

Nonstochastic Bandits with Composite Anonymous Feedback

Nicolò Cesa-Bianchi, Tommaso R Cesari, Roberto Colomboni, Claudio Gentile, Yishay Mansour

▶ **To cite this version:**

Nicolò Cesa-Bianchi, Tommaso R Cesari, Roberto Colomboni, Claudio Gentile, Yishay Mansour. Nonstochastic Bandits with Composite Anonymous Feedback. 2021. hal-03466606

HAL Id: hal-03466606
https://hal.science/hal-03466606

Preprint submitted on 6 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Nonstochastic Bandits with Composite Anonymous Feedback*

Nicolò Cesa-Bianchi nicolo.cesa-bianchi@unimi.it
Dept. of Computer Science and DSRC, Università degli Studi di Milano, Italy

Tommaso Cesari tommaso.cesari@tse-fr.eu
Institut de Mathématiques de Toulouse (IMT), Paul Sabatier University (UT3), Toulouse, France
Toulouse School of Economics (TSE), Toulouse, France

Roberto Colomboni roberto.colomboni@unimi.it
Istituto Italiano di Tecnologia (IIT), Genova, Italy
Università degli Studi di Milano, Italy

Claudio Gentile cgentile@google.com
Google Research, New York, USA

Yishay Mansour mansour@tau.ac.il
Tel Aviv University, Tel Aviv, Israel
Google research, Tel Aviv, Israel

Editor:

Abstract

We investigate a nonstochastic bandit setting in which the loss of an action is not immediately charged to the player, but rather spread over the subsequent rounds in an adversarial way. The instantaneous loss observed by the player at the end of each round is then a sum of many loss components of previously played actions. This setting encompasses as a special case the easier task of bandits with delayed feedback, a well-studied framework where the player observes the delayed losses individually.

Our first contribution is a general reduction transforming a standard bandit algorithm into one that can operate in the harder setting: We bound the regret of the transformed algorithm in terms of the stability and regret of the original algorithm. Then, we show that the transformation of a suitably tuned FTRL with Tsallis entropy has a regret of order $\sqrt{(d+1)KT}$, where d is the maximum delay, K is the number of arms, and T is the time horizon. Finally, we show that our results cannot be improved in general by exhibiting a matching (up to a log factor) lower bound on the regret of any algorithm operating in this setting.

Keywords: Multi-armed bandits, non-stochastic losses, composite losses, delayed feedback, online learning,

1. Introduction

Multiarmed bandits, originally proposed for managing clinical trials, are now routinely applied to a variety of other tasks, including computational advertising, e-commerce, and beyond. Typical examples of e-commerce applications include content recommendation

*. A preliminary version of this paper appeared in the Proceedings of the 31st Conference on Learning Theory (COLT). PMLR 75:750-773, 2018.

systems, like the recommendation of products to visitors of merchant websites and social media platforms. A common pattern in these applications is that the response elicited in a user by the recommendation system is typically not instantaneous, and might occur some time in the future, well after the recommendation was issued. This delay, which might depend on several unknown factors, implies that the reward obtained by the recommender at time t can actually be seen as the combined effect of many previous recommendations to that user.

The more specific scenario of bandits with delayed rewards has been investigated in the literature under the assumption that the contributions of past recommendations to the combined reward is individually discernible —see, e.g., (Neu et al., 2010; Joulani et al., 2013; Cesa-Bianchi et al., 2016; Vernade et al., 2017). Pike-Burke et al. (2018) revisited the problem of bandits with delayed feedback under the more realistic assumption that only the combined reward is available to the system, while the individual reward components remain unknown. This model captures a much broader range of practical settings where bandits are successfully deployed. Consider for example an advertising campaign which is spread across several channels simultaneously (e.g., radio, tv, web, social media). A well-known problem faced by the campaign manager is to disentangle the contribution of individual ads deployed in each channel from the overall change in sales. Pike-Burke et al. (2018) formalized this harder delayed setting in a bandit framework with stochastic rewards, where they introduced the notion of *delayed anonymous feedback* to emphasize the fact that the reward received at any point in time is the sum of rewards of an unknown subset of past selected actions. More specifically, choosing action $I_t \in [K]$ at time t generates a stochastic reward $Y_t(I_t) \in [0,1]$ and a stochastic delay $\tau_t \in \{0,1,\dots\}$, where $\big\{Y_t(i), \tau_t\big\}_{i\in[K], t\in\mathbb{N}}$ is a family of independent random variables such that $Y_1(i), Y_2(i), \dots$ have a common distribution $\nu_Y(i)$ (for all arms $i \in [K]$) and $\tau_1, \tau_2, \dots$ have a common distribution ν_τ with expectation μ_τ. The delayed anonymous feedback assumption entails that the reward observed at time t by the algorithm is the sum of t components of the form $Y_s(I_s)\mathbb{I}\{\tau_s = t-s\}$ for $s \in \{1,\dots,t\}$. The main result in (Pike-Burke et al., 2018) is that, when the expected delay μ_τ is known, the regret is at most of order of $K\big((\ln T)/\Delta + \mu_\tau\big)$, where Δ is the suboptimality gap. This bound is of the same order as the corresponding bound for the setting where the feedback is stochastically delayed, but not anonymous (Joulani et al., 2013), and cannot be improved in general.

In this work, we study a bandit setting similar to delayed anonymous feedback, but with two important differences. First, we work in a nonstochastic bandit setting, where rewards (or losses, in our case) are generated by some unspecified deterministic mechanism. Second, we relax the assumption that the loss of an action is charged to the player at a single instant in the future. More precisely, we assume that the loss for choosing an action at time t is adversarially spread over at most $d+1$ consecutive time steps $t, t+1, \dots, t+d$. Hence, the loss observed by the player at time t is a *composite loss*, that is, the sum of $(d+1)$-many loss components $\ell_t^{(0)}(I_t), \ell_{t-1}^{(1)}(I_{t-1}), \dots, \ell_{t-d}^{(d)}(I_{t-d})$, where $\ell_{t-s}^{(s)}(I_{t-s})$ defines the s-th loss component from the selection of action I_{t-s} at time $t-s$. Note that in the special case when $\ell_t^{(s)}(i) = 0$ for all $s \neq d_t$, and $\ell_t^{(d_t)}(i) = \ell_t(i)$, we recover the model of nonstochastic bandits with delays $d_1, d_2, \dots \leq d$ (which, in particular, reduces to the standard nonstochastic bandits when $d = 0$). Our setting, which we call *composite anonymous feedback*, can accomodate scenarios where actions have a lasting effect which combines additively over

time. Online businesses provide several use cases for this setting. For instance, an impression that results in an immediate clickthrough, later followed by a conversion, or a user that interacts with a recommended item —such as media content— multiple times over several days, or the free credit assigned to a user of a gambling platform which might not be used all at once.

Our main contribution is a general reduction technique (Composite Loss Wrapper, or CoLoWrapper, Algorithm 1) turning a base nonstochastic bandit algorithm into one operating within the composite anonymous feedback setting. We then show that the regret of CoLoWrapper can be upper bounded in terms of the stability and the regret of the base algorithm (Theorem 4). Choosing as a base algorithm Follow the Regularized Leader (FTRL) with Tsallis entropy, Theorem 4 gives an upper bound of order $\sqrt{(d+1)KT}$ on the regret of nonstochastic bandits with composite anonymous feedback (Corollary 7), where $d \geq 0$ is a known upper bound on the delay, K is the number of actions, and T is the time horizon. This result relies on a nontrivial stability analysis of FTRL with Tsallis entropy that could be of independent interest (Theorem 5). Finally, we show the optimality of the $\sqrt{(d+1)KT}$ rate by proving a matching lower bound (up to a logarithmic factor, Theorem 10). In particular, this shows that, in the nonstochastic case with delay d, anonymous feedback is strictly harder than nonanonymous feedback, whose minimax regret was characterized by Cesa-Bianchi et al. (2016) as $\sqrt{(d+K)T}$. See the table below for a summary of results for nonstochastic K-armed bandits (all rates are optimal ignoring logarithmic factors).

NO DELAY	DELAYED FEEDBACK	ANONYMOUS COMPOSITE FEEDBACK
$\sqrt{KT}$ (Auer et al., 2002)	$\sqrt{(d+K)T}$ (Cesa-Bianchi et al., 2016)	$\sqrt{(d+1)KT}$ (this paper)

We now give an idea of the proof techniques. Similar to (Pike-Burke et al., 2018), we play the same action for a block of at least $2d+1$ time steps, hence the feedback we get in the last $d+1$ steps contains only loss components pertaining to the same action, so that we can estimate in those steps the "true loss" of that action. Unfortunately, although the original losses are in $[0,1]$, the composite losses can be as large as $d+1$ (a composite loss sums $d+1$ loss components, and each component can be as large as 1). This causes a corresponding scaling in the regret, compromising optimality. However, we observe that the total composite loss relative to the same action over any $d+1$ consecutive steps can be at most $2d+1$ (Lemma 2). Hence, we can normalize the total composite loss relative to the same action over $d+1$ consecutive steps, simply dividing by $2d+1$, obtaining an average loss in the range $[0,1]$. This idea leads to the right dependence on d in the regret. The last problem is how to avoid suffering a big regret in the first d steps of each block, where the composite losses mix loss components belonging to more than one action. We solve this issue by borrowing an idea of Dekel et al. (2014b). We build blocks with random endpoints so that their length is (always) at least $2d+1$ and (on average) not much bigger. This random positioning and length of the blocks is the key to prevent the oblivious adversary from causing a large regret in the first half of each block. Moreover, as we prove in Theorem 4, if the distribution over actions maintained by the base algorithm is *stable* (Definition 3), then the algorithm is not significantly affected by the uncertainty in the positioning of the blocks. Extending our results to the case where d is unknown, Wang et al. (2021) show a

regret bound of order $T^{2/3}$. When d is known, however, their analysis does not guarantee our faster $\sqrt{T}$ rate.

