D\log(k)}\})$. Thune et al. (2019) and Bistritz et al. (2019) independently derived an algorithm that can handle non-uniform delays and achieves an $\mathcal{O}(\sqrt{kn\log(k)} + \sqrt{D\log(k)})$ regret bound under the assumption that n and D are known in advance. Thune et al. further provide a doubling scheme that achieves the same regret bound under the assumption that the delays d_t are known "at action time", i.e., at time t, but n and D are unknown, whereas Bistritz et al. provide a doubling scheme that achieves an $\mathcal{O}(\sqrt{k^2n\log(k)} + \sqrt{D\log(k)})$ regret bound when n and D are unknown and the delays d_t are observed together with the observations, i.e., at time $t + d_t$.

Thune et al. further observe that if the delays are highly unbalanced it may be worth "skipping" rounds with excessively large delays. "Skipping" means that the regret in the corresponding round is trivially bounded by 1 and the observation is ignored by the algorithm. The skipping approach of Thune et al. requires knowledge of the delays "at action time". Under the assumption that this information is available, Thune et al. provide an algorithm that achieves $\mathcal{O}(\min_\beta |S_\beta| + \beta\log(k) + \beta^{-1}(kn + D_{\bar{S}_\beta}))$ regret guaran-

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

Table 1: Overview of state-of-the-art regret bounds for multi-armed bandits with delayed feedback. (*) requires oracle knowledge of the time horizon n and the total delay D; the result appeared independently in two papers. (**) requires advance knowledge of the delays d_t "at action time" t.

Setting	Regret upper and lower bounds		Reference
Uniform delays d	$\Omega(\max\{\sqrt{kn}, \sqrt{dn\log(k)}\})$		Cesa-Bianchi et al. (2016)
	$\mathcal{O}(\sqrt{kn\log(k)} + \sqrt{dn\log(k)})$		Cesa-Bianchi et al. (2016)
	$\mathcal{O}(\sqrt{kn} + \sqrt{dn\log(k)})$		This paper
Arbitrary delays,	$\mathcal{O}(\sqrt{kn\log(k)} + \sqrt{D\log(k)})$	(*)	Thune et al. (2019); Bistritz et al. (2019)
non-adaptive bounds	$\mathcal{O}(\sqrt{k^2n\log(k)} + \sqrt{D\log(k)})$		Bistritz et al. (2019)
	$\mathcal{O}(\sqrt{kn} + \sqrt{D\log(k)})$		This paper
Arbitrary delays,	$\mathcal{O}(\min_\beta \lvert S_\beta\rvert + \beta\log(k) + \beta^{-1}(kn + D_{\bar{S}_\beta}))$	(**)	Thune et al. (2019)
adaptive bounds	$\mathcal{O}(\sqrt{kn} + \min_S(\lvert S\rvert + \sqrt{D_{\bar{S}}\log(k)}))$		This paper

tee, where β is the skipping threshold (the rounds with delays $d_t \geq \beta$ are skipped), S_β is the set of skipped rounds and $|S_\beta|$ is their number, $\bar{S}_\beta = [n] \setminus S_\beta$ are the remaining rounds (where $[n] = \{1, \dots, n\}$), and $\mathrm{D}_{\bar{S}_\beta} = \sum_{t\in\bar{S}_\beta} d_t$ is their total delay. Thune et al. provide an example, where the first $\lfloor\sqrt{kn/\log(k)}\rfloor$ rounds have delays of order n and the remaining rounds have zero delays. By skipping the first rounds, the dependence of the regret bound on n improves from order $n^{3/4}$ to $n^{1/2}$. The skipping procedure of Thune et al. crucially depends on availability of delays "at action time" in order to make the skipping decision and the skipping threshold β is tuned by doubling. Relaxation of the assumption on early availability of delays, as well as replacement of doubling with anytime strategies (i.e., algorithms without resets) were left as open questions.

We resolve both open questions and make the following contributions:

1. We provide an anytime FTRL algorithm based on a novel hybrid regularizer. The regularizer combines $\frac{1}{2}$-Tsallis entropy and negative entropy, each with its own learning rate. The algorithm requires no advance knowledge of the delays and achieves a regret bound of $\mathcal{O}(\sqrt{kn} + \sqrt{D\log(k)})$, which matches the lower bound within constants.

2. We provide a novel "skipping" technique, which allows to "ignore" rounds with excessively large delays with no advance knowledge of the delays. We put "skipping" and "ignore" in quotation marks, because the observations are still used by the algorithm and the "skipped" rounds are only excluded from updates of the learning rate. We prove an $\mathcal{O}(\sqrt{kn} + \min_S(|S| + \sqrt{D_{\bar{S}}\log(k)}))$ regret bound for the refined algorithm. The bound is slightly tighter than the refined regret bound of Thune et al. (2019), but most importantly it requires no advance knowledge of the delays. [1]

In Table 1 we provide a comparison of state-of-the art bounds with our new results. Additional prior work in other online learning settings with delayed feedback includes the full information setting studied by Joulani et al. (2016), who derived a general reduction to a non-delayed problem. To the best of our knowledge, no similar reduction under bandit feedback has been found yet. Another related setting are bandits with anonymous composite feedback, where the learner is not informed about the round from which the delayed observation is coming from, neither the identity of the action it corresponds to, and delayed observations from several rounds may be composed together with no possibility to separate them. This harder setting was studied by Cesa-Bianchi et al., who derived an $\mathcal{O}(\sqrt{kd_{max}n\log(k)})$ regret bound, where d_{max} is a known upper bound on the delays. We refer the reader to Thune et al. (2019) for further review of prior work in related settings.

