Nonstochastic Bandits with Composite Anonymous Feedback

Nicolo Cesa-Bianchi, Claudio Gentile, Yishay Mansour

▶ **To cite this version:**

Nicolo Cesa-Bianchi, Claudio Gentile, Yishay Mansour. Nonstochastic Bandits with Composite Anonymous Feedback. COLT 2018 - 31st Annual Conference on Learning Theory, Jul 2018, Stockholm, Sweden. pp.1 - 23. hal-01916981

HAL Id: hal-01916981
https://inria.hal.science/hal-01916981

Submitted on 9 Nov 2018

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.

Proceedings of Machine Learning Research vol 75:1–23, 2018 31st Annual Conference on Learning Theory

Nonstochastic Bandits with Composite Anonymous Feedback

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Dipartimento di Informatica, Università degli Studi di Milano, Italy

Claudio Gentile CLA.GENTILE@GMAIL.COM
INRIA Lille Nord Europe (France) and Google LLC (USA)

Yishay Mansour MANSOUR.YISHAY@GMAIL.COM
Blavatnik School of Computer Science, Tel Aviv University and Google

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

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 at most d consecutive steps in an adversarial way. This implies that the instantaneous loss observed by the player at the end of each round is a sum of as many as d loss components of previously played actions. Hence, unlike the standard bandit setting with delayed feedback, here the player cannot observe the individual delayed losses, but only their sum. Our main contribution is a general reduction transforming a standard bandit algorithm into one that can operate in this harder setting. We also show how the regret of the transformed algorithm can be bounded in terms of the regret of the original algorithm. Our reduction cannot be improved in general: we prove a lower bound on the regret of any bandit algorithm in this setting that matches (up to log factors) the upper bound obtained via our reduction. Finally, we show how our reduction can be extended to more complex bandit settings, such as combinatorial linear bandits and online bandit convex optimization.

Keywords: Nonstochastic bandits, composite losses, delayed feedback, bandit convex optimization

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 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.

The 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). In a recent paper, Pike-Burke et al. (2017) revisited the problem of bandits with delayed feedback under the more realistic assumption that only the combined reward is available to

© 2018 N. Cesa-Bianchi, C. Gentile & Y. Mansour.

the system, while the individual reward components remain unknown. This model captures a much broader range of practical scenarios 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 to the overall change in sales. Pike-Burke et al. (2017) formalized this harder delayed setting in a bandit framework with stochastic rewards, whereby 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 $I_t \in \{1, \dots, K\}$ at time t generates a stochastic reward $X_t(I_t) \in [0, 1]$ and a stochastic delay $\tau_t(I_t) \in \{0, 1, \dots\}$, where, for each $i \in \{1, \dots, K\}$, $X_t(i)$ and τ_t are drawn i.i.d. from fixed distributions $\nu_X(i)$ and ν_τ, respectively. 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 $X_s(I_s)\mathbb{I}\{\tau_s = t - s\}$ for $s = 1, \dots, t$. The main result in (Pike-Burke et al., 2017) is that, when the expected delay μ_τ is known, the regret is at most of order of $K\big((\ln T)/\Delta + \mu_\tau\big)$. 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 consecutive time steps in the future, $t, t + 1, \dots, t + d - 1$. Hence, the loss observed by the player at time t is a *composite loss*, that is, the sum of d-many loss components $\ell_t^{(0)}(I_t), \ell_{t-1}^{(1)}(I_{t-1}), \dots, \ell_{t-d+1}^{(d-1)}(I_{t-d+1})$, 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 $s = 0, \dots, d - 2$, and $\ell_t^{(d-1)}(i) = \ell_t(i)$, we recover the model of nonstochastic bandits with delayed feedback. 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 turning a base nonstochastic bandit algorithm into one operating within the composite anonymous feedback setting. We apply our reduction to the standard nonstochastic bandit setting with no delay to provide an upper bound of order $\sqrt{dKT}$ (ignoring factors logarithmic in K) on the regret of nonstochastic bandits with composite anonymous feedback, where d is a known delay parameter and T is the time horizon. We also prove a matching lower bound (up to logarithmic factors), thereby showing 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). See the table below for a summary of results for nonstochastic K-armed bandits (all rates are optimal ignoring factors logarithmic in K).

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

Our results can be extended to nonstochastic bandit settings that are more general than the standard K-armed bandit problem. In fact, our algorithm applies to any bandit problem for which there exists a suitably stable base algorithm whose regret bound is a sublinear and concave function of time. We give concrete examples for the settings of combinatorial linear bandits and online bandit convex optimization.

We now give an idea of the proof techniques, specifically referring to the K-armed bandit problem. Similarly to (Pike-Burke et al., 2017), we play the same action for a block of at least $2d$ time steps, hence the feedback we get in the last d 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 (a composite loss sums d 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 over any d consecutive steps can be at most $2d-1$. Hence, we can normalize the total composite loss simply by dividing by $2d$ so as to obtain an average loss in the range $[0,1]$. This gives the right scaling for the regret in the second half of each block, since the bandit regret $\sqrt{KT}$ becomes $d\sqrt{KT/d} = \sqrt{dKT}$. 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 from Dekel et al. (2014b), who extend the block size by adding a random number of steps having geometric distribution with expectation $2d$ (for technical reasons, in this paper we us a larger expectation of about $4d$). This random positioning of the blocks is the key to preventing the oblivious adversary from causing a large regret in the first half of each block. On the other hand, as we prove in Sections 3 and 5, the distribution over actions maintained by the base algorithm is "backward stable". This implies that the algorithm is not significantly affected by the uncertainty in the positioning of the blocks.

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), 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., 2017). Our delayed composite loss function generalizes the composite loss function setting of Dekel et al. (2014a) —see the discussion at the end of Section 3 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 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 $i \in \{1,\dots,K\}$ is defined by the sum

$$\ell_t(i) = \sum_{s=0}^{d-1} \ell_t^{(s)}(i)$$

of d-many components $\ell_t^{(s)}(i) \geq 0$ for $s = 0,\dots,d-1$. 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 $t+d-1$. Yet, what the player observes at time t is only the combined loss incurred at time t, which is the sum $\ell_t^{(0)}(I_t) + \ell_{t-1}^{(1)}(I_{t-1}) + \cdots + \ell_{t-d+1}^{(d-1)}(I_{t-d+1})$ of the past d loss contributions, where $\ell_t^{(s)}(i) = 0$ for all i and s when $t \leq 0$. Since the setting $d = 1$ recovers the standard nonstochastic oblivious bandit model, in the following we assume $d \geq 2$. For all sequences of actions $i_1,\dots,i_d \in \{1,\dots,K\}$, define the d-delayed *composite* loss function

$$\ell_t^\circ(i_1, i_2, \dots, i_d) = \sum_{s=0}^{d-1} \ell_{t-s}^{(s)}(i_{d-s}) \,, \tag{1}$$

with $\ell_t^{(s)}(i) = 0$ for all i and s when $t \leq 0$. 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 $\ell_t^\circ(I_{t-d+1}, I_{t-d+2}, \dots, I_t)$. Note that, whereas the losses $\ell_t(i)$ are in $[0,1]$, the composite loss can take values as large as d. On the other hand, the cumulative composite loss of any action i over d consecutive steps is at most $2d-1$:

$$\sum_{\tau=t-d+1}^{t} \ell_\tau^\circ(i,\dots,i) = \sum_{\tau=t-d+1}^{t} \sum_{s=0}^{d-1} \ell_{\tau-s}^{(s)}(i) \leq \sum_{\tau=t-2d+2}^{t} \sum_{s=0}^{d-1} \ell_\tau^{(s)}(i) = \sum_{\tau=t-2d+2}^{t} \ell_\tau(i) \leq 2d-1 \,. \tag{2}$$

The goal of the algorithm is to bound its regret R_T against the best fixed action in hindsight,

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^\circ(I_{t-d+1},\dots,I_t)\right] - \min_k \sum_{t=1}^{T} \ell_t^\circ(k,\dots,k) \,.$$

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 in 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 $d-1$.