Further related work Online learning with delayed feedback was studied in the full information (non-bandit) setting by Weinberger and Ordentlich (2002); Mesterharm (2005); Langford et al. (2009); Joulani et al. (2013); Quanrud and Khashabi (2015); Khashabi et al. (2016); Joulani et al. (2016); Garrabrant et al. (2016); Mann et al. (2019), see also (Shamir and Szlak, 2017) for an interesting variant. The bandit setting with delay was investigated in (Neu et al., 2010; Joulani et al., 2013; Mandel et al., 2015; Cesa-Bianchi et al., 2016; Vernade et al., 2017; Pike-Burke et al., 2018; Li et al., 2019; Thune et al., 2019; Zimmert and Seldin, 2020; Vernade et al., 2020; Ito et al., 2020; Gael et al., 2020; Agrawal and Tulabandula, 2020). Our delayed composite loss function generalizes the composite loss function setting of Dekel et al. (2014a)—see the discussion at the end of Section 2 for details—and is also related to the notion of loss functions with memory. This latter setting has been investigated, e.g., by Arora et al. (2012), who showed how to turn an online algorithm with regret guarantee of $\mathcal{O}(T^q)$ into one attaining $\mathcal{O}(T^{1/(2-q)})$-policy regret, also adopting a blocking scheme. A more recent paper in this direction is (Anava et al., 2015), where the authors considered a more general loss framework than ours, though with the benefit of counterfactual feedback, in that the algorithm is aware of the loss it would incur had it played any sequence of d decisions in the previous d rounds, thereby making their results incomparable to ours.

2. Preliminaries

We denote the set of positive integers by $\mathbb{N}$ and the set of integers by $\mathbb{Z}$. For all $n \in \mathbb{N}$ we denote the set $\{1, \dots, n\}$ of the first n integers by $[n]$. We will use the handy convention that, if $(c_t)_{t\in\mathbb{Z}} \subset \mathbb{R}$ and $m, n \in \mathbb{Z}$ are such that $m > n$, then $\sum_{t=m}^{n} c_t = 0$ and $\prod_{t=m}^{n} c_t = 1$. For any $x \in \mathbb{R}$, we denote its positive part $\max\{x, 0\}$ by x^+.

We start by considering a nonstochastic multiarmed bandit problem on K actions with oblivious losses in which the loss $\ell_t(i) \in [0, 1]$ at time t of an action $i \in [K]$ is defined by the sum

$$\ell_t(i) \coloneqq \sum_{s=0}^{d} \ell_t^{(s)}(i)$$

of $(d + 1)$-many components $\ell_t^{(s)}(i) \geq 0$ for $s \in \{0, \dots, d\}$. Let I_t denote the action chosen by the player at the beginning of round t. If $I_t = i$, then the player incurs loss $\ell_t^{(0)}(i)$ at time t, loss $\ell_t^{(1)}(i)$ at time $t + 1$, and so on until time $t + d$. Yet, what the player observes at time t is only the combined loss incurred at time t, which is the sum

$$\ell_t^{(0)}(I_t) + \ell_{t-1}^{(1)}(I_{t-1}) + \cdots + \ell_{t-d}^{(d)}(I_{t-d})$$

of the past $d + 1$ loss contributions, where $\ell_t^{(s)}(i) = 0$ for all i and s when $t \leq 0$. Then, we define the d-delayed *composite* loss at time t of a sequence of $d+1$ actions $i_{t-d}, \dots, i_t \in [K]$ as

$$\ell_t^{\circ}(i_{t-d}, \dots, i_t) \coloneqq \sum_{s=0}^{d} \ell_{t-s}^{(s)}(i_{t-s}) \,. \tag{1}$$

With this notation, the d-delayed composite anonymous feedback assumption states that what the player observes at the end of each round t is only the composite loss $\ell_t^\circ(I_{t-d}, \dots, I_t)$. The goal of the algorithm is to bound its regret R_T against the best fixed action in hindsight,

$$R_T \coloneqq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^\circ(I_{t-d}, \dots, I_t)\right] - \min_{i\in[K]} \sum_{t=1}^{T} \ell_t^\circ(i, \dots, i) \ .$$

We define the regret in terms of the composite losses ℓ_t° rather than the true losses ℓ_t because in our model ℓ_t° is what the algorithm pays overall on round t. It is easy to see that a bound on R_T implies a bound on the more standard notion of regret $\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_k \sum_{t=1}^{T} \ell_t(k)$ up to an additive term of at most $\mathcal{O}(d)$.

Our setting generalizes the composite loss function setting of Dekel et al. (2014a). Specifically, the linear composite loss function therein can be seen as a special case of the composite loss (1) once we remove the superscripts s from the loss function components. In fact, in the linear case, the feedback in (Dekel et al., 2014a) allows one to easily reconstruct each individual loss component in a recursive manner. This is clearly impossible in our more involved scenario, where the new loss components that are observed in round t need not have occurred in past rounds.

3. The CoLoWrapper Algorithm

Our Composite Loss Wrapper algorithm (Algorithm 1) takes as input a standard K-armed bandit algorithm A and a Boolean sequence $\boldsymbol{B}$. The base algorithm A operates on standard (noncomposite) losses with values in $[0, 1]$, producing probability distributions $\boldsymbol{q}_1, \boldsymbol{q}_2, \dots$ over the action set $[K]$. The wrapper calls the base algorithm A only in a subset of rounds determined by the Boolean sequence $\boldsymbol{B}$, which we call update rounds.

Definition 1 (Update round) *We say that $t \in \mathbb{N}$ is an update round with respect to a Boolean sequence $\boldsymbol{B} = (b_t)_{t\in\mathbb{N}} \subset \{0, 1\}^{\mathbb{N}}$ if $t \geq 2d + 1$ and $b_t \prod_{s=1}^{2d}(1 - b_{t-s}) = 1$.*

Note that if $d > 0$, the condition is equivalent to $b_t = 1$, and $b_{t-1} = \ldots = b_{t-2d} = 0$. If $d = 0$, by our convention, the condition is equivalent to $b_t = 1$.

To help understand our algorithm, we will also define two other types of rounds. We say that t is a *draw round* if $t = 1$ or the previous round $t - 1$ was an update round. If t is not a draw round, we say that it is a *stay round*. Note that, if $d = 0$, both draw and stay rounds can be update rounds, while if $d \geq 1$, only stay rounds can be update rounds.

The CoLoWrapper algorithm proceeds in blocks of (random) length of at least $2d + 1$ rounds in which it constantly plays the same action (Figure 1). Blocks in Algorithm 1 are counted by variable n_t. Each block n_t consists of a draw round followed by ($2d$ or more) stay rounds, with the last round of the block being also an update round. During a draw round t, CoLoWrapper uses its current distribution $\boldsymbol{p}_t$ to draw and play an action I_t. During stay rounds, it keeps playing the action that was drawn during the latest draw round. After playing the action I_t for the current round t, if t is an update round, CoLoWrapper asks the base algorithm A to make an update of its base distribution $\boldsymbol{q}_{n_t} \to \boldsymbol{q}_{n_t+1}$ as if A played action I_t and observed as the loss of I_t the quantity $\frac{1}{2d+1} \sum_{s=t-d}^{t} \ell_s^\circ(I_{s-d}, \dots, I_s)$. Then,

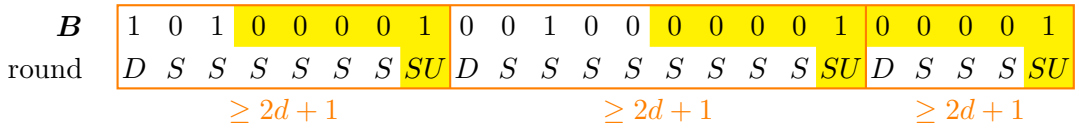


Figure 1: Sequence of rounds the algorithm is undergoing when $d = 2$. The top line contains the values of the Boolean sequence $\boldsymbol{B} = (B_t)_{t\in\mathbb{N}}$. The bottom line shows the corresponding types of rounds: each block begins with a (D)raw round, followed by a variable number of (S)tay rounds, the last of which is also an (U)pdate round. Since a round t is an update round only if $B_t = 1$ and $B_s = 0$ for the $2d$ previous rounds s, the length of each block is at least $2d + 1$.

Algorithm 1: CoLoWrapper (Composite Loss Wrapper)

input: Base K-armed bandit algorithm A and Boolean sequence $\boldsymbol{B}$
initialize: let $n_0 = 0$ and $\boldsymbol{q}_1$ be the initial distribution over $[K]$ of A
for *round* $t = 1, 2, \ldots$ **do**
  if *either* $t = 1$ *or* $t - 1$ *was an update round (w.r.t.* $\boldsymbol{B}$*)* **then**
    let $n_t = n_{t-1} + 1$, $\boldsymbol{p}_t = \boldsymbol{q}_{n_t}$, and draw $I_t \sim \boldsymbol{p}_t$ `// draw`
  else
    let $n_t = n_{t-1}$, $\boldsymbol{p}_t = \boldsymbol{p}_{t-1}$, and $I_t = I_{t-1}$ `// stay`
  play I_t and observe loss $\ell_t^\circ(I_{t-d}, \ldots, I_t)$
  if t *is an update round (w.r.t.* $\boldsymbol{B}$*)* **then** `// update`
    feed A with arm I_t and loss $\frac{1}{2d+1}\sum_{s=t-d}^{t} \ell_s^\circ(I_{s-d}, \ldots, I_s)$
    use the update rule $\boldsymbol{q}_{n_t} \to \boldsymbol{q}_{n_t+1}$ of A to obtain a new base distribution $\boldsymbol{q}_{n_t+1}$

the block ends and the distribution of CoLoWrapper at the beginning of the next block $n_{t+1} = n_t + 1$ is $\boldsymbol{p}_{t+1} = \boldsymbol{q}_{n_t+1}$.

Note that if t is an update round, the quantity $\frac{1}{2d+1}\sum_{s=t-d}^{t} \ell_s^\circ(I_{s-d}, \ldots, I_s)$ that is fed back to A relates only to the current action I_t, because blocks contain at least $2d+1$ rounds and the same action is played in all of them.