The paper is structured in the following way: Section 2 provides a formal definition of the problem setting. Section 3 explains in detail our algorithm and two versions of learning rate tuning. Section 4 contains our main theorems, as well as an intuition behind the refined learning rate tuning. Section 5 presents a general analysis of FTRL for multi-armed bandits with delays and formally proves the theorems from the previous section. Finally, Section 6 provides a summary and directions for future work.

[1] We note that the new skipping technique could also be combined with the doubling scheme of Thune et al. to eliminate the need in advanced knowledge of delays there as well. However, the anytime FTRL algorithm presented here is much more elegant than doubling.

2 Problem setting

Adversarial bandits with delay is a sequential game between a learner and an environment with k fixed actions. At time steps $t = 1, \dots, n$ the learner picks actions $A_t \in [k]$ and immediately suffers the loss ℓ_{t,A_t}, where $(\ell_t)_{t=1,\dots,n}$ are vectors in $[0,1]^k$. Unlike in the regular bandit problem, the learner does not necessarily observe the loss ℓ_{t,A_t} at the end of round t. Instead, the environment chooses a sequence of delays $(d_t)_{t=1,\dots,n}$ and the player observes the tuples (s, ℓ_{s,A_s}) for each s such that $s + d_s = t$ at the end of round t. Without loss of generality, we assume that all outstanding tuples are observed at the end of the game, i.e., $t + d_t \leq n$ for all t. We focus on the oblivious adversarial setting (sometimes called "non-stochastic"), which means that both the sequence of losses $(\ell_t)_{t=1,\dots,n}$ and the sequence of delays $(d_t)_{t=1,\dots,n}$ are chosen by the environment at the beginning of the game. We use $D = \sum_{t=1}^n d_t$ to denote the total delay. The learner has no prior knowledge of the quantities n, D, or $(d_t)_{t=1,\dots,n}$. The performance of the algorithm is measured by its expected regret

$$\mathfrak{R}_n := \mathbb{E}\left[\sum_{t=1}^n \ell_{t,A_t}\right] - \min_{i\in[k]} \sum_{t=1}^n \ell_{t,i}\,.$$

Some technical definitions We use $\Delta([k]) = \{x \in \mathbb{R}_+^k \,|\, \|x\|_1 = 1\}$ to denote the $(k-1)$-simplex. For a set $S \subset [n] = \{1, \dots, n\}$, we denote its complement by $\bar{S} = [n] \setminus S$. For a convex function F we use F^* to denote its convex conjugate (a.k.a. Fenchel conjugate) and $\overline{F}^*$ to denote the constrained convex conjugate. They are defined, respectively, by

$$F^*(y) = \max_{x\in\mathbb{R}^k} \langle x, y\rangle - F(x),$$
$$\overline{F}^*(y) = \max_{x\in\Delta([k])} \langle x, y\rangle - F(x)\,.$$

3 Algorithm

Our Algorithm 1 is a standard Follow the Regularized Leader (FTRL) algorithm that works with importance weighted loss estimators of all observations available up to the current point in time. The loss estimators are defined by

$$\hat{\ell}_s = \frac{\ell_{s,A_s}}{x_{s,A_s}} \mathbf{e}_{A_s}\,,$$

where x_{s,A_s} is the algorithm's probability of selecting action A_s at round s and $\mathbf{e}_{A_s}$ is a standard basis vector. We define the cumulative observed loss estimator at time t by

$$\hat{L}_t^{obs} = \sum_{s:s+d_s<t} \hat{\ell}_s\,.$$

Given a convex regularizer $F_t : \mathbb{R}^k \to \mathbb{R}$, FTRL samples action A_t according to the distribution

$$x_t = \operatorname*{arg\,min}_{x\in\Delta([k])} \langle x, \hat{L}_t^{obs}\rangle + F_t(x)\,.$$

x_t can be equivalently expressed as $x_t = \nabla \overline{F}_t^*(-\hat{L}_t^{obs})$.

We are using a hybrid regularizer $F_t = F_{t,1} + F_{t,2}$, where in contrast to most prior work each of the two parts of the regularizer has its own learning rate.

$$\underbrace{F_t(x)}_{=\sum_{i=1}^k f_t(x_i)} = \underbrace{-\sum_{i=1}^k 2\sqrt{t}x_i^{1/2}}_{F_{t,1}(x)=\sum_{i=1}^k f_{t,1}(x_i)} + \underbrace{\eta_t^{-1}\sum_{i=1}^k x_i\log(x_i)}_{F_{t,2}(x)=\sum_{i=1}^k f_{t,2}(x_i)}\,.$$

The first part of the regularizer $F_{t,1}(x) = \sqrt{t}F_1(x)$ is the $\frac{1}{2}$-Tsallis entropy $F_1(x) = -2\sum_{i=1}^k \sqrt{x_i}$ with learning rate $\frac{1}{\sqrt{t}}$, which is non-adaptive to the problem. The second part of the regularizer $F_{t,2}(x) = \eta_t^{-1}F_2(x)$ is the negative entropy $F_2(x) = \sum_{i=1}^k x_i\log(x_i)$ with adaptive learning rate η_t. We call a sequence of learning rates $(\eta_t)_{t=1,\dots,n}$ *proper* if it is non-increasing and can be defined using information available at the beginning of round t.