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.

Algorithm 1: The Composite Loss Wrapper.

Input: Base MAB algorithm A with parameter $\eta \in (0, 1]$.

Initialize:

- Draw I_0 from the uniform distribution $\boldsymbol{p}_1$ over $\{1, \dots, K\}$;
- If $B_0 = 1$ then $t = 0$ is an update round.

For $t = 1, 2, \dots$:

1. If $t-1$ was an update round, then draw $I_t \sim \boldsymbol{p}_t$ and play it without updating $\boldsymbol{p}_t$ (draw round, $\boldsymbol{p}_{t+1} = \boldsymbol{p}_t$);
2. Else if an update round was in the interval $\{t-2d+1, \dots, t-2\}$ then play $I_t = I_{t-1}$ without updating $\boldsymbol{p}_t$ (stay round, $\boldsymbol{p}_{t+1} = \boldsymbol{p}_t$);
3. Else play $I_t = I_{t-1}$ (stay round), and if $B_t = 1$ then the stay round becomes an update round. In such a case:
   - Feed Base MAB $A(\eta)$ with average composite loss[a]
   $$\overline{\ell}_t = \frac{1}{2d} \sum_{\tau=t-d+1}^{t} \ell_\tau^\circ(I_{\tau-d+1}, \dots, I_\tau)$$
   - Use the update rule $\boldsymbol{p}_t \to \boldsymbol{p}_{t+1}$ of Base MAB to obtain the new distribution $\boldsymbol{p}_{t+1}$.

a. Recall that when $t \le 0$, we defined $\ell_t^{(s)} = 0$, so the initial stretch of $2d-2$ actions $I_1, \dots, I_{2d-2}$ can be disregarded here at the price of an extra additive $\mathcal{O}(d)$ regret in the analysis.

3. Wrapper Algorithm for Composite Losses

Our "Composite Loss Wrapper" algorithm (Algorithm 1) wraps a standard bandit algorithm called here Base MAB (Base Multi-Armed Bandit). Base MAB operates on standard (noncomposite) losses with values in $[0, 1]$, producing probability distributions $\boldsymbol{p}_t$ over the action set $\{1, \dots, K\}$. The wrapper, which has access to a sequence $B_0, B_1, \dots$ of i.i.d. Bernoulli random variables of parameter q (to be chosen later), experiences three kinds of online rounds: a *draw*, an *update*, and a *stay* round. If round t is a draw round, the algorithm draws action I_t according to the current distribution $\boldsymbol{p}_t$ maintained by Base MAB, but without having Base MAB update $\boldsymbol{p}_t$. If t is an update round, then the algorithm's action I_t is the same as I_{t-1} (in particular, the algorithm does not draw I_t from $\boldsymbol{p}_t$), but then a distribution update $\boldsymbol{p}_t \to \boldsymbol{p}_{t+1}$ takes place by invoking the update rule of Base MAB over an average of the observed losses. Finally, if t is a stay round, then both $I_t = I_{t-1}$ and $\boldsymbol{p}_{t+1} = \boldsymbol{p}_t$. The way these three kinds of rounds are interleaved is illustrated in Figure 1.

Note that the algorithm's pseudocode corresponds to the description in Figure 1 in that update and draw rounds are interleaved, and an update round is immediately followed by a draw round. If an update round occurs at time $t \ge 1$, then no update round can occur during the next $2d-1$ rounds; the next update takes place at time $t + 2d + G$ where $G \ge 0$ is a Geometric random variable with parameter q. Hence a stretch of stay rounds is $2d - 2 + G$ round long. Moreover,

- If t is not a draw round (i.e., it is either an update or a stay round), then the last action is played again.

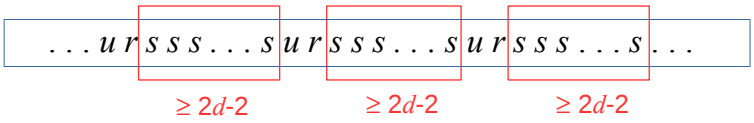

Figure 1: Sequence of rounds the algorithm is undergoing. Each (*u*)pdate round is always followed by a d(*r*)aw round, and then by a stretch of (*s*)tay rounds whose length is random, but is at least $2d-2$. The actual length of the stay stretch is ruled by the realizations of the Bernoulli random variables B_t.

- If t is an update round, then we are guaranteed that $I_t = \cdots = I_{t-2d+1}$, since the last draw could have only occurred at time $t-2d+1$ or earlier.

In order for our analysis to go through, we make mild assumptions on Base MAB. The first assumption (fulfilled by many standard K-armed bandit algorithms —see below) is a stability condition described in the following definition.

Definition 1 *Let $A(\eta)$ be a Base MAB with learning rate η, and $\{\boldsymbol{p}_t\}_{t=1}^T$ be the sequence of probability distributions over actions $\{1,\dots,K\}$ produced by $A(\eta)$ during a run over T rounds. We say that $A(\eta)$ is ξ-*stable *if for any round t we have that*

$$\mathbb{E}\left[\sum_{i\,:\,p_{t+1}(i)>p_t(i)} p_{t+1}(i)-p_t(i)\right] \le \xi$$

holds, where $\xi=\xi(K,\eta,\dots)$ is a function of K, η, and possibly other relevant parameters of the Base MAB.

The second assumption is that $A(\eta)$ is *nontrivial* for any $\eta>0$: when operating in the standard (non-delayed) bandit setting, $A(\eta)$ enjoys a concave (possibly linear) regret bound as a function of the time horizon T. Specifically, if we let $R_A(T,K,\eta)$ be a regret bound for A when the time horizon is T, the number of actions is K and the learning rate is η, we have for any $K\ge 1$ and $\eta>0$ that $R_A(T,K,\eta)$ is a concave function of T. For example, $R_A(T,K,\eta)=\mathcal{O}((\ln K)/\eta+\eta KT)$, which is linear in T. We have the following theorem, whose proof is in the appendix.

Theorem 2 *Let $A(\eta)$ be a ξ-stable and nontrivial Base MAB algorithm with learning rate η and regret bound $R_A(T,K,\eta)$ for standard K-armed bandits. Then Algorithm 1 with input $A(\eta)$ achieves regret*

$$R_T \le T\,\xi + 8(2d-1)R_A(T/2d,K,\eta)+\mathcal{O}(d)$$

for K-armed bandits with d-delayed composite anonymous feedback.

We can now derive corollaries for various algorithms using Theorem 2. Consider for instance, the well-known Exp3 algorithm of Auer et al. (2002). When operating with losses, the algorithm maintains a probability distribution $\boldsymbol{p}_t=(p_t(1),\dots,p_t(K))$ over $\{1,\dots,K\}$ of the form $p_t(i)=$

$w_t(i)/\sum_{j=1}^K w_t(j)$, while the update rule $\boldsymbol{p}_t \to \boldsymbol{p}_{t+1}$ can be described as follows:

$$w_{t+1}(i) = p_t(i)\, e^{-\eta \widehat{\ell}_t(i)}, \quad \widehat{\ell}_t(i) = \frac{\ell_t(i)\mathbb{I}\{I_t = i\}}{p_t(i)}, \quad i = 1, \dots, K\,. \tag{3}$$

When the losses $\ell_t(i)$ are in $[0,1]$ we have the regret bound $R_{\text{Exp3}}(T, K, \eta) \le \frac{\ln K}{\eta} + \frac{\eta}{2} K T$. Moreover, the following simple stability property holds (proof in the appendix).

Lemma 3 *Exp3 with learning rate η is ξ-stable with $\xi = \eta$.*

Combined with Theorem 2, this implies the following regret bound with composite losses.