The following lemma shows that this quantity is indeed in $[0, 1]$, so that it is a legitimate feedback to pass to the base algorithm A.

Lemma 2 *For all* $t \geq 2d + 1$ *and* $i \in [K]$,

$$\sum_{\tau=t-d}^{t} \ell_\tau^\circ(i, \ldots, i) \leq 2d + 1 \,.$$

Proof For all $t \geq 2d + 1$ and $i \in [K]$,

$$\sum_{\tau=t-d}^{t} \ell_\tau^\circ(i, \ldots, i) = \sum_{\tau=t-d}^{t}\sum_{s=0}^{d} \ell_{\tau-s}^{(s)}(i) = \sum_{s=0}^{d}\sum_{\tau=t-d}^{t} \ell_{\tau-s}^{(s)}(i) = \sum_{s=0}^{d}\sum_{\rho=t-d-s}^{t-s} \ell_\rho^{(s)}(i)$$

$$\leq \sum_{s=0}^{d} \sum_{\rho=t-2d}^{t} \ell_\rho^{(s)}(i) = \sum_{\rho=t-2d}^{t} \sum_{s=0}^{d} \ell_\rho^{(s)}(i) = \sum_{\rho=t-2d}^{t} \ell_\rho(i) \leq t - (t-2d) + 1 = 2d+1 \ .$$

■

As a final remark, we point out that, albeit the algorithm is parameterized with an entire sequence $\boldsymbol{B}$, at each time t, it does not require the knowledge of the sequence at future times $t+1, t+2, \ldots$. This implies in particular that these Boolean values could be produced and fed to CoLoWrapper in an on-line fashion.

4. Upper Bound

We begin by formalizing the notion of stability (of the base algorithm), in terms of which we express the performance of the CoLoWrapper algorithm.

Definition 3 (ξ-stability) *Let $\xi > 0$, A be a K-armed bandit algorithm, and $(\boldsymbol{q}_n)_{n\in\mathbb{N}}$ be the (random) sequence of probability distributions over actions $[K]$ produced by A during a run over rounds $\{1, 2, \ldots\}$. We say that A is ξ-stable if for any round n, we have*

$$\mathbb{E}\left[\sum_{i\in[K]} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big)^+\right] \leq \xi \ .$$

We can now state our main result of this section.

Theorem 4 *If we run CoLoWrapper with a ξ-stable base K-armed bandit algorithm A and an i.i.d. sequence $\boldsymbol{B} = (B_t)_{t\in\mathbb{N}}$ of Bernoulli random variables with bias $\beta \in (0,1)$ (independent of the randomization of A), then, for any time horizon $T \geq 2d+1$, the regret R_T satisfies*

$$R_T \leq 2d + \frac{2d+1}{d+1}\left(3d + 2d\beta(1-\beta)^{2d}\xi T + \frac{1}{\beta(1-\beta)^{2d}}\mathcal{R}_{\lfloor T/(2d+1)\rfloor}\right)$$

where $\mathcal{R}_{\lfloor T/(2d+1)\rfloor}$ is the worst-case regret after $\lfloor T/(2d+1)\rfloor$ rounds of A (for an adversarial setting with $[0,1]$-valued losses).

Proof Fix an arbitrary horizon $T \geq 2d+1$ and an arm $i^\star \in [K]$. Let, for all [1] $t \geq 2d+1$,

$$c_t \coloneqq \mathbb{E}\big[\ell_t^\circ(I_{t-d}, \ldots, I_t) - \ell_t^\circ(i^\star, \ldots, i^\star)\big] \ , \qquad a \coloneqq 2d+1 \ , \qquad b \coloneqq T \ .$$

Applying the elementary identity in Lemma 11 (Appendix A) in step $(*)$ below, we obtain

$$R_T = \sum_{t=1}^{2d} c_t + \sum_{t=2d+1}^{T} c_t \leq \sum_{t=1}^{2d} \mathbb{E}\big[\ell_t(I_t)\big] + \sum_{t=2d+1}^{T} c_t \leq 2d + \sum_{t=2d+1}^{T} c_t$$

1. Here, we refer to the infinite sequence of losses $(\ell_t^{(s)})_{s\in\{0,\ldots,d\},t\in\mathbb{N}}$ (and the respective composite losses $(\ell_t^\circ)_{t\in\mathbb{N}}$). If the problem is formalized only with a finite sequence $(\ell_t^{(s)})_{s\in\{0,\ldots,d\},t\in[T]}$, it is sufficient to define an arbitrary sequence $(\ell_t^{(s)})_{s\in\{0,\ldots,1\},t>T}$ with $\ell_t^{(s)} \geq 0$ and $\sum_{s=0}^{d} \ell_t^{(s)} \leq 1$ (and the corresponding composite losses $(\ell_t^\circ)_{t>T}$) and proceed as we do. This trick is needed to invoke Lemma 2, for which it is handy to sum d rounds into the future.

$$
\begin{aligned}
&= 2d + \frac{1}{d+1}\left(\sum_{t=a-d}^{a-1}(t-a+d+1)\,c_t + (d+1)\sum_{t=a}^{b} c_t + \sum_{t=b+1}^{b+d}(b+d+1-t)\,c_t\right) \\
&\quad - \frac{1}{d+1}\sum_{t=a-d}^{a-1}(t-a+d+1)\,c_t - \frac{1}{d+1}\sum_{t=b+1}^{b+d}(b+d+1-t)\,c_t \\
&\overset{(*)}{=} 2d + \frac{1}{d+1}\sum_{\tau=a}^{b+d}\sum_{t=\tau-d}^{\tau} c_t - \frac{1}{d+1}\sum_{t=a-d}^{a-1}(t-a+d+1)c_t - \frac{1}{d+1}\sum_{t=b+1}^{b+d}(b+d+1-t)c_t \\
&\le 2d + \frac{1}{d+1}\sum_{\tau=a}^{b+d}\sum_{t=\tau-d}^{\tau} c_t + \frac{1}{d+1}\sum_{t=a-d}^{a-1}(t-a+d+1)\ell_t^\circ(i^\star,\dots,i^\star) \\
&\quad + \frac{1}{d+1}\sum_{t=b+1}^{b+d}(b+d+1-t)\ell_t^\circ(i^\star,\dots,i^\star) =: (\blacktriangledown)
\end{aligned}
$$

Now, applying Lemma 2 in steps ($\circ$) below, we get

$$
\begin{aligned}
(\blacktriangledown) &\overset{(\circ)}{\le} 2d + 2\frac{2d+1}{d+1}d + \frac{1}{d+1}\sum_{\tau=a}^{b+d}\sum_{t=\tau-d}^{\tau} c_t \\
&= 2d + 2\frac{2d+1}{d+1}d + \frac{1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{t=\tau-d}^{\tau}\mathbb{E}\big[\ell_t^\circ(I_{t-d},\dots,I_t) - \ell_t^\circ(I_{\tau-2d},\dots,I_{\tau-2d})\big] \\
&\quad + \frac{2d+1}{d+1}\mathbb{E}\left[\sum_{\tau=2d+1}^{T+d}\frac{1}{2d+1}\sum_{t=\tau-d}^{\tau}\big(\ell_t^\circ(I_{\tau-2d},\dots,I_{\tau-2d}) - \ell_t^\circ(i^\star,\dots,i^\star)\big)\right] \\
&\overset{(\circ)}{\le} 2d + 3\frac{2d+1}{d+1}d + \frac{1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{t=\tau-d}^{\tau}\mathbb{E}\big[\ell_t^\circ(I_{t-d},\dots,I_t) - \ell_t^\circ(I_{\tau-2d},\dots,I_{\tau-2d})\big] \\
&\quad + \frac{2d+1}{d+1}\mathbb{E}\left[\sum_{\tau=2d+1}^{T}\frac{1}{2d+1}\sum_{t=\tau-d}^{\tau}\big(\ell_t^\circ(I_{\tau-2d},\dots,I_{\tau-2d}) - \ell_t^\circ(i^\star,\dots,i^\star)\big)\right] \\
&=: 2d + 3\frac{2d+1}{d+1}d + (\mathrm{I}) + \frac{2d+1}{d+1}\times(\mathrm{II})\ .
\end{aligned}
$$

We upper bound the two terms (I) and (II) separately. First, let $\mathcal{U}$ be the (random) set of update rounds

$$
\mathcal{U} \coloneqq \big\{\tau \in [T] : \tau \text{ is an update round (w.r.t. } \boldsymbol{B})\big\}\ .
$$

For the first term (I), we have

$$
\begin{aligned}
(\mathrm{I}) &= \frac{1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{i\in[K]}\mathbb{E}\left[\sum_{t=\tau-d}^{\tau}\sum_{s=0}^{d}\ell_{t-s}^{(s)}(i)\big(\boldsymbol{p}_{t-s}(i) - \boldsymbol{p}_{\tau-2d}(i)\big)\right] \\
&\le \frac{1}{d+1}\sum_{\tau=2d+1}^{T+d}\sum_{i\in[K]}\mathbb{E}\left[\max_{t\in\{\tau-d,\dots,\tau\},s\in\{0,\dots,d\}}\big(\boldsymbol{p}_{t-s}(i) - \boldsymbol{p}_{\tau-2d}(i)\big)\right]\sum_{t=\tau-d}^{\tau}\ell_t^\circ(i,\dots,i)
\end{aligned}
$$

$$\overset{(\circ)}{\leq} \frac{2d+1}{d+1} \sum_{\tau=2d+1}^{T+d} \sum_{i\in[K]} \mathbb{E}\left[\max_{t\in\{\tau-d,\dots,\tau\},s\in\{0,\dots,d\}} \left(\boldsymbol{p}_{t-s}(i) - \boldsymbol{p}_{\tau-2d}(i)\right)\right]$$