3.1 Intuition behind the regularizer

Hybrid regularizers have been successfully used in adaptive regret bounds for sparse bandits, online portfolio selection, adversarially robust semi-bandits, and adaptive first order bounds for multi-armed bandits (Bubeck et al., 2018; Luo et al., 2018; Zimmert et al., 2019; Pogodin and Lattimore, 2019). They are useful for targeting multiple objectives. In our case, the regret lower bound for bandits with fixed delay d is $\Omega(\max\{\sqrt{kn}, \sqrt{dn\log(k)}\})$ (Cesa-Bianchi et al., 2016). The first part of the bound is the standard regret lower bound for multi-armed bandits with no delays, which is clearly also a lower bound for the problem with delays. The second part of the bound is achieved by grouping the game rounds into batches of size d and reducing the game to a full information game over n/d rounds with loss range $[0, d]$. The second part is then a lower bound on the regret in the full information game.

Our regularizer uses the same decomposition of the problem. We combine the optimal regularizer for the standard bandit problem with no delay, the $\frac{1}{2}$-Tsallis Entropy, with the optimal regularizer for the full information problems, the negative entropy. We further tune the learning rate for the second part to the actual delay sequence $(d_t)_{t=1,\dots,n}$.

3.2 Tuning of the learning rate

We propose and analyze two versions of learning rate tuning. The *simple tuning* is given in Algorithm 1. For *advanced tuning*, replace the colored blocks **Initialize** and **determine** η_t in Algorithm 1 with the corresponding blocks from Algorithm 2.

Algorithm 1: FTRL for bandits with delay

Input: Proper learning rate rule η_t
Initialize $\hat{L}_1^{obs} = 0$
Initialize $\mathfrak{D}_0 = 0$ (simple tuning)
for $t = 1, \dots, n$ **do**
  determine η_t
    Set $\mathfrak{D}_t = \mathfrak{D}_{t-1} + \mathfrak{d}_t$
    Set $\eta_t^{-1} = \sqrt{2\mathfrak{D}_t / \log(k)}$ } (simple tuning)
  Set $x_t = \arg\min_{x \in \Delta([k])} \langle x, \hat{L}_t^{obs} \rangle + F_t(x)$
  Sample $A_t \sim x_t$
  for $s : s + d_s = t$ **do**
    Observe (s, ℓ_{s,A_s})
    Construct $\hat{\ell}_s$ and update $\hat{L}_t^{obs}$

Simple tuning We define the key quantity, which is used for tuning the learning rate.

Definition 1. *The* number of outstanding observations *at round t is defined by*

$$\mathfrak{d}_t = \sum_{s=1}^{t-1} \mathbb{I}\{s + d_s \geq t\},$$

where $\mathbb{I}$ is the indicator function.

$\mathfrak{d}_t$ counts how many observations from rounds $s < t$ are still missing at the beginning of round t. Note that $\mathfrak{d}_t$ is an observable quantity, unlike the delays d_t. Therefore, $\mathfrak{d}_t$ can be used for online tuning of the learning rate. The learning rate under the *simple tuning* is given by

$$\mathfrak{D}_t = \sum_{s=1}^{t} \mathfrak{d}_t \quad , \quad \eta_t^{-1} = \sqrt{2\mathfrak{D}_t / \log(k)}\,.$$

The algorithm only uses the inverse of the learning rate. If $\mathfrak{D}_t = 0$, then $\eta_t^{-1} = 0$ and the algorithm is well-defined, even though $\eta_t = \infty$.

Advanced tuning In the advanced tuning, we maintain a running estimate $\tilde{\mathfrak{D}}_t$ of the optimal truncated delay $D_{\bar{S}}$. To achieve that, we modify the quantity $\mathfrak{d}_t$ by "skipping" some outstanding observations. To be precise, we keep indicator variables $a_s^t \in \{0, 1\}$, where a_s^t indicates whether an outstanding observation from round s should still be counted at round t:

$$\tilde{\mathfrak{d}}_t = \sum_{s=1}^{t-1} a_s^t \mathbb{I}\{s + d_s \geq t\}.$$

We define

$$\tilde{\mathfrak{D}}_t = \sum_{s=1}^{t} \tilde{\mathfrak{d}}_t \quad , \quad \eta_t^{-1} = \sqrt{\tilde{\mathfrak{D}}_t / \log(k)}\,.$$

The algorithm initially waits for all observations, but as soon as the waiting time exceeds a threshold the round is "skipped". If we observe a delay such that $d_s > \sqrt{\tilde{\mathfrak{D}}_t / \log(k)}$, we set $(a_s^{t'})_{t' > t}$ to 0. The indicators are not changed retrospectively, which means that the initial waiting time still counts toward $\tilde{\mathfrak{D}}_t$. The intuition behind advanced tuning is explained in Section 4.3.