Corollary 4 *If Algorithm 1 is run with Exp3(η) with $\eta = 4\sqrt{\frac{d \ln K}{(4K+1)T}}$ as Base MAB, then its regret for K-armed bandits with d-delayed composite anonymous feedback satisfies*

$$R_T \le 8\sqrt{d\,(4K+1)\,T\,\ln K} + \mathcal{O}(d) = \mathcal{O}(\sqrt{dKT \ln K})\,.$$

K-armed bandits are a special case of combinatorial linear bandits (Cesa-Bianchi and Lugosi, 2012), a setting where actions are incidence vectors $\boldsymbol{v} \in \mathcal{K} \subset \{0,1\}^n$ describing elements in some combinatorial space (e.g., spanning trees of a given graph) and loss vectors $\boldsymbol{\ell}_t \in [0,1]^n$ satisfy $\boldsymbol{\ell}_t^\top \boldsymbol{v} \in [0,1]$ for all $\boldsymbol{v} \in \mathcal{K}$ (in K-armed bandits, $\mathcal{K}$ is simply the canonical basis of $\{0,1\}^n$). Let $\boldsymbol{v}_t \in \mathcal{K}$ be the action played at time t. The generalization of Exp3 for the combinatorial bandit setting uses the Exp2 algorithm with loss estimates of the form $\widehat{\boldsymbol{\ell}}_t = P_t^+ \boldsymbol{v}_t \boldsymbol{v}_t^\top \boldsymbol{\ell}_t$, where $P_t = \mathbb{E}_{\boldsymbol{V}\sim\boldsymbol{p}_t}\left[\boldsymbol{V}\boldsymbol{V}^\top\right]$ and P_t^+ is the pseudo inverse of P_t —see (Dani et al., 2008). Note that these estimates are unbiased: $\mathbb{E}_t\left[\widehat{\boldsymbol{\ell}}_t^\top \boldsymbol{v}\right] = \boldsymbol{\ell}_t^\top \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathcal{K}$. The distribution $\boldsymbol{p}_t$ is a mixture $\boldsymbol{p}_t = (1-\gamma)\boldsymbol{q}_t + \boldsymbol{\mu}$, where $0 < \gamma < 1$, $\boldsymbol{q}_t$ has the exponential form (3)

$$q_t(\boldsymbol{v}) = \frac{w_t(\boldsymbol{v})}{\sum_{\boldsymbol{v}'\in\mathcal{K}} w_t(\boldsymbol{v}')}\,, \quad w_{t+1}(\boldsymbol{v}) = q_t(\boldsymbol{v})\, e^{-\eta \widehat{\boldsymbol{\ell}}_t^\top \boldsymbol{v}}\,, \quad \boldsymbol{v} \in \mathcal{K}\,,$$

and $\boldsymbol{\mu}$ is a fixed exploration distribution on $\mathcal{K}$. When run with an appropriate exploration distribution $\boldsymbol{\mu}$ and $\gamma = \eta n < 1$, Exp2(η) has the following regret bound —see, e.g., (Bubeck et al., 2012, Theorem 4), $R_{\text{Exp2}}(T, \mathcal{K}, \eta) \le \big(\ln |\mathcal{K}|\big)/\eta + 3\eta n T$. Now, similarly to Lemma 3, we can prove the following (the proof is provided in the appendix):

Lemma 5 *Exp2 with learning rate η and mixing coefficient γ is ξ-stable with $\xi = (1-\gamma)\eta$.*

Combining again with Theorem 2, the above implies the following regret bound with composite losses.

Corollary 6 *If Algorithm 1 is run with Exp2(η) with $\eta = 4\sqrt{\frac{d \ln |\mathcal{K}|}{(24n+1)T}}$ as Base MAB, then its regret for $\mathcal{K}$-combinatorial bandits, $\mathcal{K} \subseteq \{0,1\}^n$, with d-delayed composite anonymous feedback satisfies*

$$R_T \le 8\sqrt{d\,(24n+1)\,T\,\ln |\mathcal{K}|} + \mathcal{O}(d) = \mathcal{O}(\sqrt{dnT \ln |\mathcal{K}|})\,.$$

Remark 7 *The proof of Lemma 3 in the appendix shows pointwise stability, a stronger notion than the expected stability of Definition 1. In fact, an outer expectation over the random variable I_t in the proof of Lemma 3 makes the stability parameter ξ be upper bounded by $\eta \sum_{i=1}^K p_t(i)\ell_t(i)$ in round t, so that the term $T\xi$ in Theorem 2 can be replaced by $\eta\, L_A(T)$, where $L_A(T)$ is the cumulative (average) loss of the Base MAB. Coupled with a "first order" regret analysis of Exp3 where T is indeed replaced by $L_A(T)$ —see (Allenberg et al., 2006, Theorem 2), this gives a regret bound in the composite anonymous feedback setting where T is likewise replaced by $L_A(T)$.*

4. 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 $\boldsymbol{p}_t \in \Delta_K$ and suffers loss $\boldsymbol{\ell}_t^\top \boldsymbol{p}_t$, 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 $\boldsymbol{\ell}_t^\top \boldsymbol{p}_t$. The regret suffered by algorithm A playing actions $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_T$ is

$$R_T^{\text{lin}} = \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p}_t - \min_{\boldsymbol{p} \in \Delta_K} \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p} = \sum_{t=1}^{T} \boldsymbol{\ell}_t^\top \boldsymbol{p}_t - \min_{i=1,\ldots,K} \sum_{t=1}^{T} \ell_t(i) \tag{4}$$

where we used the fact that a linear function on the simplex is minimized at one of the corners. Let $R_T^{\text{lin}}(A, \Delta_K)$ denote the worst case regret (over the oblivious choice of $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T$) of algorithm A. Similarly, let $R_T(A_d, K, d)$ 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. 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 $R_T(A_d, K, d) \geq d\, R_{T/d}^{\text{lin}}(A, \Delta_K)$.*

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

Lemma 9 *The regret of any algorithm A for linear bandits in the simplex satisfies $R_T^{\text{lin}}(A, \Delta_K) = \widetilde{\Omega}\big(\sqrt{KT}\big)$.*

Using the above two lemmas we can prove the following theorem.

Theorem 10 *For any algorithm A_d for K-armed bandits with d-delayed composite anonymous feedback, $R_T(A_d, K, d) = \widetilde{\Omega}\big(\sqrt{dKT}\big)$.*

Proof Fix an algorithm A_d. Using the reduction of Lemma 8 gives an algorithm A such that $R_T(A_d, K, d) \geq d\, R_{T/d}^{\text{lin}}(A, \Delta_K) = \widetilde{\Omega}\big(\sqrt{dKT}\big)$, where we used Lemma 9 with horizon T/d 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 independent of the algorithm, thus preserving the oblivious nature of the adversary).

5. Extensions: Bandit Convex Optimization

We now show that a similar reduction as the one in Section 3 can be made to work in the more general Bandit Convex Optimization (BCO) framework. This learning setting is defined by a convex and compact domain $\Omega \subseteq \mathbb{R}^n$ and a sequence of loss functions $f_1, f_2, \ldots, f_T$, where each $f_t : \Omega \to [0,1]$ is convex over Ω. We assume each function f_t is the cumulated effect of d-many convex loss components $f_t^{(0)}, \ldots, f_t^{(d-1)}$, with $f_t^{(s)} : \Omega \to [0,1]$ so that, for any $\boldsymbol{w} \in \Omega$,

$$f_t(\boldsymbol{w}) = \sum_{s=0}^{d-1} f_t^{(s)}(\boldsymbol{w}) \in [0,1] \,.$$

To be concrete, we shall view f_t's components $f_t^{(s)}$ as constant fractions of f_t, specifically,

$$f_t^{(s)}(\boldsymbol{w}) = \alpha_t^{(s)} f_t(\boldsymbol{w}) \,, \qquad s = 0, \ldots, d-1 \,, \qquad t = 1, \ldots, T \,,$$

for nonnegative constant coefficients $\alpha_t^{(s)}$ such that $\sum_{s=0}^{d-1} \alpha_t^{(s)} = 1$, for $t = 1, \ldots, T$.