$$\overset{(\heartsuit)}{=} \frac{2d+1}{d+1} \sum_{\tau=2d+1}^{T+d} \sum_{i\in[K]} \mathbb{E}\left[\mathbb{I}\left\{\bigcup_{\sigma=\tau-2d}^{\tau-1} \{\sigma\in\mathcal{U}\}\right\} \left(\boldsymbol{p}_{\tau}(i) - \boldsymbol{p}_{\tau-2d}(i)\right)^+\right]$$

$$= \frac{2d+1}{d+1} \sum_{\tau=2d+1}^{T+d} \sum_{i\in[K]} \sum_{\sigma=\tau-2d}^{\tau-1} \mathbb{E}\left[\mathbb{I}\{\sigma\in\mathcal{U}\} \left(\boldsymbol{p}_{\tau}(i) - \boldsymbol{p}_{\tau-2d}(i)\right)^+\right] =: (\blacklozenge)$$

where $(\circ)$ follows by Lemma 2 together with the fact that the max in the previous line is always nonnegative (to see this, simply observe that picking $t=\tau-d$ and $s=d$ within the max makes $\boldsymbol{p}_{t-s}(i)=\boldsymbol{p}_{\tau-2d}(i)$), and $(\heartsuit)$ follows by the facts that the max in the previous line is always greater than or equal to zero, it can be strictly positive only if there is an update in a round σ with $\tau-2d\leq\sigma\leq\tau-1$, and there can be at most a single update in $2d+1$ consecutive time steps. Now, using the facts that $\boldsymbol{p}_\tau=\boldsymbol{q}_{n_\tau}$ and that $n_\tau=n_{\tau-2d}+1$ on the event $\{\sigma\in\mathcal{U}\}$, we get

$$(\blacklozenge) = \frac{2d+1}{d+1} \sum_{\tau=2d+1}^{T+d} \sum_{i\in[K]} \sum_{\sigma=\tau-2d}^{\tau-1} \mathbb{E}\left[\mathbb{I}\{\sigma\in\mathcal{U}\} \left(\boldsymbol{q}_{n_{\tau-2d}+1}(i) - \boldsymbol{q}_{n_{\tau-2d}}(i)\right)^+\right]$$

$$= \frac{2d+1}{d+1} \sum_{\tau=2d+1}^{T+d} \sum_{i\in[K]} \sum_{\sigma=\tau-2d}^{\tau-1} \sum_{n\in\mathbb{N}} \mathbb{E}\left[\mathbb{I}\{\sigma\in\mathcal{U}\}\mathbb{I}\{n_{\tau-2d}=n\} \left(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_{n}(i)\right)^+\right]$$

$$\overset{(\clubsuit)}{=} \frac{2d+1}{d+1} \sum_{\tau=2d+1}^{T+d} \sum_{\sigma=\tau-2d}^{\tau-1} \sum_{n\in\mathbb{N}} \mathbb{E}\big[\mathbb{I}\{\sigma\in\mathcal{U}\}\mathbb{I}\{n_{\tau-2d}=n\}\big]\, \mathbb{E}\left[\sum_{i\in[K]} \left(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_{n}(i)\right)^+\right]$$

$$\overset{(\spadesuit)}{\leq} \frac{2d+1}{d+1}\xi \sum_{\tau=2d+1}^{T+d} \sum_{\sigma=\tau-2d}^{\tau-1} \sum_{n\in\mathbb{N}} \mathbb{E}\big[\mathbb{I}\{\sigma\in\mathcal{U}\}\mathbb{I}\{n_{\tau-2d}=n\}\big]$$

$$= \frac{2d+1}{d+1}\,\xi \sum_{\tau=2d+1}^{T+d} \sum_{\sigma=\tau-2d}^{\tau-1} \mathbb{P}[\sigma\in\mathcal{U}] \leq \frac{2d+1}{d+1} 2d\beta(1-\beta)^{2d}\xi T$$

where $(\clubsuit)$ follows by the independence of the Bernoulli sequence $\boldsymbol{B}$ and the randomization of base algorithm A; $(\spadesuit)$ by the ξ-stability of A; and the last inequality by the fact that the probability that a time step σ is an update round is 0 if $\sigma\leq 2d$ and $\beta(1-\beta)^{2d}$ otherwise.

For the second term (II), set for brevity

$$r_\tau\colon [K]\to\mathbb{R}\,, \qquad i\mapsto \frac{1}{2d+1}\sum_{t=\tau-d}^{\tau} \left(\ell_t^\circ(i,\dots,i) - \ell_t^\circ(i^\star,\dots,i^\star)\right)$$

for all $\tau\geq 2d+1$. We first note that

$$\mathbb{E}\left[\sum_{\tau\in\mathcal{U}} r_\tau(I_\tau)\right] = \mathbb{E}\left[\sum_{\tau\in\mathcal{U}} r_\tau(I_{\tau-2d})\right] = \mathbb{E}\left[\sum_{\tau=2d+1}^{T} r_\tau(I_{\tau-2d}) B_\tau \prod_{s=1}^{2d}(1-B_{\tau-s})\right]$$

$$= \beta(1-\beta)^{2d} \sum_{\tau=2d+1}^{T} \mathbb{E}\big[r_\tau(I_{\tau-2d})\big]$$

where the first identity is a consequence of the fact that if $\tau \in \mathcal{U}$, then the action I_τ played by Algorithm 1 at round τ coincides with the actions played in the previous $2d$ rounds, i.e., $I_\tau = \cdots = I_{\tau-2d}$, while the last equality follows by the independence of $I_{\tau-2d}$ and the vector of Bernoulli random variables $(B_\tau, \dots, B_{\tau-2d})$. Thus,

$$\begin{aligned}
\text{(II)} &= \mathbb{E}\left[\sum_{\tau=2d+1}^{T} r_\tau(I_{\tau-2d})\right] = \frac{1}{\beta(1-\beta)^{2d}}\mathbb{E}\left[\sum_{\tau\in\mathcal{U}} r_\tau(I_\tau)\right] \\
&= \frac{1}{\beta(1-\beta)^{2d}}\mathbb{E}\left[\sum_{\tau\in\mathcal{U}} \frac{1}{2d+1}\sum_{t=\tau-d}^{\tau} \ell_t^\circ(I_\tau,\dots,I_\tau) - \sum_{\tau\in\mathcal{U}} \frac{1}{2d+1}\sum_{t=\tau-d}^{\tau} \ell_t^\circ(i^\star,\dots,i^\star)\right] \quad (2)
\end{aligned}$$

Now define for any $\tau \geq 2d+1$,

$$\tilde{\ell}_\tau \colon [K] \to [0,1]\ , \quad i \mapsto \frac{1}{2d+1}\sum_{t=\tau-d}^{\tau} \ell_t^\circ(i,\dots,i)\ .$$

(Note that $\tilde{\ell}_\tau(i) \in [0,1]$ for all $i \in [K]$ by Lemma 2.) Leveraging again the fact that $\tau \in \mathcal{U}$ implies $I_\tau = \cdots = I_{\tau-2d}$, yields, for each $\tau \in \mathcal{U}$,

$$\frac{1}{2d+1}\sum_{t=\tau-d}^{\tau} \ell_t^\circ(I_{t-d},\dots,I_t) = \frac{1}{2d+1}\sum_{t=\tau-d}^{\tau} \ell_t^\circ(I_\tau,\dots,I_\tau) = \tilde{\ell}_\tau(I_\tau)\ .$$

This shows that the loss Algorithm 1 feeds the base algorithm A (in line 8) at each update round τ is a bandit feedback (for A) for the arm I_τ with respect to the $[0,1]$-valued loss $\tilde{\ell}_\tau$. Therefore, by Equation (2) and the regret guarantees of the base algorithm, we obtain

$$\text{(II)} = \frac{1}{\beta(1-\beta)^{2d}}\mathbb{E}\left[\sum_{\tau\in\mathcal{U}} \tilde{\ell}_\tau(I_\tau) - \sum_{\tau\in\mathcal{U}} \tilde{\ell}_\tau(i^\star)\right] \leq \frac{1}{\beta(1-\beta)^{2d}}\mathcal{R}_{\lfloor T/(2d+1)\rfloor}$$

where in the last inequality we also used the monotonicity of the worst-case regret $\tau \mapsto \mathcal{R}_\tau$ and the fact that $|\mathcal{U}| \leq \left\lceil \frac{T-(2d+1)+1}{2d+1} \right\rceil \leq \left\lfloor \frac{T}{2d+1} \right\rfloor$.

In conclusion, we have

$$\begin{aligned}
R_T &\leq 2d + 3\frac{2d+1}{d+1}d + \text{(I)} + \frac{2d+1}{d+1}\cdot\text{(II)} \\
&\leq 2d + \frac{2d+1}{d+1}\left(3d + 2d\beta(1-\beta)^{2d}\xi T + \frac{1}{\beta(1-\beta)^{2d}}\mathcal{R}_{\lfloor T/(2d+1)\rfloor}\right)
\end{aligned}$$

hence concluding the proof. ■

We can derive corollaries for various algorithms using Theorem 4. Consider for instance as a base algorithm A the well-known Exp3 algorithm of Auer et al. (2002). It can be easily

shown that Exp3 is η-stable (where η is its learning rate, see Lemma 14 in Appendix C). Combining this fact with its worst-case regret bound $\mathcal{R}_N \leq \frac{\ln K}{\eta} + \frac{\eta}{2}KN$, we obtain that the horizon-T regret of the CoLoWrapper algorithm with $A = \text{Exp3}(\eta)$, $\eta = \sqrt{\frac{d \ln K}{KT}}$, and an i.i.d. sequence $\boldsymbol{B}$ of Bernoulli random variables with bias $\beta = \frac{1}{2d+1}$ (independent of the randomization of A), satisfies $R_T = \mathcal{O}\big(\sqrt{(d+1)KT \log K}\big)$.