Algorithm 2: Advanced tuning of η_t for Alg. 1

Initialize $\tilde{\mathfrak{D}}_0 = 0$ and $(a_s^t)_{s=1,\dots,n;t=1,\dots,n} = 1$
determine η_t
  Set $\tilde{\mathfrak{d}}_t = \sum_{s=1}^{t-1} \mathbb{I}\{s + d_s \geq t\} a_s^t$
  Update $\tilde{\mathfrak{D}}_t = \tilde{\mathfrak{D}}_{t-1} + \tilde{\mathfrak{d}}_t$
  Set $\eta_t^{-1} = \sqrt{\tilde{\mathfrak{D}}_t / \log(k)}$
  for $s = 1, \dots, t-1$ **do**
    if $\min\{d_s, t - s\} > \eta_t^{-1}$ **then**
      $(a_s^{t'})_{t' > t} = 0$ (At most one index s satisfies the **if**-condition, see Lemma 7)

4 Main results

In this section, we present regret upper bounds for Algorithm 1 with *simple tuning* and *advanced tuning*. The first result confirms the conjecture of Cesa-Bianchi et al. (2016) that an upper bound of $\mathcal{O}(\sqrt{kn} + \sqrt{D \log(k)})$ is achievable with a simple algorithm. The second result shows that it is possible to obtain a refined bound of $\mathcal{O}(\sqrt{kn} + \min_{S \subset [n]}(|S| + \sqrt{D_{\bar{S}} \log(k)}))$ by a more careful tuning of the learning rate.

4.1 Adaptation to the total delay D

The following theorem provides a regret bound for Algorithm 1 with *simple tuning*.

Theorem 1. *The regret of Algorithm 1 with any non-increasing positive sequence of learning rates $(\eta_t)_{t=1,\dots,n}$ satisfies*

$$\mathfrak{R}_n \leq 4\sqrt{kn} + \eta_n^{-1} \log(k) + \sum_{t=1}^{n} \eta_t \mathfrak{d}_t\,.$$

In particular, the simple tuning $\eta_t^{-1} = \sqrt{2\mathfrak{D}_t/\log(k)} = \sqrt{2(\sum_{s=1}^{t}\mathfrak{d}_s)/\log(k)}$ *is proper and leads to a regret bound of*

$$\mathfrak{R}_n \le 4\sqrt{kn} + \sqrt{8D\log(k)}\,.$$

Proof. The first statement is a special case of Theorem 3, which is presented in Section 5. For the second statement we use a standard summation lemma, by which for a sequence of positive $\mathfrak{d}_1, \dots, \mathfrak{d}_n$ we have $\sum_{t=1}^{n}\left(\mathfrak{d}_t/\sqrt{\sum_{s=1}^{t}\mathfrak{d}_s}\right) \le 2\sqrt{\sum_{s=1}^{t}\mathfrak{d}_t}$ (Seldin et al., 2014, Lemma 8) and the convention that if $\mathfrak{d}_t = 0$ then $\eta_t\mathfrak{d}_t = 0$ (so that zero terms naturally fall out of the summation). By substituting the definition of the learning rate in the second statement into the first statement and using the summation lemma we obtain

$$\mathfrak{R}_n \le 4\sqrt{kn} + \sqrt{8\mathfrak{D}_n\log(k)}\,.$$

Finally, note that an observation from round t with delay d_t contributes 1 to each of $\mathfrak{d}_t, \dots, \mathfrak{d}_{t+d_t}$, i.e., it contributes d_t to the total sum of the number of outstanding observations $\sum_{t=1}^{n}\mathfrak{d}_t$. Since we have assumed that $t + d_t \le n$ for all t, we have $\sum_{t=1}^{n}\mathfrak{d}_t = \sum_{t=1}^{n} d_t = D$. □

The main advantage of Algorithm 1 and Theorem 1 compared to the work of Thune et al. (2019) is that the tuning requires neither the knowledge of D and n, nor doubling.

4.2 Refined bounds for unbalanced delays

Thune et al. (2019) observed that if the delays are highly unbalanced it may be worth skipping rounds with overly large delays rather than keeping them in the analysis. Let S denote the set of skipped rounds and $|S|$ their number. The regret in every skipped round is trivially bounded by 1 and, assuming we knew which rounds to skip, we could reduce the regret bound to $\mathcal{O}\left(\sqrt{kn} + |S| + \sqrt{D_{\bar{S}}\log(k)}\right)$. As shown by Thune et al., this could potentially be much smaller than the regret bound in Theorem 1. For example, if the delay in the first $\theta(\sqrt{kn})$ rounds is of order n and the delay in the remaining rounds is zero, then the regret bound in Theorem 1 is of order $n^{3/4}$, whereas the refined regret bound is of order $n^{1/2}$ (ignoring the dependence on k). The challenge faced by Thune et al. was that they had to know the delays in advance (more precisely, "at action time") in order to tune the parameters of their algorithm and make the skipping decision. Since we have an anytime algorithm, we are able to obtain the refinement with no need in advance knowledge of the delay information. Strictly speaking, we even do not need to skip observations and we can obtain the refinement by using all observations and only adjusting the learning rate appropriately, although technically the "no-skipping" solution yields the same regret bound as skipping.

The following theorem provides our adaptive bound.