Since we are working with oblivious adversaries, we assume that all losses $\{f_t\}_{t=1\ldots T}$ and coefficients $\{\alpha_t^{(s)}\}_{t=1\ldots T, s=0\ldots d-1}$ are generated before the game starts. At each round $t = 1, 2, \ldots, T$, the learner picks $\tilde{\boldsymbol{w}}_t \in \Omega$ and suffers loss $f_t^{(0)}(\tilde{\boldsymbol{w}}_t) = \alpha_t^{(0)} f_t(\tilde{\boldsymbol{w}}_t)$ at time t, loss $f_t^{(1)}(\tilde{\boldsymbol{w}}_t) = \alpha_t^{(1)} f_t(\tilde{\boldsymbol{w}}_t)$ at time $t+1, \ldots$, loss $f_t^{(d-1)}(\tilde{\boldsymbol{w}}_t) = \alpha_t^{(d-1)} f_t(\tilde{\boldsymbol{w}}_t)$ at time $t+d-1$. However, what the algorithm really observes at time t is the cumulated effect of present and past actions quantified by the composite loss $f_t^\circ(\tilde{\boldsymbol{w}}_{t-d+1}, \tilde{\boldsymbol{w}}_{t-d+2}, \ldots, \tilde{\boldsymbol{w}}_t)$ with

$$f_t^\circ(\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d) = \sum_{s=0}^{d-1} f_{t-s}^{(s)}(\boldsymbol{w}_{d-s}) = \sum_{s=0}^{d-1} \alpha_{t-s}^{(s)} f_{t-s}(\boldsymbol{w}_{d-s}) \,,$$

where in the above $\alpha_t^{(s)} = 0$ for all s if $t \leq 0$. The aim of the algorithm is to minimize its regret

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} f_t^\circ(\tilde{\boldsymbol{w}}_{t-d+1}, \ldots, \tilde{\boldsymbol{w}}_t)\right] - \min_{\boldsymbol{w}} \sum_{t=1}^{T} f_t^\circ(\boldsymbol{w}, \ldots, \boldsymbol{w}) \,.$$

As in previous sections, we build a wrapper around a base Bandit Convex Optimization algorithm (Base BCO) which operates in the standard BCO framework with standard losses with range in $[0,1]$. Base BCO maintains at each round t a state variable $\boldsymbol{w}_t$ which is randomly perturbed to obtain the actual play $\tilde{\boldsymbol{w}}_t \in \Omega$. The wrapper algorithm is described as Algorithm 2. The notion of stability of the Base BCO has now to refer also to the sequence of loss functions the algorithm is operating with. Notice that, unlike the standard notion of stability in Online Convex Optimization, the kind of stability we need here is a *backward* stability, for it involves the backward differences $f_{t+1}(\tilde{\boldsymbol{w}}_{t+1}) - f_{t+1}(\tilde{\boldsymbol{w}}_t)$, rather than the forward differences $f_t(\tilde{\boldsymbol{w}}_t) - f_t(\tilde{\boldsymbol{w}}_{t+1})$. Moreover, we have to consider only the positive part of the backward difference.

Definition 11 *Let $A(\eta)$ be a Base BCO with learning rate η, and $\{\tilde{\boldsymbol{w}}_t\}_{t=1}^T$ be the sequence of plays produced by $A(\eta)$ during a run over T rounds on the sequence of convex losses $\{f_t\}_{t=1}^T$. We say that $A(\eta)$ is* ξ-stable *w.r.t. $\{f_t\}_{t=1}^T$ if for any round t we have that*[1]

$$\left[\mathbb{E}\left[f_{t+1}(\tilde{\boldsymbol{w}}_{t+1}) - f_{t+1}(\tilde{\boldsymbol{w}}_t)\right]\right]_+ \leq \xi$$

1. Here and throughout, $[x]_+ = \max\{x, 0\}$. The outer $[\cdot]_+$ in Definition 11 forces ξ to be nonnegative.

Algorithm 2: The Composite Loss Wrapper for BCO.

Input: Base BCO algorithm A with parameter $\eta \in (0, 1]$.

Initialize:

- Play any $\boldsymbol{w}_1 \in \Omega$;
- If $B_0 = 1$ then $t = 0$ is an update round.

For $t = 1, 2, \ldots$:

1. If $t-1$ was an update round, then play $\tilde{\boldsymbol{w}}_t$ by randomly perturbing state variable $\boldsymbol{w}_t$ without updating $\boldsymbol{w}_t$ (draw round, $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$);
2. Else if an update round was in the interval $\{t-2d+1, \ldots, t-2\}$ then play $\tilde{\boldsymbol{w}}_t = \tilde{\boldsymbol{w}}_{t-1}$ without updating $\boldsymbol{w}_t$ (stay round, $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$);
3. Else play $\tilde{\boldsymbol{w}}_t = \tilde{\boldsymbol{w}}_{t-1}$ (stay round), and if $B_t = 1$ then the stay round becomes an update round. In such a case:
   - Feed Base BCO $A(\eta)$ with average composite loss[a]

$$\overline{f}_t = \frac{1}{2d} \sum_{\tau=t-d+1}^{t} f_\tau^\circ(\tilde{\boldsymbol{w}}_{\tau-d+1}, \ldots, \tilde{\boldsymbol{w}}_\tau)$$

   - Use the update rule $\boldsymbol{w}_t \to \boldsymbol{w}_{t+1}$ of Base BCO to obtain the new state variable $\boldsymbol{w}_{t+1}$.

a. Recall that when $t \leq 0$, we defined $\alpha_t^{(s)} = 0$, for all s, so the initial stretch of $2d-2$ actions $\tilde{\boldsymbol{w}}_1, \ldots, \tilde{\boldsymbol{w}}_{2d-2}$ can be disregarded here at the price of an extra additive $\mathcal{O}(d)$ regret in the analysis.

holds, where ξ may depend on the input dimension n, the learning rate η, as well as on relevant properties of the loss functions $\{f_t\}_{t=1}^T$ and parameters of the algorithm.

We call a Base BCO algorithm A *nontrivial* w.r.t. the sequence of losses $\{f_t\}_{t=1}^T$ if, when applied to the standard setting on $\{f_t\}_{t=1}^T$, A has a regret bound $R_A(T, n, \eta)$ which is concave (possibly linear) in T for any $n \geq 1$, $\eta > 0$, and the other relevant parameters of the algorithm. Theorem 13 below rests on the assumption that the properties of the loss functions $\{f_t\}_{t=1\ldots T}$ that make the Base BCO algorithm A work are inherited by the average composite loss functions

$$\overline{f}_t(\boldsymbol{w}) = \frac{1}{2d} \sum_{\tau=t-d+1}^{t} f_\tau^\circ(\boldsymbol{w}, \ldots, \boldsymbol{w}) \,, \qquad \text{for } t \geq 2d-2,$$

the wrapper feeds to A. For the sake of concreteness, let us simply focus on boundedness, Lipschitzness, and β-smoothness w.r.t. the Euclidean norm $||\cdot||$. Recall that a convex function $f : \Omega \to [0, 1]$ is said to be β-smooth (or, equivalently, to have β-Lispchitz continuous gradient) w.r.t. $||\cdot||$ if for all $\boldsymbol{w}, \boldsymbol{w}' \in \Omega$ we have $||\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{w}')|| \leq \beta ||\boldsymbol{w} - \boldsymbol{w}'||$, where $\beta \geq 0$. Moreover, given constants $\beta_1, \beta_2, b_1, b_2 \geq 0$, if f_1 is β_1-smooth w.r.t. $||\cdot||$ and f_2 is β_2-smooth w.r.t. $||\cdot||$, then it is easy to see that $b_1 f_1 + b_2 f_2$ is $(b_1\beta_1 + b_2\beta_2)$-smooth w.r.t. $||\cdot||$. The following proposition lists the relevant properties of the functions $\overline{f}_t$ as immediate consequences of the properties of the functions f_t (proven in the appendix).