Using Follow the Regularized Leader (FTRL) with $\frac{1}{2}$-Tsallis Entropy (Algorithm 2), we can remove the $\log K$ term in the above bound.[2]

Algorithm 2: Follow The Regularized Leader (FTRL) with (1/2)-Tsallis entropy

input: learning rate $\eta \geq 0$
initialize: $\widehat{L}_0 = 0$
for *round* $n = 1, 2, \ldots$ **do**
  Play action J_n drawn according to

$$\boldsymbol{q}_n \in \operatorname*{argmin}_{\boldsymbol{q} \in \Delta_K} \left(\sum_{i \in [K]} \widehat{L}_{n-1}(i) \boldsymbol{q}(i) - 2\eta \sum_{i \in [K]} \sqrt{\boldsymbol{q}(i)} \right)$$

  where Δ_K is the probability simplex in $\mathbb{R}^K$
  Observe loss $\ell_n(J_n)$ and update $\widehat{L}_n = \widehat{L}_{n-1} + \widehat{\ell}_n$, where

$$\widehat{\ell}_n(i) = \frac{\ell_n(i)}{\boldsymbol{q}_n(i)} \mathbb{I}\{J_n = i\} \qquad \forall i \in [K]$$

However, we still need to prove the stability of Algorithm 2. This is established by the following result, whose (non-trivial) proof is given in Appendix B.

Theorem 5 *Algorithm 2 run with any learning rate $\eta > 0$ is ξ-stable, with*

$$\xi \leq 2 \frac{1 + \ln K}{\eta}$$

The worst-case regret guarantees of FTRL with $\frac{1}{2}$-Tsallis entropy are as follows.[3]

Theorem 6 (Abernethy et al. 2015) *For each $N \in \mathbb{N}$, the worst-case regret $\mathcal{R}_N$ after N rounds of Algorithm 2 run with learning rate $\eta = \sqrt{\frac{N}{2}}$ (in an adversarial setting with $[0,1]$-valued losses) satisfies*

$$\mathcal{R}_N \leq 2\sqrt{2KN}$$

Combining Theorems 4 to 6, we obtain the following regret bound for composite losses.

2. The argmin where Algorithm 2 picks each distribution $\boldsymbol{q}_n$ is always a singleton if $\eta > 0$. This is a consequence of Lemma 12, in Appendix B. For the sake of convenience, we also allow the learning rate $\eta = 0$, corresponding to a (non-regularized) Follow The Leader algorithm. In this case, multiple minimizers could exist but, for the sake of our results, ties could be broken in any (measurable) way.
3. The analysis of Abernethy et al. (2015) is presented for a more general class of algorithms. A straightforward application of their Corollary 3.2 shows the validity of Theorem 6 for Algorithm 2.

Corollary 7 *For any time horizon $T \in \mathbb{N}$, if we run CoLoWrapper using:*

- *As A, Algorithm 2 with learning rate $\eta = \sqrt{\frac{1}{2}\lfloor T/(2d+1)\rfloor}$*
- *As $\boldsymbol{B}$, an i.i.d. sequence of Bernoulli random variables with bias $\beta = \frac{1}{2d+1}$ (independent of the randomization of A)*

then, its regret satisfies

$$R_T \le c\sqrt{(d+1)KT}$$

where $c = 2\sqrt{2}$ if $d = 0$, and $c = 28$ otherwise.

Proof If $d = 0$, then $\beta = 1$ and Algorithm 1 reduces to the base algorithm. The result in this case is therefore implied immediately by Theorem 6.

For the second part, we can assume without loss of generality that $T \ge 2d+1$, so that $\eta > 0$. Then, plugging $\xi \le 2\frac{1+\ln K}{\eta}$ (Theorem 5) and $\mathcal{R}_{\lfloor T/(2d+1)\rfloor} \le 2\sqrt{2K\lfloor T/(2d+1)\rfloor}$ (Theorem 6) into Theorem 4 gives the result. ■

5. Lower bound

In this section we derive a lower bound for bandits with composite anonymous feedback. We do that through a reduction from the setting of linear bandits (in the probability simplex) to our setting. This reduction allows us to upper bound the regret of a linear bandit algorithm in terms of (a suitably scaled version of) the regret of an algorithm in our setting. Since the reduction applies to any instance of a linear bandit problem, we can use a known lower bound for the linear bandit setting to derive a corresponding lower bound for our composite setting.

Let Δ_K be the probability simplex in $\mathbb{R}^K$. At each round t, an algorithm A for linear bandit optimization chooses an action $\boldsymbol{q}_t \in \Delta_K$ and suffers loss $\boldsymbol{\ell}_t^\top \boldsymbol{q}_t$, where $\boldsymbol{\ell}_t \in [0,1]^K$ is some unknown loss vector. The feedback observed by the algorithm at the end of round t is the scalar $\boldsymbol{\ell}_t^\top \boldsymbol{q}_t$. The regret suffered by algorithm A playing actions $\boldsymbol{q}_1, \dots, \boldsymbol{q}_T$ is

$$R_T^{\text{lin}} = \sum_{t=1}^T \boldsymbol{\ell}_t^\top \boldsymbol{q}_t - \min_{\boldsymbol{q}\in\Delta_K} \sum_{t=1}^T \boldsymbol{\ell}_t^\top \boldsymbol{q} = \sum_{t=1}^T \boldsymbol{\ell}_t^\top \boldsymbol{q}_t - \min_{i=1,\dots,K} \sum_{t=1}^T \boldsymbol{\ell}^\top \boldsymbol{e}_i \tag{3}$$

where $\boldsymbol{e}_1, \dots, \boldsymbol{e}_K$ are the elements of the canonical basis of $\mathbb{R}^K$ and we used the fact that a linear function on the simplex is minimized at one of the corners. Let $\mathcal{R}_T^{\text{lin}}(A, \Delta_K)$ denote the worst case regret (over the oblivious choice of $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T$) of algorithm A. Similarly, let $\mathcal{R}_T(A_d, K, d)$ be the worst case regret (over the oblivious choice of loss components $\ell_t^{(s)}(i)$ for all t, s, and i) of algorithm A_d for nonstochastic K-armed bandits with d-delayed composite anonymous feedback. For the sake of clarity, we assume below that the time horizon T is a multiple of $d+1$. If this is not the case, we can straightforwardly stop at the highest multiple of $d+1$ (smaller than T) up to paying an additive $\mathcal{O}(d+1)$ regret. Our reduction shows the following.

Lemma 8 *For any algorithm A_d for K-armed bandits with d-delayed composite anonymous feedback, there exists an algorithm A for linear bandits in Δ_K such that $\mathcal{R}_T(A_d, K, d) \ge (d+1)\,\mathcal{R}^{\text{lin}}_{T/(d+1)}(A, \Delta_K)$.*

Our reduction, described in detail in the proof of the above lemma (see Appendix D), essentially builds the probability vectors $\boldsymbol{q}_t$ played by A based on the empirical distribution of actions played by A_d during blocks of size $d+1$. Now, an additional lemma is needed (whose proof is also given in the Appendix D).

Lemma 9 *The regret of any algorithm A for linear bandits on Δ_K satisfies $\mathcal{R}_T^{\mathrm{lin}}(A, \Delta_K) = \widetilde{\Omega}\big(\sqrt{KT}\big)$.*

In the previous lemma, as well as the following result, the $\widetilde{\Omega}$ notation is only hiding a $\sqrt{\log T}$ denominator. Using the two lemmas above we can finally prove the lower bound.

Theorem 10 *For any algorithm A_d for K-armed bandits with d-delayed composite anonymous feedback, $\mathcal{R}_T(A_d, K, d) = \widetilde{\Omega}\big(\sqrt{(d+1)KT}\big)$.*

Proof Fix an algorithm A_d. Using the reduction of Lemma 8 gives an algorithm A such that $\mathcal{R}_T(A_d, K, d) \geq (d+1)\,\mathcal{R}^{\mathrm{lin}}_{T/(d+1)}(A, \Delta_K) = \widetilde{\Omega}\big(\sqrt{(d+1)KT}\big)$, where we used Lemma 9 with horizon $T/(d+1)$ to prove the $\widetilde{\Omega}$-equality. ■

Although the loss sequence used to prove the lower bound for linear bandits in the simplex is stochastic i.i.d., the loss sequence achieving the lower bound in our delayed setting is not independent due to the deterministic loss transformation in the proof of Lemma 8 (which is defined independently of the algorithm, thus preserving the oblivious nature of the adversary).

6. Conclusions

We have investigated the setting of d-delayed composite anonymous feedback as applied to nonstochastic bandits. Composite anonymous feedback lends itself to formalize scenarios where the actions performed by the online decision-maker produce long-lasting effects that combine additively over time. A general reduction technique was introduced that enables the conversion of a stable algorithm working in a standard bandit framework into one working in the composite feedback framework. Applying our reduction to the FTRL algorithm with Tsallis entropy, we obtain an upper bounded on the regret of order $\sqrt{(d+1)KT}$. We then showed the optimality of this rate (up to a logarithmic factor) relying on a lower bound for bandit linear optimization in the probability simplex.

Acknowledgments

NCB and RC were partially supported by the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR) and the EU Horizon 2020 ICT-48 research and innovation action under grant agreement 951847, project ELISE (European Learning and Intelligent Systems Excellence). The work of TC has benefited from the AI Interdisciplinary Institute ANITI, which is funded by the French "Investing for the Future – PIA3" program under the Grant agreement ANR-19-P3IA-0004. TC also acknowledges the support of the project BOLD from the French national research agency (ANR). YM was supported in part by 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), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University.