Theorem 2. *Algorithm 1 with advanced learning rate tuning provided in Algorithm 2 satisfies*

$$\mathfrak{R}_n \le 4\sqrt{kn} + 10\max\begin{cases}\min_{S\subset[n]}|S| + \sqrt{D_{\bar{S}}\log(k)},\\ 2\log(k).\end{cases}$$

The proof is postponed to Section 5

4.3 Intuition behind the "skipping" procedure

In order to give an intuition behind the refined algorithm we provide a simple back-of-the-envelope calculation. If we skip $|S|$ rounds and trivially bound their regret by 1 and apply Theorem 1 to the remaining rounds, then the regret bound is $\mathcal{O}(\sqrt{kn} + \sqrt{D_{\bar{S}}\log(k)} + |S|)$. Thus, the number of skipped rounds can be as large as $\sqrt{D_{\bar{S}}\log(k)}$ without significantly impacting the bound. Obviously, we want to skip rounds with the largest delays, but how should we determine the skipping threshold X? If we want to achieve a significant reduction in the regret bound, the skipped delay $D_S = \sum_{t\in S} d_t \ge X|S|$ should be at least as large as the remaining delay $D_{\bar{S}}$, because $D = D_S + D_{\bar{S}}$ and our aim is to reduce the $\sqrt{D\log k}$ term. Thus, if we put a threshold at X and skip $\sqrt{D_{\bar{S}}\log(k)}$ rounds we want to have $X\sqrt{D_{\bar{S}}\log(k)} \ge D_{\bar{S}}$. Therefore, we aim at $X = \sqrt{D_{\bar{S}}/\log(k)}$. However, there are two challenges: (a) we do not know the delays d_t in advance and, therefore, we do not know which rounds to skip, and (b) the threshold definition is recursive: X depends on $D_{\bar{S}}$ and $D_{\bar{S}}$ depends on X.

The strategy that we take in Algorithm 2 is the following: we keep a running estimate $\tilde{\mathfrak{D}}_t$ of $D_{\bar{S}}$. For an observation from round s we initially start waiting and count it in the number of outstanding observations $\tilde{\mathfrak{d}}_t$ for the initial rounds. However, we constantly monitor the waiting time and if the observation has not arrived within $\sqrt{\tilde{\mathfrak{D}}_t/\log(k)}$ rounds we stop waiting. The initial rounds we have been waiting for still count for the estimate $\tilde{\mathfrak{D}}_t$. Another quick back-of-the-envelope calculation shows that if $\tilde{\mathfrak{D}}_t$ is indeed a good approximation of $D_{\bar{S}}$, then the extra delay from the initial waiting rounds is of order $\sqrt{D_{\bar{S}}\log(k)}\sqrt{D_{\bar{S}}/\log(k)} = D_{\bar{S}}$, where the first term is a rough estimate of the number of rounds that we skip and the second term is a

rough estimate of the initial waiting time for each of the observations. Thus, the initial waiting time has no significant impact on the final bound.

Algorithm 2 follows this intuitive approach. We use indicator variables $(a_s^t)_{(s,t)\in[n]^2}$ to keep track of which observations ℓ_{s,A_s} we are still waiting for at round t (expressed by $a_s^t = 1$) and which not (expressed by $a_s^t = 0$). We use $\tilde{\mathfrak{d}}_t$ to count the truncated number of outstanding observations, where those observations we are no longer waiting for at round t are excluded from counting. We provide a detailed analysis in Section 5.2, but before we get there we provide a refined version of Theorem 1, which allows us to use all observations and only use skipping in the tuning of the learning rate. (Though, as already mentioned, complete skipping of the observations would lead to the same regret bound as in Theorem 2.)

5 Analysis of FTRL for bandits with delays

In this section we develop a novel analysis of FTRL-style algorithms and present a generalization of the first part of Theorem 1. The analysis is based on a permuted counting of losses, similar to the techniques used by Joulani et al. (2013) and Thune et al. (2019). Afterward, we use the general regret bound to prove Theorem 2.

5.1 Dependency preserving permutations

Reordering of losses by a permutation $\rho : [n] \to [n]$ is a useful tool in the analysis of online learning with delays. Joulani et al. (2013) have used "ordering by arrival", where the losses $\hat{\ell}_s$ are sorted by the time of arrival $s + d_s$ with ties broken arbitrarily. We generalize this type of analysis by studying a general class of admissible permutations. This also provides insights into why it is useful to consider permutations.

Definition 2. *A permutation $\rho : [n] \to [n]$ is* dependency preserving *if it satisfies:*

$$\forall\, s,t \in [n] :\; s + d_s < t \;\Rightarrow\; \rho(s) < \rho(t)\,.$$

It means that if at the beginning of round t the loss ℓ_{s,A_s} has been already observed (and thus can influence the selection of A_t), then s must come before t under the permutation.
Furthermore, we define the ρ-number of outstanding observations *at time t by*

$$\mathfrak{d}_t^\rho = \sum_{s:\rho(s)<\rho(t)} \mathbb{I}\{s + d_s \geq t\}\,.$$

Example 1. *The identity function* $\mathrm{id}(t) = t$ *is dependency preserving, since ℓ_{s,A_s} being observed before t implies* $\mathrm{id}(s) = s < t = \mathrm{id}(t)$.