Proposition 12 *Let $f_1, \ldots, f_T : \Omega \subseteq \mathbb{R}^n \to [0,1]$ be a sequence of convex loss functions, and $\overline{f}_{2d-2}, \ldots, \overline{f}_T : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^+$ be the corresponding sequence of average composite losses. Then the following holds.*

1. *$\overline{f}_t(\boldsymbol{w}) \in [0,1]$ for all $\boldsymbol{w} \in \Omega$.*
2. *If, for some constant $L \geq 0$, the loss functions $f_1, \ldots, f_T$ are L-Lipschitz on Ω w.r.t. $||\cdot||$, then so are $\overline{f}_{2d-2}, \ldots, \overline{f}_T$.*
3. *If, for some constant $\beta \geq 0$, the loss functions $f_1, \ldots, f_T$ are β-smooth w.r.t. $||\cdot||$, then so are $\overline{f}_{2d-2}, \ldots, \overline{f}_T$.*

The following theorem, whose proof sketch is in the appendix, is the BCO counterpart to Theorem 2.

Theorem 13 *Let $A(\eta)$ be a ξ-stable and nontrivial Base BCO algorithm with learning rate η and regret bound $R_A(T,n,\eta)$ in the standard BCO setting on a sequence of convex losses $\{f_t\}_{t=1}^T$ enjoying Properties P (e.g., a subset of those listed in Proposition 12). If Properties P are inherited by $\{\overline{f}_t\}_{t=2d-2}^T$, then Algorithm 2 with input $A(\eta)$ achieves regret R_T satisfying*

$$R_T \leq T\,\xi + 8(2d-1)R_A(T/2d, n, \eta) + \mathcal{O}(d)\,.$$

As an example, consider the Base BCO algorithm by Saha and Tewari (2011) that works under the assumption of β-smoothness w.r.t. $||\cdot||$. This algorithm is a BCO variant of the SCRIBLe algorithm by Abernethy et al. (2012). The algorithm takes in input a learning rate η, a scaling parameter $\delta \in (0,1]$ (which will be set as a function of η), and a ν-self-concordant (barrier) function Ψ which we assume to be strongly convex w.r.t. $||\cdot||$. For instance, if Ω is defined by a set of m linear constraints $\Omega = \{\boldsymbol{w} \in \mathbb{R}^n : A\boldsymbol{w} \leq b\}$, a standard choice of Ψ is the sum of negative log distances to each boundary, i.e., $\Psi(\boldsymbol{w}) = -\sum_{i=1}^m \log(b_i - \boldsymbol{e}_i^\top A\boldsymbol{w})$, where $b = (b_1, \ldots, b_m)^\top$, and $\boldsymbol{e}_i$ is the i-th unit vector in the canonical basis of $\mathbb{R}^m$. Then Ψ is strongly convex w.r.t. $||\cdot||$, up to a strong convexity constant. The algorithm maintains at each round t the state variable $\boldsymbol{w}_t \in \Omega$, of the form

$$\boldsymbol{w}_t = \operatorname*{argmin}_{\boldsymbol{w}\in\Omega}\ \eta \sum_{\tau=1}^{t-1} \boldsymbol{w}^\top \hat{g}_\tau + \Psi(\boldsymbol{w})\,. \tag{5}$$

Then, it computes a perturbed version $\tilde{\boldsymbol{w}}_t$ of $\boldsymbol{w}_t$ as $\tilde{\boldsymbol{w}}_t = \boldsymbol{w}_t + \delta H_t^{-1/2} s_t$, where H_t is the Hessian matrix $\nabla^2\Psi(\boldsymbol{w}_t)$, s_t is drawn uniformly at random from the surface of the Euclidean n-dimensional unit ball $\mathbb{B}^n$, and $\delta = \delta(\eta) \in (0,1]$ is a scaling parameter. Finally, the update $\boldsymbol{w}_t \to \boldsymbol{w}_{t+1}$ amounts to computing the next vector $\hat{g}_t$ in (5) as $\hat{g}_t = \frac{n}{\delta} f_t(\tilde{\boldsymbol{w}}_t) H_t^{1/2} s_t$, an unbiased estimate of the gradient at $\boldsymbol{w}_t$ of a smoothed version of f_t. From (Saha and Tewari, 2011) one can bound $R_A(T,n,\eta) = R_A(T,n,\eta,\delta(\eta))$ as follows:

$$R_A(T,n,\eta) \leq \beta T\delta^2\mathcal{D}^2 + \eta T\left(\frac{n}{\delta}\right)^2 + \frac{2\nu\log T}{\eta} + \left(\frac{2}{\mathcal{D}} + \mathcal{D}\beta\right)\sqrt{T}\,, \tag{6}$$

where $\mathcal{D} = \max_{\boldsymbol{w},\boldsymbol{w}'\in\Omega} ||\boldsymbol{w} - \boldsymbol{w}'||$ is the diameter of Ω. Moreover, the following stability lemma can be shown (proven in the appendix).

Lemma 14 *Let $f_1, \ldots, f_T : \Omega \subseteq \mathbb{R}^n \to [0,1]$ be a sequence of β-smooth convex losses w.r.t. $||\cdot||$, and $\mathcal{D}$ be the diameter of Ω. Then the Base BCO algorithm by Saha and Tewari (2011) is ξ-stable, with $\xi = \mathcal{O}\left((1/\mathcal{D} + \mathcal{D}\beta)\frac{\eta n}{\delta} + \beta\delta^2\mathcal{D}^2\right)$.*

Combining (6) with Theorem 13 and Lemma 14 implies the following regret bound for composite losses.

Corollary 15 *If Algorithm 2 is run with the abovementioned algorithm by Saha and Tewari (2011) as Base BCO algorithm, with* $\eta = \mathcal{O}\left(\left(\frac{d\log(T/d)}{nT}\right)^{2/3}\right)$ *and* $\delta = \mathcal{O}(\eta^{1/4}\, n^{1/2})$, *then its regret for BCO with* d*-delayed composite anonymous feedback satisfies* $R_T = \mathcal{O}\left(\left(d\log(T/d)\right)^{1/3}(nT)^{2/3} + \sqrt{dT}\right)$, *where the* $\mathcal{O}$ *notation in the tuning of* η, δ *and in the bound on* R_T *hides the constants* β, $\mathcal{D}$ *and* ν.

Remark 16 *A similar statement can be made in the special case of bandit linear optimization, where the losses* f_t *are* β*-smooth with* $\beta = 0$. *In this case, Corollary 15 with* $\delta = 1$ *and* η *set appropriately gives a bound of the form* $\mathcal{O}\left(\sqrt{d\, n^2\, T\, \log(T/d)}\right)$. *The rate* $T^{2/3}$ *shown in Corollary 15 is the same as the one achieved by the Base BCO algorithm of Saha and Tewari (2011). Likewise, the rate* $T^{1/2}$ *achieved by Corollary 15 for the linear case is the same as the one obtained by the analyses in (Abernethy et al., 2012; Saha and Tewari, 2011). In both cases (and in line with the results in Sections 3 and 4) we have an extra factor* $\sqrt{d}$ *introduced by the composite anonymous feedback.*

6. Conclusions

We have investigated the setting of d-delayed composite anonymous feedback as applied to nonstochastic bandits. A general reduction technique was introduced that enables the conversion of a (backward stable) algorithm working in a standard bandit framework into one working in the composite feedback framework. In the case of K-armed bandits, we relied on a lower bound for bandit linear optimization in the probability simplex to show that no algorithm in the composite feedback framework can do better than $\mathcal{O}(\sqrt{dKT})$. In turn, up to log factors, this is what we obtain as an upper bound by applying our reduction to the standard Exp3 algorithm. We showed the generality and flexibility of our conversion technique by further applying it to Combinatorial Bandits (the Exp2 algorithm) and to Bandit Convex Optimization (the self-concordant barrier-based algorithms by Abernethy et al. (2012) and Saha and Tewari (2011)) with smooth/linear loss functions.

Three main directions for extending our work are:

- Proving an upper bound for the case of nonoblivious adversaries;
- Investigating the setting where the delay parameter d is not perfectly known;
- Extending our results to the nonstochastic contextual case.

Acknowledgments

YM was supported in part by a grant from the Israel Science Foundation (ISF).