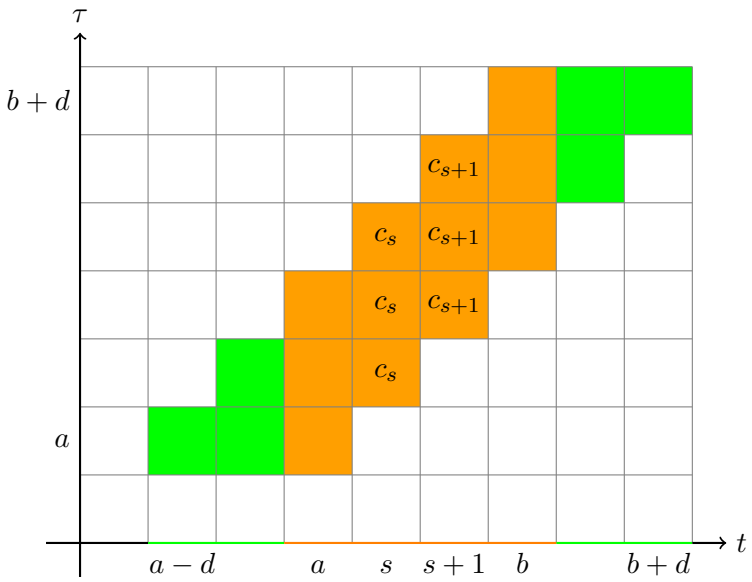


Figure 2: On the right-hand side of the equation in Lemma 11, we are summing each row of the squares in the picture. On the left hand side, we are summing the columns. Each column s contains the same constant c_s in each component.

Appendix A. An Accountants' Lemma

The next elementary lemma can be proved straightforwardly by swapping the order of the sums (see Figure 2).

Lemma 11 *If $(c_t)_{t\in\mathbb{Z}} \subset \mathbb{R}$, $a, b \in \mathbb{Z}$ are such that $a \le b$ and $d \ge 0$ then*

$$\sum_{t=a-d}^{a-1} (t-a+d+1)c_t + (d+1)\sum_{t=a}^{b} c_t + \sum_{t=b+1}^{b+d} (b+d+1-t)c_t = \sum_{\tau=a}^{b+d} \sum_{t=\tau-d}^{\tau} c_t \ .$$

Appendix B. Stability of FTRL with Tsallis entropy

In this section, we prove a key stability property of FTRL with Tsallis entropy that could be of independent interest. We begin by showing that, when $\eta > 0$, there exists a unique solution of the optimization problem that defines $\boldsymbol{q}_n$ in Algorithm 2 and provide a formula for it in terms of the corresponding Lagrange multiplier λ. An analogous result was stated in (Zimmert and Seldin, 2021, Section 3.3) for the related algorithm Tsallis-INF.

Lemma 12 *Let $\boldsymbol{w} \in \mathbb{R}^K$ and $\eta > 0$. Then*

$$\exists! \lambda_{\boldsymbol{w},\eta} < \min_{i\in[K]} \boldsymbol{w}(i) \ , \quad \sum_{i\in[K]} \frac{\eta^2}{\big(\boldsymbol{w}(i) - \lambda_{\boldsymbol{w},\eta}\big)^2} = 1 \ .$$

Furthermore, defining for all $i \in [K]$,

$$\boldsymbol{q}^{\mathrm{opt}}_{\boldsymbol{w},\eta}(i) \coloneqq \frac{\eta^2}{\big(\boldsymbol{w}(i) - \lambda_{\boldsymbol{w},\eta}\big)^2} \,,$$

we have that $\boldsymbol{q}^{\mathrm{opt}}_{\boldsymbol{w},\eta}$ *is the unique global minimizer of the function:*

$$\Delta_K \to \mathbb{R} \,, \quad \boldsymbol{q} \mapsto \sum_{i\in[K]} \boldsymbol{w}(i)\boldsymbol{q}(i) - 2\eta \sum_{i\in[K]} \sqrt{\boldsymbol{q}(i)} \,.$$

Proof Define the two auxiliary functions

$$f_{\boldsymbol{w},\eta}\colon [0,\infty)^K \to \mathbb{R} \,, \qquad \boldsymbol{q} \mapsto \sum_{i\in[K]} \boldsymbol{w}(i)\boldsymbol{q}(i) - 2\eta \sum_{i\in[K]} \sqrt{\boldsymbol{q}(i)}$$

$$\varphi\colon [0,\infty)^K \to \mathbb{R} \,, \qquad \boldsymbol{q} \mapsto \sum_{i\in[K]} \boldsymbol{q}(i)$$

Note that $f_{\boldsymbol{w},\eta}$ is strictly convex and continuous on $[0,\infty)^K$ and differentiable on $(0,\infty)^K$. Thus, by the Lagrange multiplier theorem, a point $\boldsymbol{q} \in (0,\infty)^K$ is the unique global minimizer of $f_{\boldsymbol{w},\eta}$ on the simplex Δ_K if and only if

$$\varphi(\boldsymbol{q}) = 1 \qquad \text{and} \qquad \exists \lambda \in \mathbb{R},\ \nabla f_{\boldsymbol{w},\eta}(\boldsymbol{q}) = \lambda \nabla \varphi(\boldsymbol{q}) \tag{4}$$

A direct verification shows that a pair $(\boldsymbol{q},\lambda) \in (0,\infty)^K \times \mathbb{R}$ satisfies condition (4) if and only if

$$\lambda < \min_{i\in[K]} \boldsymbol{w}(i) \,, \quad \sum_{i\in[K]} \frac{\eta^2}{\big(\boldsymbol{w}(i) - \lambda\big)^2} = 1 \,, \quad \text{and} \quad \forall i \in [K] \,,\ \boldsymbol{q}(i) = \frac{\eta^2}{\big(\boldsymbol{w}(i) - \lambda\big)^2} \tag{5}$$

Note that, letting $m \coloneqq \min_{i\in[K]} \boldsymbol{w}(i)$, the function

$$g\colon (-\infty, m) \to \mathbb{R} \,, \quad \lambda \mapsto \sum_{i\in[K]} \frac{\eta^2}{\big(\boldsymbol{w}(i) - \lambda\big)^2}$$

is continuous, strictly increasing, and it satisfies

$$\lim_{\lambda\to-\infty} g(\lambda) = 0 \qquad \text{and} \qquad \lim_{\lambda\to m^-} g(\lambda) = +\infty$$

hence, there exists a unique $\lambda_{\boldsymbol{w},\eta} \in (-\infty, m)$ such that $g(\lambda_{\boldsymbol{w},\eta}) = 1$. Therefore, $\boldsymbol{q}^{\mathrm{opt}}_{\boldsymbol{w},\eta}$ is the unique global minimizer of $f_{\boldsymbol{w},\eta}$ on the simplex Δ_K. ■

This lemma controls the variation of the Lagrange multipliers corresponding to two points that vary by quantity δ only in a single coordinate.

Lemma 13 *Let* $\boldsymbol{w} \in \mathbb{R}^K$ *such that* $\boldsymbol{w}(1) \le \cdots \le \boldsymbol{w}(K)$ *and* $\eta > 0$. *Then, for all* $\delta > 0$ *and* $j \in [K]$,

$$\lambda_{\boldsymbol{w}+\delta\boldsymbol{e}_j,\eta} - \lambda_{\boldsymbol{w},\eta} \le \frac{1}{j}\delta \,,$$

where $\lambda_{\boldsymbol{w}+\delta\boldsymbol{e}_j,\eta}$ *and* $\lambda_{\boldsymbol{w},\eta}$ *are defined as in Lemma 12 and* $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K$ *is the canonical basis of* $\mathbb{R}^K$.

Proof Define $\mathcal{X} \coloneqq \big\{\big(\boldsymbol{u}(1),\dots,\boldsymbol{u}(K),\lambda\big) \in \mathbb{R}^K \times \mathbb{R} \mid \lambda < \min_{i\in[K]} \boldsymbol{u}(i)\big\}$ and

$$\psi\colon \mathcal{X} \to \mathbb{R}\,, \quad \big(\boldsymbol{u}(1),\dots,\boldsymbol{u}(K),\lambda\big) \mapsto \sum_{i\in[K]} \frac{\eta^2}{\big(\boldsymbol{u}(i)-\lambda\big)^2}$$
$$\Lambda\colon \mathbb{R}^K \to \mathbb{R}\,, \quad \boldsymbol{u} \mapsto \lambda_{\boldsymbol{u},\eta}$$

where $\lambda_{\boldsymbol{u},\eta}$ is defined as in Lemma 12. By the implicit function theorem, we have that Λ is infinitely differentiable and for all $\boldsymbol{u} \in \mathbb{R}^K$,

$$\begin{aligned} D_j\Lambda(\boldsymbol{u}) &= -\frac{D_j\psi\big(\boldsymbol{u}(1),\dots,\boldsymbol{u}(K),\lambda_{\boldsymbol{u},\eta}\big)}{D_{K+1}\psi\big(\boldsymbol{u}(1),\dots,\boldsymbol{u}(K),\lambda_{\boldsymbol{u},\eta}\big)} \\ &= -\frac{-2\frac{\eta^2}{(\boldsymbol{u}(j)-\lambda_{\boldsymbol{u},\eta})^3}}{2\sum_{i\in[K]}\frac{\eta^2}{(\boldsymbol{u}(i)-\lambda_{\boldsymbol{u},\eta})^3}} = \frac{1}{1+\sum_{i\in[K],i\neq j}\left(\frac{\boldsymbol{u}(j)-\lambda_{\boldsymbol{u},\eta}}{\boldsymbol{u}(i)-\lambda_{\boldsymbol{u},\eta}}\right)^3}\,, \end{aligned}$$

where we denoted the partial derivative with respect to the j-th coordinate by D_j. Now, by the fundamental theorem of calculus,

$$\begin{aligned} &\lambda_{\boldsymbol{w}+\delta\boldsymbol{e}_j,\eta} - \lambda_{\boldsymbol{w},\eta} = \Lambda(\boldsymbol{w}+\delta\boldsymbol{e}_j) - \Lambda(\boldsymbol{w}) = \delta\int_0^1 D_j\Lambda(\boldsymbol{w}+s\delta\boldsymbol{e}_j)\,\mathrm{d}s \\ &= \delta\int_0^1 \frac{1}{1+\sum_{i\in[K],i\le j-1}\left(\frac{\boldsymbol{w}(j)+s\delta-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}{\boldsymbol{w}(i)-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}\right)^3 + \sum_{i\in[K],i\ge j+1}\left(\frac{\boldsymbol{w}(j)+s\delta-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}{\boldsymbol{w}(i)-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}\right)^3}\,\mathrm{d}s \\ &\le \delta\int_0^1 \frac{1}{1+\sum_{i\in[K],i\le j-1}\left(\frac{\boldsymbol{w}(i)-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}{\boldsymbol{w}(i)-\lambda_{\boldsymbol{w}+s\delta\boldsymbol{e}_j,\eta}}\right)^3}\,\mathrm{d}s = \frac{1}{j}\delta\,. \end{aligned}$$

■

We can finally prove the stability of FTRL with $\frac{1}{2}$-Tsallis entropy.