Example 2. *"Ordering by arrival" is dependency preserving, since $s + d_s < t \;\Rightarrow\; s + d_s < t + d_t \;\Rightarrow\; \rho(s) < \rho(t)$.*

The ρ-number of outstanding observations extends the previous definition of the number of outstanding observations in the sense that $\mathfrak{d}_t = \mathfrak{d}_t^{id}$. Furthermore, the property $\sum_{t=1}^n \mathfrak{d}_t^\rho = D$ holds for any dependency preserving permutation ρ (refer to Lemma 6 in the supplementary material, Section 7.1).

Next we present a general regret bound which holds for any dependency preserving permutation ρ.

Theorem 3. *For any dependency preserving permutation ρ, the regret of Algorithm 1 with non-increasing positive learning rates $(\eta_t)_{t=1,\dots,n}$ satisfies*

$$\mathfrak{R}_n \leq 4\sqrt{kn} + \eta_n^{-1}\log(k) + \sum_{t=1}^n \min\{1, \eta_t \mathfrak{d}_t^\rho\}\,.$$

Remark 1. *The first part of Theorem 1 is a direct corollary using $\rho = \mathrm{id}$.*

The proof uses Lemmas 1, 2, and 3. In order to motivate them we first present the proof and then the lemmas.

Proof. We define cumulative losses $\hat{L}_t^\rho = \sum_{s:\rho(s)<\rho(t)} \hat{\ell}_s$ (and by convention $\hat{L}_{n+1}^\rho = \sum_{s=1}^n \hat{\ell}_s$) and $i^* = \arg\min \sum_{t=1}^n \ell_{t,i}$. We decompose the regret into three terms:

$$\begin{aligned}
\mathfrak{R}_n &= \mathbb{E}\left[\sum_{t=1}^n \ell_{t,A_t} - \ell_{t,i^*}\right] = \mathbb{E}\left[\sum_{t=1}^n \langle x_t, \hat{\ell}_t\rangle - \langle \mathbf{e}_{i^*}, \hat{\ell}_t\rangle\right] \\
&= \mathbb{E}\Bigg[\underbrace{\sum_{t=1}^n \left(\overline{F}_t^*(-\hat{L}_t^{obs} - \hat{\ell}_t) - \overline{F}_t^*(-\hat{L}_t^{obs}) + \langle x_t, \hat{\ell}_t\rangle\right)}_{(A)} \\
&\quad + \underbrace{\sum_{t=1}^n \Big(\overline{F}_t^*(-\hat{L}_t^{obs}) - \overline{F}_t^*(-\hat{L}_t^{obs} - \hat{\ell}_t) - \overline{F}_t^*(-\hat{L}_t^\rho) + \overline{F}_t^*(-\hat{L}_{t+1}^\rho)\Big)}_{(B)} \\
&\quad + \underbrace{\sum_{t=1}^n \left(\overline{F}_t^*(-\hat{L}_t^\rho) - \overline{F}_t^*(-\hat{L}_{t+1}^\rho) - \langle \mathbf{e}_{i^*}, \hat{\ell}_t\rangle\right)}_{(C)}\Bigg].
\end{aligned}$$

Term (A) is a typical Bregman divergence term from the classical FTRL/OMD analysis and depends on the local norm of the regularizer. Lemma 1 directly gives

$$\mathbb{E}[(A)] \leq \sum_{t=1}^n \sqrt{k}/\sqrt{t} \leq 2\sqrt{kn}\,.$$

Term (C) can also be bounded by standard techniques. Lemma 2 gives us

$$(C) \leq 2\sqrt{kn} + \eta_n^{-1} \log(k) \,.$$

Term (B) requires a novel analysis, which is presented in Lemma 3. This allows to bound the second term by

$$\mathbb{E}[(B)] \leq \sum_{t=1}^{n} \min\{1, \eta_t \mathfrak{d}_t^\rho\} \,.$$

Combining everything finishes the proof. □

Support lemmas for the proof of Theorem 3 The proofs for all the support lemmas are given in the supplementary material, Section 7.4. The first Lemma is a small modification of the classical result that bounds the Bregman divergence by the local norm of the regularizer. We show that we can bound the local norm by the contribution of the Tsallis entropy.

Lemma 1. *For any t it holds that*

$$\mathbb{E}\left[\overline{F}_t^*(-\hat{L}_t^{obs} - \hat{\ell}_t) - \overline{F}_t^*(-\hat{L}_t^{obs}) + \langle x_t, \hat{\ell}_t \rangle\right] \leq \frac{\sqrt{k}}{\sqrt{t}} \,.$$

The second Lemma bounds the so-called "penalty" term coming from the regularization penalty. It appears in almost identical form in the literature (Lattimore and Szepesvári, 2019, Exercise 28.12).

Lemma 2. *For any non-increasing learning rate η_t, it holds that*

$$\sum_{t=1}^{n} \left(\overline{F}_t^*(-\hat{L}_t^\rho) - \overline{F}_t^*(-\hat{L}_{t+1}^\rho) - \langle \mathbf{e}_{i^*}, \hat{\ell}_t \rangle \right) \leq 2\sqrt{kn} + \eta_n^{-1} \log(k) \,.$$

The third quantity does not show up in the regular analysis without delays. We show that similarly to the Bregman divergence, it depends on the local norm of the regularizer. However, it is beneficial to use the norm of the negative entropy instead of the Tsallis entropy.