References

J.D. Abernethy, E. Hazan, and A. Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012.

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *International Conference on Algorithmic Learning Theory*, pages 229–243. Springer, 2006.

Oren Anava, Elad Hazan, and Shie 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.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Annual Conference on Learning Theory*, volume 23, pages 41.1–41.14. Microtome, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

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

Varsha Dani, Sham M Kakade, and Thomas P Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, pages 345–352, 2008.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Online learning with composite loss functions. In *Conference on Learning Theory*, pages 1214–1231, 2014a.

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

Scott Garrabrant, Nate Soares, and Jessica Taylor. Asymptotic convergence in online learning with unbounded delays. *arXiv preprint arXiv:1604.05280*, 2016.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3–4):157–325, 2016.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461, 2013.

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

Daniel Khashabi, Kent Quanrud, and Amirhossein Taghvaei. Adversarial delays in online strongly-convex optimization. *arXiv preprint arXiv:1605.06201*, 2016.

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

Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popovic. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *AAAI*, pages 2849–2856, 2015.

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

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

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed anonymous feedback. *arXiv preprint arXiv:1709.06853*, 2017.

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

A. Saha and A. Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 636–642, 2011.

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

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

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

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

Appendix A. Proof of Theorem 2

Proof Let $\mathcal{U} \subseteq \{1, \dots, T\}$ be the (random) subset of update rounds. Let us call for brevity an update round a u-round, and similarly for the other two kinds. First, observe that if t is a u-round, we have $\overline{\ell}_t \in [0, 1]$. This is because, due to (2) and the fact that $I_t = I_{t-1} = \ldots = I_{t-2d+1}$,

$$\overline{\ell}_t = \frac{1}{2d} \sum_{\tau=t-d+1}^{t} \ell_\tau^\circ(I_{\tau-d+1}, \dots, I_\tau) = \frac{1}{2d} \sum_{\tau=t-d+1}^{t} \ell_\tau^\circ(I_t, \dots, I_t) \leq \frac{1}{2d}(2d-1) < 1 \,.$$

Since a u-round is followed by an r-round, and during the stretch of s-rounds between an r-round and the next u-round the action played by Algorithm 1 does not change, the algorithm behaves

exactly as $A(\eta)$ on the steps in $\mathcal{U}$. Therefore, if we set for brevity

$$\Delta_t^k = \frac{1}{2d} \sum_{\tau=t-d+1}^{t} \Big(\ell_\tau^\circ(I_{\tau-d+1}, \dots, I_{\tau-d+1}) - \ell_\tau^\circ(k, \dots, k) \Big) , \qquad \text{for } t \geq 2d-2,$$

we have, for any action k,

$$\mathbb{E}\left[\sum_{t \in \mathcal{U},\, t \geq 2d-2} \Delta_t^k \right] \leq \mathbb{E}\big[R_A(|\mathcal{U}|, K, \eta) \big] \leq R_A\big(\mathbb{E}[|\mathcal{U}|], K, \eta \big) \leq R_A(T/2d, K, \eta) , \tag{7}$$

where the second-last inequality is due to the concavity of $R_A(\cdot, K, \eta)$, and the last inequality derives from $|\mathcal{U}| \leq T/2d$, for there can be at most one u-round every $2d$ rounds. Now, notice that by definition of a u-round we have, for all t,

$$\mathbb{I}\{t \in \mathcal{U}\} = \mathbb{I}\{B_t = 1\}\, \mathbb{I}\left\{ \bigwedge_{s=1}^{2d-1} (t - s \notin \mathcal{U}) \right\} .$$

Moreover,

$$\begin{aligned}
\sum_{t=2d-2}^{T} \Delta_t^k &= \frac{1}{2d} \sum_{t=2d-2}^{T} \sum_{\tau=t-d+1}^{t} \Big(\ell_\tau^\circ(I_{\tau-d+1}, \dots, I_{\tau-d+1}) - \ell_\tau^\circ(k, \dots, k) \Big) \\
&= \frac{1}{2d} \sum_{t=d-1}^{2d-3} (t-d+2) \Big(\ell_t^\circ(I_{t-d+1}, \dots, I_{t-d+1}) - \ell_t^\circ(k, \dots, k) \Big) \\
&\quad + \frac{d}{2d} \sum_{t=2d-2}^{T-d+1} \Big(\ell_t^\circ(I_{t-d+1}, \dots, I_{t-d+1}) - \ell_t^\circ(k, \dots, k) \Big) \\
&\quad + \frac{1}{2d} \sum_{t=T-d+2}^{T} (T+1-t) \Big(\ell_t^\circ(I_{t-d+1}, \dots, I_{t-d+1}) - \ell_t^\circ(k, \dots, k) \Big) \\
&\geq \frac{1}{2} \sum_{t=2d-2}^{T-d+1} \Big(\ell_t^\circ(I_{t-d+1}, \dots, I_{t-d+1}) - \ell_t^\circ(k, \dots, k) \Big) \\
&\quad - \frac{1}{2d} \sum_{t=d-1}^{2d-3} (t-d+2) \ell_t^\circ(k, \dots, k) - \frac{1}{2d} \sum_{t=T-d+2}^{T} (T+1-t) \ell_t^\circ(k, \dots, k) \\
&\geq \frac{1}{2} \sum_{t=2d-2}^{T-d+1} \Big(\ell_t^\circ(I_{t-d+1}, \dots, I_{t-d+1}) - \ell_t^\circ(k, \dots, k) \Big) - 2(d-1) ,
\end{aligned} \tag{8}$$

where the last inequality holds because, due to (2),

$$\sum_{t=d-1}^{2d-3} (t-d+2) \ell_t^\circ(k, \dots, k) \leq (d-1) \sum_{t=d-1}^{2d-3} \ell_t^\circ(k, \dots, k) \leq (d-1)(2d-1)$$

and

$$\sum_{t=T-d+2}^{T}(T+1-t)\ell_t^\circ(k,\ldots,k) \le (d-1)\sum_{t=T-d+2}^{T}\ell_t^\circ(k,\ldots,k) \le (d-1)(2d-1)\,.$$

Now, for any action k we have,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in\mathcal{U},\,t\ge 2d-2}\Delta_t^k\right] &= \mathbb{E}\left[\sum_{t=2d-2}^{T}\mathbb{I}\{t\in\mathcal{U}\}\Delta_t^k\right] \\
&= \mathbb{E}\left[\sum_{t=2d-2}^{T}\mathbb{I}\{B_t=1\}\,\mathbb{I}\left\{\bigwedge_{s=1}^{2d-1}(t-s\notin\mathcal{U})\right\}\Delta_t^k\right] \\
&= q\,\mathbb{E}\left[\sum_{t=2d-2}^{T}\mathbb{E}\left[\mathbb{I}\left\{\bigwedge_{s=1}^{2d-1}(t-s\notin\mathcal{U})\right\}\Delta_t^k\,\middle|\,B_0,\ldots,B_{t-2d},I_0,\ldots,I_{t-2d+1}\right]\right] \\
&= q\,\mathbb{E}\left[\sum_{t=2d-2}^{T}\Delta_t^k\,\mathbb{E}\left[\mathbb{I}\left\{\bigwedge_{s=1}^{2d-1}(t-s\notin\mathcal{U})\right\}\,\middle|\,B_0,\ldots,B_{t-2d},I_0,\ldots,I_{t-2d+1}\right]\right] \\
&= q\,\mathbb{E}\left[\sum_{t=2d-2}^{T}\Delta_t^k\,\mathbb{E}\left[\mathbb{I}\left\{\bigwedge_{s=1}^{2d-1}(t-s\notin\mathcal{U})\right\}\,\middle|\,B_0,\ldots,B_{t-2d}\right]\right] \\
&= q\,\mathbb{E}\left[\sum_{t=2d-2}^{T}\Delta_t^k\,\mathbb{P}'\left(\bigwedge_{s=1}^{2d-1}(t-s\notin\mathcal{U})\right)\right]\,,
\end{aligned} \tag{9}$$

where we set for brevity $\mathbb{P}'(\cdot)=\mathbb{P}\big(\cdot\mid B_0,B_1\ldots,B_{t-2d}\big)$. We can thus write

$$\begin{aligned}
1-\mathbb{P}'\left(\bigwedge_{s=1}^{2d-1}(t-s\notin\mathcal{U})\right) &= \mathbb{P}'\left(\bigvee_{s=1}^{2d-1}(t-s\in\mathcal{U})\right) \\
&\le \sum_{s=1}^{2d-1}\mathbb{P}'\big(t-s\in\mathcal{U}\big) \\
&\le \sum_{s=1}^{2d-1}\mathbb{P}'\big(B_{t-s}=1,t-s-1\notin\mathcal{U},\ldots,t-s-2d+1\notin\mathcal{U}\big) \\
&\le q\,(2d-1)\,.
\end{aligned}$$