Proof of Theorem 5 Consider an arbitrary sequence of losses $(\ell_m)_{m\in\mathbb{N}} \subset [0,1]$. Fix any $\eta > 0$. For each $\boldsymbol{w} \in \mathbb{R}^K$, let $\lambda_{\boldsymbol{w}} \coloneqq \lambda_{\boldsymbol{w},\eta}$, where $\lambda_{\boldsymbol{w},\eta}$ is defined as in Lemma 12. Let $\mathcal{F}_0$ be the trivial σ-algebra (containing only the sample space and the empty set) and for all $n \in \mathbb{N}$, $\mathcal{F}_n \coloneqq \sigma(J_1,\dots J_n)$. Note that, for each $n \in \mathbb{N}$, we have that $\widehat{L}_{n-1}$ is $\mathcal{F}_{n-1}$-measurable and, as a consequence of Lemma 12, for all $i \in [K]$,

$$\boldsymbol{q}_n(i) = \frac{\eta^2}{\big(\widehat{L}_{n-1}(i) - \lambda_{\widehat{L}_{n-1}}\big)^2}\,.$$

Let $\boldsymbol{e}_1,\dots,\boldsymbol{e}_K$ be the canonical basis of $\mathbb{R}^K$ and fix any $n \in \mathbb{N}$. Define for each $j \in [K]$, $\widehat{\ell}_{n,j} \coloneqq \frac{\ell_n(j)}{\boldsymbol{q}_n(j)}\boldsymbol{e}_j$ and note that $\widehat{\ell}_n = \widehat{\ell}_{n,J_n}$. To make the notation more compact, we also let $\mathbb{E}_n[\cdot] \coloneqq \mathbb{E}[\cdot \mid \mathcal{F}_{n-1}]$. Then

$$\mathbb{E}_n\left[\sum_{i\in[K]}\big(\boldsymbol{q}_{n+1}(i)-\boldsymbol{q}_n(i)\big)^+\right] = \mathbb{E}_n\left[\sum_{i\in[K]}\left(\frac{\eta^2}{\big(\widehat{L}_n(i)-\lambda_{\widehat{L}_n}\big)^2} - \frac{\eta^2}{\big(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}\big)^2}\right)^+\right]$$

$$
\begin{aligned}
&= \mathbb{E}_n\left[\sum_{i\in[K]}\left(\frac{\eta^2}{\left(\widehat{L}_{n-1}(i)+\widehat{\ell}_{n,J_n}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,J_n}}\right)^2}-\frac{\eta^2}{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}\right)^2}\right)^+\right]\\
&= \sum_{j\in[K]}\mathbb{E}_n\left[\mathbb{I}\{J_n=j\}\sum_{i\in[K]}\left(\frac{\eta^2}{\left(\widehat{L}_{n-1}(i)+\widehat{\ell}_{n,j}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\right)^2}-\frac{\eta^2}{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}\right)^2}\right)^+\right]\\
&= \sum_{j\in[K]}\mathbb{E}_n\big[\mathbb{I}\{J_n=j\}\big]\sum_{i\in[K]}\left(\frac{\eta^2}{\left(\widehat{L}_{n-1}(i)+\widehat{\ell}_{n,j}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\right)^2}-\frac{\eta^2}{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}\right)^2}\right)^+\\
&= \sum_{j\in[K]}\boldsymbol{q}_n(j)\sum_{i\in[K]}\left(\frac{\eta^2}{\left(\widehat{L}_{n-1}(i)+\widehat{\ell}_{n,j}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\right)^2}-\frac{\eta^2}{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}\right)^2}\right)^+ = (\star)\,.
\end{aligned}
$$

Now, note that, for each $j\in[K]$:

- $\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\geq\lambda_{\widehat{L}_{n-1}}$ (because, as we show in the proof of Lemma 13, the directional derivatives of $\boldsymbol{u}\mapsto\lambda_{\boldsymbol{u}}$ are positive).
- For each $i\in[K]\backslash\{j\}$, $\frac{\eta^2}{\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}}\leq\frac{\eta^2}{\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}}=\frac{\eta^2}{\widehat{L}_{n-1}(i)+\widehat{\ell}_{n,j}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}}$ (by the previous point).
- $\frac{\eta^2}{\widehat{L}_{n-1}(j)-\lambda_{\widehat{L}_{n-1}}}\geq\frac{\eta^2}{\widehat{L}_{n-1}(j)+\widehat{\ell}_{n,j}(j)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}}$ (by the previous point and the fact that $\sum_{i\in[K]}\frac{\eta^2}{\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}}=1=\sum_{i\in[K]}\frac{\eta^2}{\widehat{L}_{n-1}(i)+\widehat{\ell}_{n,j}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}}$).

It follows that

$$
\begin{aligned}
(\star) &= \sum_{j\in[K]}\boldsymbol{q}_n(j)\sum_{i\in[K],i\neq j}\left(\frac{\eta^2}{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\right)^2}-\frac{\eta^2}{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}\right)^2}\right)\\
&= \eta^2\sum_{j\in[K]}\boldsymbol{q}_n(j)\sum_{i\in[K],i\neq j}\frac{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}\right)^2-\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\right)^2}{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\right)^2\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}\right)^2}\\
&= \eta^2\sum_{j\in[K]}\boldsymbol{q}_n(j)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}-\lambda_{\widehat{L}_{n-1}})\sum_{i\in[K],i\neq j}\frac{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}\right)+\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\right)}{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\right)^2\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}}\right)^2}\\
&\leq 2\eta^2\sum_{j\in[K]}\boldsymbol{q}_n(j)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}-\lambda_{\widehat{L}_{n-1}})\sum_{i\in[K],i\neq j}\frac{1}{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\right)^3}\\
&\leq 2\eta^2\sum_{j\in[K]}\boldsymbol{q}_n(j)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}-\lambda_{\widehat{L}_{n-1}})\left(\sum_{i\in[K],i\neq j}\frac{1}{\left(\widehat{L}_{n-1}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\right)^2}\right)^{3/2}\\
&= 2\eta^2\sum_{j\in[K]}\boldsymbol{q}_n(j)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}-\lambda_{\widehat{L}_{n-1}})\left(\sum_{i\in[K],i\neq j}\frac{1}{\left(\widehat{L}_{n-1}(i)+\widehat{\ell}_{n,j}(i)-\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\right)^2}\right)^{3/2}
\end{aligned}
$$

$$\leq 2\eta^2 \sum_{j\in[K]} \boldsymbol{q}_n(j)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}} - \lambda_{\widehat{L}_{n-1}}) \left(\sum_{i\in[K]} \frac{1}{\big(\widehat{L}_{n-1}(i) + \widehat{\ell}_{n,j}(i) - \lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}}\big)^2} \right)^{3/2}$$
$$= \frac{2}{\eta} \sum_{j\in[K]} \boldsymbol{q}_n(j)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,j}} - \lambda_{\widehat{L}_{n-1}}) = (\star\star) \,.$$

Now, let σ be a random permutation of $[K]$ such that $\widehat{L}_{n-1}\big(\sigma(1)\big) \leq \cdots \leq \widehat{L}_{n-1}\big(\sigma(K)\big)$. Then, using Lemma 13, we have

$$\begin{aligned}
(\star\star) &= \frac{2}{\eta} \sum_{j\in[K]} \boldsymbol{q}_n\big(\sigma(j)\big)(\lambda_{\widehat{L}_{n-1}+\widehat{\ell}_{n,\sigma(j)}} - \lambda_{\widehat{L}_{n-1}}) \\
&= \frac{2}{\eta} \sum_{j\in[K]} \boldsymbol{q}_n\big(\sigma(j)\big) \left(\lambda_{\widehat{L}_{n-1}+\frac{\ell_n(\sigma(j))}{\boldsymbol{q}_n(\sigma(j))}\boldsymbol{e}_{\sigma(j)}} - \lambda_{\widehat{L}_{n-1}} \right) \\
&\leq \frac{2}{\eta} \sum_{j\in[K]} \boldsymbol{q}_n\big(\sigma(j)\big) \frac{1}{j} \frac{\ell_n(\sigma(j))}{\boldsymbol{q}_n(\sigma(j))} \leq \frac{2}{\eta} \sum_{j\in[K]} \frac{1}{j} \leq 2\frac{1+\ln K}{\eta} \,.
\end{aligned}$$

In conclusion:

$$\mathbb{E}\left[\sum_{i\in[K]} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big)^+ \mid \mathcal{F}_{n-1} \right] \leq 2\frac{1+\ln K}{\eta} \,.$$

It follows that:

$$\mathbb{E}\left[\sum_{i\in[K]} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big)^+ \right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i\in[K]} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big)^+ \mid \mathcal{F}_{n-1} \right] \right] \leq 2\frac{1+\ln K}{\eta} \,.$$

Being n and $(\ell_m)_{m\in\mathbb{N}}$ arbitrary, the result follows. ■

Appendix C. Stability of Exp3

In this section, we prove the stability of Exp3.