Lemma 3. *For any t it holds that*

$$\mathbb{E}\Bigg[\overline{F}_t^*(-\hat{L}_t^{obs}) - \overline{F}_t^*(-\hat{L}_t^{obs} - \hat{\ell}_t) - \overline{F}_t^*(-\hat{L}_t^\rho) + \overline{F}_t^*(-\hat{L}_{t+1}^\rho)\Bigg] \leq \min\{1, \eta_t \tilde{\mathfrak{d}}_t^\rho\} \,.$$

5.2 Refined regret bound

The reason why it is beneficial to consider permutations in the analysis is the following lemma.

t	1	2	3	4	5	6	7	8	9	10
d_t	9	0	6	0	5	0	0	0	0	0
$\mathfrak{d}_t^{\rho_0}$	0	1	1	2	2	3	3	3	3	2
t	2	3	4	5	6	7	8	9	10	1
d_t	0	6	0	5	0	0	0	0	0	9
$\mathfrak{d}_t^{\rho_1}$	0	0	1	1	2	2	2	2	1	9
t	2	4	5	6	7	8	9	3	10	1
d_t	0	0	5	0	0	0	0	6	0	9
$\mathfrak{d}_t^{\rho_2}$	0	0	0	1	1	1	1	6	1	9
t	2	4	6	7	8	9	3	10	5	1
d_t	0	0	0	0	0	0	6	0	5	9
$\mathfrak{d}_t^{\rho_3}$	0	0	0	0	0	0	5	0	6	9

Figure 1: An iterative construction of the permutation in Lemma 4. Colored columns are elements in S.

Lemma 4. *For any $S \subset [n]$ there exists a dependency preserving permutation ρ, such that*

$$\forall t \in \bar{S}: \quad \mathfrak{d}_t^\rho = \sum_{s:s<t} \mathbb{I}\{s \in \bar{S}\} \mathbb{I}\{s + d_s \geq t\} \,.$$

Furthermore, this implies $\sum_{t\in\bar{S}} \mathfrak{d}_t^\rho \leq \sum_{t\in\bar{S}} d_t$.

An iterative procedure for construction of ρ is given in the supplementary material, Section 7.1. The lemma allows to split the rounds into sets S and $\bar{S}$ and construct a permutation, so that the number of outstanding delays for rounds in $\bar{S}$ only depends on the delays in other rounds in $\bar{S}$, but not on rounds in S. Fig. 1 provides an example of construction of such a permutation. The lemma is particularly useful for splitting the rounds into a set S containing large delays and the complementary set $\bar{S}$ containing small delays. Then the lemma allows to "push" the contributions to the ρ-number of outstanding observations away from the elements in $\bar{S}$ to the elements in S. Skipping the rounds in S yields the highest benefit.

Combining Lemma 4 with Theorem 3 and a suitable learning rate leads directly to the bound

$$\mathfrak{R}_n \leq 4\sqrt{kn} + |S| + 2\sqrt{\sum_{t\in\bar{S}} d_s \log(k)} \,.$$

In the following proof, we show that the learning rate in Algorithm 2 brings us within a constant of the minimum of the above bound, $4\sqrt{kn} + \min_S(|S| + 2\sqrt{D_{\bar{S}} \log(k)})$.

From now on, let S be the set

$$S = \{t \in [n] \,|\, a_t^n = 0\} \,,$$

which is the set of rounds "skipped" by Algorithm 2, and let ρ be the associated permutation from Lemma 4. Since $(a_s^t)_{t=1,\dots,n}$ is non-increasing, we have for any $t \in \bar{S}$: $\mathfrak{d}_t^\rho \le \tilde{\mathfrak{d}}_t$. Furthermore, the following lemma bounds the magnitude of $|S|$:

Lemma 5. *For any sequence of delays d_t, Algorithm 2 satisfies*

$$|S| = \sum_{t=1}^{n} \mathbb{I}\{a_t^n = 0\} \le 2\sqrt{\tilde{\mathfrak{D}}_n \log(k)}\,.$$

The proof is provided in the supplementary material, Section 7.2.

Finally we have all the prerequisites to prove Theorem 2 .

Proof of Theorem 2. Using Theorem 3 and Lemma 5 with ρ constructed for S, we have

$$\begin{aligned}
\mathfrak{R}_n &\le 4\sqrt{kn} + \eta_n^{-1}\log(k) + \sum_{t=1}^{n} \min\{1, \eta_t \mathfrak{d}_t^\rho\} \\
&\le 4\sqrt{kn} + \eta_n^{-1}\log(k) + |S| + \sum_{t\in\bar{S}} \eta_t \tilde{\mathfrak{d}}_t \\
&\le 4\sqrt{kn} + 5\sqrt{\tilde{\mathfrak{D}}_n \log(k)}\,.
\end{aligned}$$