Hence, substituting into (9) and combining with (8), we conclude that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in\mathcal{U},\,t\ge 2d-2}\Delta_t^k\right] &\ge q(1-q(2d-1))\sum_{t=2d-2}^{T}\mathbb{E}\big[\Delta_t^k\big] \\
&\ge \frac{q}{2}(1-q(2d-1))\left(\sum_{t=2d-2}^{T-d+1}\Big(\mathbb{E}\big[\ell_t^\circ(I_{t-d+1},\ldots,I_{t-d+1})\big]-\ell_t^\circ(k,\ldots,k)\Big)-2(d-1)\right)\,.
\end{aligned} \tag{10}$$

Next, using $\mathbb{E}_t[\cdot]$ to denote expectation conditioned on all random events at time steps $1, \ldots, t-1$, we observe that

$$\begin{aligned}
&\mathbb{E}\Big[\ell_t^\circ(I_{t-d+1}, \ldots, I_t) - \ell_t^\circ(I_{t-d+1}, \ldots, I_{t-d+1})\Big] \\
&= \mathbb{E}\left[\sum_{s=0}^{d-1}\Big(\mathbb{E}_{t-s}\big[\ell_{t-s}^{(s)}(I_{t-s})\big] - \mathbb{E}_{t-d+1}\big[\ell_{t-s}^{(s)}(I_{t-d+1})\big]\Big)\right] \\
&= \mathbb{E}\left[\sum_{s=0}^{d-1}\sum_{i=1}^{K}\ell_{t-s}^{(s)}(i)\big(p_{t-s}(i) - p_{t-d+1}(i)\big)\right] \\
&\le \mathbb{E}\left[\sum_{s=0}^{d-1}\sum_{i\,:\,p_{t-s}(i)>p_{t-d+1}(i)}\big(p_{t-s}(i) - p_{t-d+1}(i)\big)\right] \le \xi\,, \qquad (11)
\end{aligned}$$

where the last inequality is because in any block of $2d$ rounds there is at most one update of distribution $\boldsymbol{p}_t$, each loss component $\ell_{t-s}^{(s)}(i)$ is in $[0,1]$, and because $A(\eta)$ is a ξ-stable Base MAB. Hence, for any k, we can write

$$\begin{aligned}
R_T &\le \mathbb{E}\left[\sum_{t=1}^{T}\ell_t^\circ(I_{t-d+1}, \ldots, I_t)\right] - \sum_{t=1}^{T}\ell_t^\circ(k, \ldots, k) \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\ell_t^\circ(I_{t-d+1}, \ldots, I_t) - \sum_{t=1}^{T}\ell_t^\circ(I_{t-d+1}, \ldots, I_{t-d+1})\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\ell_t^\circ(I_{t-d+1}, \ldots, I_{t-d+1})\right] - \sum_{t=1}^{T}\ell_t^\circ(k, \ldots, k) \\
&\le T\,\xi + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\ell_t^\circ(I_{t-d+1}, \ldots, I_{t-d+1})\right] - \sum_{t=1}^{T}\ell_t^\circ(k, \ldots, k)}_{(\star)} \qquad \text{(from (11))}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
(\star) &\le \mathbb{E}\left[\sum_{t=2d-2}^{T-d+1}\ell_t^\circ(I_{t-d+1}, \ldots, I_{t-d+1})\right] - \sum_{t=2d-2}^{T-d+1}\ell_t^\circ(k, \ldots, k) + 3d \\
&\le 2(d-1) + \frac{2}{q(1-q(2d-1))}\,\mathbb{E}\left[\sum_{t\in\mathcal{U},\,t\ge 2d-2}\Delta_t^k\right] + 3d \qquad \text{(from (10))} \\
&\le 2(d-1) + \frac{2}{q(1-q(2d-1))}\,R_A(T/2d, K, \eta) + 3d\,. \qquad \text{(from (7))}
\end{aligned}$$

By picking $q = \frac{1}{2(2d-1)}$ so as to maximize the denominator in the second term of the right-most side yields

$$(*) \le 8(2d-1)\,R_A(T/2d, K, \eta) + \mathcal{O}(d)\,,$$

so that

$$R_T \le T\,\xi + 8(2d-1)\,R_A(T/2d, K, \eta) + \mathcal{O}(d)\,,$$

as claimed. ■

Appendix B. Proof of Lemma 3

Proof In this case, stability holds pointwise (for all realizations of $I_1, \dots, I_T$) rather that in expectation (yet, see Remark 7). From (Cesa-Bianchi et al., 2016, Lemma 1) we have, for any round t,

$$p_{t+1}(i) - p_t(i) \le \eta\, p_{t+1}(i) \sum_{j=1}^{K} p_t(j) \widehat{\ell}_t(j)\ .$$

Hence we can write

$$\begin{aligned}
\sum_{i\,:\,p_{t+1}(i)>p_t(i)} p_{t+1}(i) - p_t(i) &\le \sum_{i\,:\,p_{t+1}(i)>p_t(i)} \eta\, p_{t+1}(i) \sum_{j=1}^{K} p_t(j) \widehat{\ell}_t(j) \\
&= \sum_{i\,:\,p_{t+1}(i)>p_t(i)} \eta\, p_{t+1}(i) \ell_t(I_t) \\
&\le \eta \sum_{i\,:\,p_{t+1}(i)>p_t(i)} p_{t+1}(i) \\
&\le \eta
\end{aligned}$$

concluding the proof. ■

Appendix C. Proof of Lemma 5

Proof Since $\boldsymbol{q}_t$ has exponential form, we can apply again (Cesa-Bianchi et al., 2016, Lemma 1) and obtain

$$p_{t+1}(\boldsymbol{v}) - p_t(\boldsymbol{v}) = (1-\gamma)\big(q_{t+1}(\boldsymbol{v}) - q_t(\boldsymbol{v})\big) \le (1-\gamma)\eta\, q_{t+1}(\boldsymbol{v}) \sum_{\boldsymbol{v}' \in \mathcal{K}} q_t(\boldsymbol{v}')\, \widehat{\boldsymbol{\ell}}_t^{\top} \boldsymbol{v}'\ .$$

Hence we can write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\boldsymbol{v}\,:\,p_{t+1}(\boldsymbol{v})>p_t(\boldsymbol{v})} p_{t+1}(\boldsymbol{v}) - p_t(\boldsymbol{v})\right] \\
&\le (1-\gamma)\eta\, \mathbb{E}\left[\sum_{\boldsymbol{v}\,:\,p_{t+1}(\boldsymbol{v})>p_t(\boldsymbol{v})} q_{t+1}(\boldsymbol{v}) \sum_{\boldsymbol{v}' \in \mathcal{K}} q_t(\boldsymbol{v}') \mathbb{E}_t\left[\widehat{\boldsymbol{\ell}}_t^{\top} \boldsymbol{v}'\right]\right] \\
&= (1-\gamma)\eta\, \mathbb{E}\left[\sum_{\boldsymbol{v}\,:\,p_{t+1}(\boldsymbol{v})>p_t(\boldsymbol{v})} q_{t+1}(\boldsymbol{v}) \sum_{\boldsymbol{v}' \in \mathcal{K}} q_t(\boldsymbol{v}') \boldsymbol{\ell}_t^{\top} \boldsymbol{v}'\right] \qquad \text{(because estimates are unbiased)} \\
&\le (1-\gamma)\eta\, \mathbb{E}\left[\sum_{\boldsymbol{v}\,:\,p_{t+1}(\boldsymbol{v})>p_t(\boldsymbol{v})} q_{t+1}(\boldsymbol{v})\right] \qquad \text{(because } \boldsymbol{\ell}_t^{\top} \boldsymbol{v} \in [0,1] \text{ for all } t \text{ and } \boldsymbol{v}) \\
&\le (1-\gamma)\eta
\end{aligned}$$

concluding the proof. ■

Appendix D. Proof of Lemma 8

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

$$\ell_t^{(s)}(i) = \begin{cases} \ell_{\lceil t/d \rceil}(i) & \text{if } t + s \pmod d = 0, \\ 0 & \text{otherwise.} \end{cases}$$