Lemma 14 *Exp3 with learning rate η is ξ-stable with $\xi = \eta$.*

Proof Consider an arbitrary sequence of losses $(\ell_n)_{n\in\mathbb{N}} \subset [0,1]$. In this case, stability holds pointwise (for all realizations of the actions $J_1, J_2, \ldots$ played by Exp3 on the sequence of losses $(\ell_n)_{n\in\mathbb{N}}$) rather that in expectation. From (Cesa-Bianchi et al., 2016, Lemma 1) we have, for any round $n \in \mathbb{N}$ and all arms $i \in [K]$,

$$\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i) \leq \eta\, \boldsymbol{q}_{n+1}(i) \sum_{j=1}^{K} \boldsymbol{q}_n(j)\widehat{\ell}_n(j)$$

where $\widehat{\ell}_n(j) = \frac{\ell_n(j)\mathbb{I}\{J_n=j\}}{\boldsymbol{q}_n(j)}$, $\boldsymbol{q}_n(j) = \boldsymbol{w}_n(j)/\sum_{k=1}^K \boldsymbol{w}_n(k)$, and if $n=1$, $\boldsymbol{w}_n(k)=1$ while, if $n \geq 2$, $\boldsymbol{w}_n(k)$ is defined inductively by $\boldsymbol{w}_n(k) = \boldsymbol{q}_{n-1}(k)e^{-\eta\widehat{\ell}_{n-1}(k)}$. Hence we can write, for any $n \in \mathbb{N}$,

$$\begin{aligned}\sum_{i\,:\,\boldsymbol{q}_{n+1}(i)>\boldsymbol{q}_n(i)} \big(\boldsymbol{q}_{n+1}(i) - \boldsymbol{q}_n(i)\big) &\leq \sum_{i\,:\,\boldsymbol{q}_{n+1}(i)>\boldsymbol{q}_n(i)} \eta\,\boldsymbol{q}_{n+1}(i)\sum_{j=1}^K \boldsymbol{q}_n(j)\widehat{\ell}_n(j) \\ &= \sum_{i\,:\,\boldsymbol{q}_{n+1}(i)>\boldsymbol{q}_n(i)} \eta\,\boldsymbol{q}_{n+1}(i)\ell_n(J_n) \leq \eta \sum_{i\,:\,\boldsymbol{q}_{n+1}(i)>\boldsymbol{q}_n(i)} \boldsymbol{q}_{n+1}(i) \leq \eta\,.\end{aligned}$$

Being $(\ell_n)_{n\in\mathbb{N}}$ arbitrary, the result follows. ■

Appendix D. Lower Bound (missing proofs)

We begin this section by showing a reduction mapping each algorithm for bandits with composite anonymous feedback to one for linear bandits with a better or equal regret.

Proof of Lemma 8 Fix an instance $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_{T/(d+1)}$ of a linear bandit problem and use it to construct an instance of the d-delayed bandit setting with loss components

$$\ell_t^{(s)}(i) = \begin{cases} \boldsymbol{\ell}_{\lceil t/(d+1)\rceil}^\top \boldsymbol{e}_i & \text{if } t+s=0 \pmod{d+1}, \\ 0 & \text{otherwise,}\end{cases}$$

where $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K$ are the elements of the canonical basis of $\mathbb{R}^K$. These components define the following composite loss incurred by any algorithm A_d playing actions $I_1, I_2, \ldots$

$$\ell_t^\circ(I_{t-d}, \ldots, I_t) = \sum_{s=0}^d \ell_{t-s}^{(s)}(I_{t-s}) = \begin{cases} (d+1)\,\boldsymbol{\ell}_{\lceil t/(d+1)\rceil}^\top \boldsymbol{q}_t & \text{if } t=0 \pmod{d+1}, \\ 0 & \text{otherwise,}\end{cases}$$

where $\boldsymbol{q}_t$ is defined from $I_{t-d}, \ldots, I_t \in [K]$ as follows

$$\boldsymbol{q}_t(j) = \frac{1}{d+1}\sum_{s=t-d}^t \mathbb{I}\{I_s = j\} \qquad j \in [K]. \tag{6}$$

Note that $\boldsymbol{q}_t(i)$ is the fraction of times action i was played by A_d in the last $d+1$ rounds. Given the algorithm A_d, we define the algorithm A for playing linear bandits on the loss sequence $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_{T/(d+1)}$ as follows. If $t \neq 0 \pmod{d+1}$, then A skips the round. On the other hand, when $t = 0 \pmod{d+1}$, A performs action $\boldsymbol{q}_t$ defined in (6), observes the loss $\boldsymbol{\ell}_{\lceil t/(d+1)\rceil}^\top \boldsymbol{q}_t$, and returns to A_d the composite loss $\ell_t^\circ(I_{t-d}, \ldots, I_t)$. Essentially, A_d observes a nonzero composite loss only every d time steps, when $t = 0 \pmod{d+1}$. When this happens, the composite loss of A_d is $d\,\boldsymbol{\ell}_{\lceil t/(d+1)\rceil}^\top \boldsymbol{q}_t$, which is $d+1$ times the loss of A.

Now it is enough to note that, using (3),

$$\min_{k=1,\ldots,K}\sum_{t=1}^T \ell_t^\circ(k,\ldots,k) = \min_{k=1,\ldots,K}(d+1)\sum_{s=1}^{T/(d+1)} \boldsymbol{\ell}_s^\top \boldsymbol{e}_k = \min_{\boldsymbol{q}\in\Delta_K}(d+1)\sum_{s=1}^{T/(d+1)} \boldsymbol{\ell}_s^\top \boldsymbol{q}\,.$$

This concludes the proof. ■

We now prove a lower bound for linear bandits.

Proof of Lemma 9 The statement is essentially proven in (Shamir, 2015, Theorem 5), where the author shows a $\Omega\big(\sqrt{K/T}\big)$ lower bound on the error of *bandit linear optimization* in the probability simplex.[4] As explained in (Shamir, 2015, Section 1.1), (cumulative) regret lower bounds for linear bandits can be obtained by multiplying the lower bounds on bandit linear optimization error by T. A possible issue is that the proof in (Shamir, 2015, Theorem 5) uses unbounded Gaussian losses. However, in (Shamir, 2015, Appendix B) it is shown how lower bounds for Gaussian losses can be converted into lower bounds for losses in $[-1, 1]$ at the cost of a $1/\sqrt{\ln T}$ factor in the regret. Finally, note that our setting requires losses in $[0, 1]$, but this is not an issue either because we are in a linear setting, and thus we can add the $(1, \dots, 1)$ constant vector to all loss vectors without affecting the regret. ■

References

J. D. Abernethy, C. Lee, and A. Tewari. Fighting bandits with a new kind of smoothness. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

P. Agrawal and T. Tulabandula. Learning by repetition: Stochastic multi-armed bandits under priming effect. In *Conference on Uncertainty in Artificial Intelligence*, pages 470–479. PMLR, 2020.

O. Anava, E. Hazan, and S. Mannor. Online learning for adversaries with memory: price of past mistakes. In *Advances in Neural Information Processing Systems*, pages 784–792, 2015.

R. Arora, O. Dekel, and A. Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proc. 29th ICML*, 2012.

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

N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622, 2016.

O. Dekel, J. Ding, T. Koren, and Y. Peres. Online learning with composite loss functions. In *Conference on Learning Theory*, pages 1214–1231, 2014a.

O. Dekel, E. Hazan, and T. Koren. The blinded bandit: Learning with adaptive feedback. In *Advances in Neural Information Processing Systems*, pages 1610–1618, 2014b.

4. It is worth stressing that the lower bound in (Shamir, 2015) is based on stochastic i.i.d. generation of losses, hence it does not violate our assumption about the obliviousness of the adversary.

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

S. Garrabrant, N. Soares, and J. Taylor. Asymptotic convergence in online learning with unbounded delays. *arXiv preprint arXiv:1604.05280*, 2016.

S. Ito, D. Hatano, H. Sumita, K. Takemura, T. Fukunaga, N. Kakimura, and K.-I. Kawarabayashi. Delay and cooperation in nonstochastic linear bandits. *Advances in Neural Information Processing Systems*, 33:4872–4883, 2020.

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

P. Joulani, A. György, and C. Szepesvári. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *AAAI*, volume 16, pages 1744–1750, 2016.

D. Khashabi, K. Quanrud, and A. Taghvaei. Adversarial delays in online strongly-convex optimization. *arXiv preprint arXiv:1605.06201*, 2016.

J. Langford, A. J. Smola, and M. Zinkevich. Slow learners are fast. *Advances in Neural Information Processing Systems*, 22:2331–2339, 2009.

B. Li, T. Chen, and G. B. Giannakis. Bandit online learning with unknown delays. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 993–1002. PMLR, 2019.

T. Mandel, Y.-E. Liu, E. Brunskill, and Z. Popovic. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *AAAI*, pages 2849–2856, 2015.

T. A. Mann, S. Gowal, A. Gyorgy, H. Hu, R. Jiang, B. Lakshminarayanan, and P. Srinivasan. Learning from delayed outcomes via proxies with applications to recommender systems. In *International Conference on Machine Learning*, pages 4324–4332. PMLR, 2019.

C. Mesterharm. On-line learning with delayed label feedback. In *Algorithmic Learning Theory*, pages 399–413. Springer, 2005.

G. Neu, A. Antos, A. György, and C. Szepesvári. Online Markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems 23*, pages 1804–1812. Curran Associates, Inc., 2010.

C. Pike-Burke, S. Agrawal, C. Szepesvari, and S. Grunewalder. Bandits with delayed, aggregated anonymous feedback. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4105–4113. PMLR, 10–15 Jul 2018.

K. Quanrud and D. Khashabi. Online learning with adversarial delays. In *Advances in Neural Information Processing Systems*, pages 1270–1278, 2015.

O. Shamir. On the complexity of bandit linear optimization. In *Conference on Learning Theory*, pages 1523–1551, 2015.

O. Shamir and L. Szlak. Online learning with local permutations and delayed feedback. In *Proc. 34th ICML*, 2017.

T. S. Thune, N. Cesa-Bianchi, and Y. Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6541–6550, 2019.

C. Vernade, O. Cappé, and V. Perchet. Stochastic bandit models for delayed conversions. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI*, 2017.

C. Vernade, A. Carpentier, T. Lattimore, G. Zappella, B. Ermis, and M. Brueckner. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pages 9712–9721. PMLR, 2020.

S. Wang, H. Wang, and L. Huang. Adaptive algorithms for multi-armed bandit with composite and anonymous feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10210–10217, 2021.

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

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

J. Zimmert and Y. Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *J. Mach. Learn. Res.*, 22:28–1, 2021.