Now we need to control the term $\sqrt{\tilde{\mathfrak{D}}_n \log(k)}$. Let's consider the case $\tilde{\mathfrak{D}}_n \le 4\sqrt{\tilde{\mathfrak{D}}_n \log(k)}$, then $\sqrt{\tilde{\mathfrak{D}}_n \log(k)} \le 4\log(k)$ and we are done. Otherwise, define $\tilde{d}_t = \sum_{s=t+1}^{t+d_t} a_t^s$, i.e., the contribution of round t to the sum $\tilde{\mathfrak{D}}_n$. Then we can decompose

$$\begin{aligned}
\tilde{\mathfrak{D}}_n &= \sum_{s=1}^{n}\sum_{t<s} \mathbb{I}\{t + d_t > s\} a_t^s \\
&= \sum_{t=1}^{n}\sum_{s>t} \mathbb{I}\{t + d_t > s\} a_t^s \\
&= \sum_{t=1}^{n}\sum_{s=t+1}^{t+d_t} a_t^s = \sum_{t=1}^{n} \tilde{d}_t\,.
\end{aligned}$$

Any element $t \in \bar{S}$ satisfies

$$\tilde{d}_t \le \sqrt{\tilde{\mathfrak{D}}_t/\log(k)} \le \sqrt{\tilde{\mathfrak{D}}_n/\log(k)}\,,$$

while any element $t \in S$ satisfies

$$\begin{aligned}
\tilde{d}_t &\le \left\lceil \sqrt{\tilde{\mathfrak{D}}_t/\log(k)} \right\rceil \le \left\lceil \sqrt{\tilde{\mathfrak{D}}_n/\log(k)} \right\rceil \\
&\le \sqrt{\tilde{\mathfrak{D}}_n/\log(k)} + 1\,.
\end{aligned}$$

Therefore, we can bound for any $R \subset [n]$:

$$\begin{aligned}
\sum_{t\in\bar{R}} d_t &\ge \sum_{t\in\bar{R}} \tilde{d}_t \ge \tilde{\mathfrak{D}}_n - |R|\sqrt{\tilde{\mathfrak{D}}_n/\log(k)} - |S| \\
&\ge \tilde{\mathfrak{D}}_n - |R|\sqrt{\tilde{\mathfrak{D}}_n/\log(k)} - 2\sqrt{\tilde{\mathfrak{D}}_n \log(k)} \\
&\ge \frac{1}{2}\tilde{\mathfrak{D}}_n - |R|\sqrt{\tilde{\mathfrak{D}}_n/\log(k)}\,.
\end{aligned}$$

This implies that

$$\begin{aligned}
&\min_{R\subset[n]} |R| + \sqrt{\sum_{t\in\bar{R}} d_t \log(k)} \\
&\ge \min_{r\in[0,\frac{1}{2}\sqrt{\tilde{\mathfrak{D}}_n\log(k)}]} r + \sqrt{\frac{1}{2}\tilde{\mathfrak{D}}_n\log(k) - r\sqrt{\tilde{\mathfrak{D}}_n\log(k)}}\,.
\end{aligned}$$

The function is concave in r so the minimum is achieved at one of the endpoints of the interval, which happens to be $r = \frac{1}{2}\sqrt{\tilde{\mathfrak{D}}_n\log(k)}$ for which the function equals $\frac{1}{2}\sqrt{\tilde{\mathfrak{D}}_n\log(k)}$. Hence, we have shown

$$\sqrt{\tilde{\mathfrak{D}}_n\log(k)} \le 2\min_{R\subset[n]} \left(|R| + \sqrt{\sum_{s\in\bar{R}} d_s \log(k)} \right)\,,$$

which concludes the proof. □

6 Discussion

We confirmed an open conjecture from Cesa-Bianchi et al. (2016) by presenting a simple FTRL algorithm for adversarial bandits with arbitrary delays and proving regret upper bound that matches the lower bound within constants. Furthermore, we proposed a refined tuning of the learning rate that achieves even tighter regret bound for highly unbalanced delays. We strictly improve on the state-of-the-art bounds and present the first anytime result requiring no doubling, skipping, or advance information about the delays.

If the delays are all 0, then our algorithm reduces to the Tsallis-INF algorithm of Zimmert and Seldin (2019), which has been proven to be simultaneously optimal in both the stochastic and the adversarial setting. We conjecture that the algorithm presented in this paper is capable of obtaining logarithmic regret in the stochastic setting, but leave the analysis for future work.

Another open question is the tightness of our adaptive bound $\mathcal{O}(\sqrt{kn} + \min_{S\subset[n]}(|S| + \sqrt{D_S \log(k)}))$. We conjecture that for a fixed set of delays $\{d_1, \dots, d_n\}$ which the adversary is allowed to permute without changing the magnitudes, the upper bound is actually tight.

Acknowledgements

We would like to thank András György and Tobias Sommer Thuner for fruitful discussions. The work was partly supported by the Independent Research Fund Denmark, grant number 9040-00361B.

References

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. Online exp3 learning in adversarial bandits with delayed feedback. In *NIPS*, pages 11345–11354, 2019.

Sébastien Bubeck, Michael B. Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2018.

Nicolò Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2016.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. In *Conference On Learning Theory*, pages 750–773, 2018.

Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press (preprint), 2019.

Haipeng Luo, Chen-Yu Wei, and Kai Zheng. Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Roman Pogodin and Tor Lattimore. On first-order bounds, variance and gap-dependent bounds for adversarial bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2019.

Yevgeny Seldin, Peter L. Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.