These components define the following composite loss incurred by any algorithm A_d playing actions $I_1, I_2, \dots$

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

where $\boldsymbol{p}_t$ is defined from $I_{t-d+1}, \dots, I_t \in \{1, \dots, K\}$ as follows

$$p_t(j) = \frac{1}{d} \sum_{s=t-d+1}^{t} \mathbb{I}\{I_s = j\} \qquad j = 1, \dots, K. \tag{12}$$

Note that $p_t(i)$ is the fraction of times action i was played by A_d in the last d rounds. Given the algorithm A_d, we define the algorithm A for playing linear bandits on the loss sequence $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{T/d}$ as follows. If $t \pmod d \neq 0$, then A skips the round. On the other hand, when $t \pmod d = 0$, A performs action $\boldsymbol{p}_t$ defined in (12), observes the loss $\boldsymbol{p}_t^\top \boldsymbol{\ell}_{\lceil t/d \rceil}$, and returns to A_d the composite loss $\ell_t^\circ(I_{t-d+1}, \dots, I_t)$. Essentially, A_d observes a nonzero composite loss only every d time steps, when $t \pmod d = 0$. When this happens, the composite loss of A_d is $d\, \boldsymbol{p}_t^\top \boldsymbol{\ell}_{\lceil t/d \rceil}$, which is d times the loss of A.

Now it is enough to note that, using (4),

$$\min_{k=1,\dots,K} \sum_{t=1}^{T} \ell_t^\circ(k, \dots, k) = \min_{k=1,\dots,K} d \sum_{s=1}^{T/d} \ell_s(k) = \min_{\boldsymbol{p} \in \Delta_K} d \sum_{s=1}^{T/d} \boldsymbol{p}^\top \boldsymbol{\ell}_s \,.$$

This concludes the proof. ■

Appendix E. Proof of Lemma 9

Proof 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.[2] 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

2. 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.

(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. ■

Appendix F. Proof of Proposition 12

Proof Simply observe that

$$\begin{aligned}
\overline{f}_t(\boldsymbol{w}) &= \frac{1}{2d} \sum_{\tau=t-d+1}^{t} \sum_{s=0}^{d-1} \alpha_{\tau-s}^{(s)} f_{\tau-s}(\boldsymbol{w}) \\
&= \frac{1}{2d} \left(\sum_{\tau=t-2d+2}^{t-d} \sum_{s=t-d+1}^{\tau+d-1} \alpha_{\tau}^{(s-\tau)} f_\tau(\boldsymbol{w}) + \sum_{\tau=t-d+1}^{t} \sum_{s=\tau}^{\tau+d-1} \alpha_{\tau}^{(s-\tau)} f_\tau(\boldsymbol{w}) \right) \\
&= \frac{1}{2d} \left(\sum_{\tau=t-2d+2}^{t-d} f_\tau(\boldsymbol{w}) \left(\sum_{s=t-d+1}^{\tau+d-1} \alpha_{\tau}^{(s-\tau)} \right) + \sum_{\tau=t-d+1}^{t} f_\tau(\boldsymbol{w}) \right) .
\end{aligned}$$

Now, since the first inner sum $\sum_{s=t-d+1}^{\tau+d-1} \alpha_{\tau}^{(s-\tau)}$ is upper bounded by $\sum_{s=0}^{d-1} \alpha_{\tau}^{(s)} = 1$, we see that $\overline{f}_t(\boldsymbol{w})$ is indeed a linear combination of the form

$$\overline{f}_t(\boldsymbol{w}) = \sum_{\tau=t-2d+2}^{t} b_\tau f_\tau(\boldsymbol{w}) ,$$

whose coefficients b_τ are nonnegative and sum to a quantity which is less than one. All the three claimed properties then immediately follow. ■

Appendix G. Proof of Theorem 13

Proof The proof is almost the same as the one of Theorem 2 (up to a change of notation), with the additional care that has to be taken when dealing with the inheritance of Properties P from $\{f_t\}_{t=1}^T$ to $\{\overline{f}_t\}_{t=2d-2}^T$. In particular, if we define

$$\Delta_t^{\boldsymbol{w}} = \frac{1}{2d} \sum_{\tau=t-d+1}^{t} \big(f_\tau^\circ(\tilde{\boldsymbol{w}}_{\tau-d+1}, \dots, \tilde{\boldsymbol{w}}_{\tau-d+1}) - f_\tau^\circ(\boldsymbol{w}, \dots, \boldsymbol{w}) \big), \qquad \text{for } t \geq 2d-2,$$

we have, for any $\boldsymbol{w} \in \Omega$,

$$\mathbb{E}\left[\sum_{t \in \mathcal{U},\, t \geq 2d-2} \Delta_t^{\boldsymbol{w}} \right] \leq R_A(T/2d, n, \eta) ,$$

since the average loss function $\overline{f}_t(\boldsymbol{w})$ enjoys the same properties as those that allow us to prove the regret bound $R_A(T,n,\eta)$ for the Base BCO algorithm A. Next,

$$\mathbb{E}\left[\sum_{t\in\mathcal{U},\, t\geq 2d-2} \Delta_t^{\boldsymbol{w}}\right]$$
$$\geq \frac{q}{2}(1-q(2d-1))\left(\sum_{t=2d-2}^{T-d+1}\Big(\mathbb{E}\left[f_t^\circ(\tilde{\boldsymbol{w}}_{t-d+1},\ldots,\tilde{\boldsymbol{w}}_{t-d+1})\right]-f_t^\circ(\boldsymbol{w},\ldots,\boldsymbol{w})\Big)-2(d-1)\right)$$

is the counterpart to (10), and is proved in exactly the same manner. Then, from the notion of stability given in Definition 11, we can write

$$\begin{aligned}
&\mathbb{E}\Big[f_t^\circ(\tilde{\boldsymbol{w}}_{t-d+1},\ldots,\tilde{\boldsymbol{w}}_t)-f_t^\circ(\tilde{\boldsymbol{w}}_{t-d+1},\ldots,\tilde{\boldsymbol{w}}_{t-d+1})\Big]\\
&\quad=\mathbb{E}\left[\sum_{s=0}^{d-1}\Big(f_{t-s}^{(s)}(\tilde{\boldsymbol{w}}_{t-s})-f_{t-s}^{(s)}(\tilde{\boldsymbol{w}}_{t-d+1})\Big)\right]\\
&\quad=\sum_{s=0}^{d-1}\alpha_{t-s}^{(s)}\,\mathbb{E}\left[f_{t-s}(\tilde{\boldsymbol{w}}_{t-s})-f_{t-s}(\tilde{\boldsymbol{w}}_{t-d+1})\right]\\
&\quad\leq\sum_{s=0}^{d-1}\alpha_{t-s}^{(s)}\Big[\mathbb{E}\left[f_{t-s}(\tilde{\boldsymbol{w}}_{t-s})-f_{t-s}(\tilde{\boldsymbol{w}}_{t-d+1})\right]\Big]_+\leq\xi\,,
\end{aligned}$$

since there is at most one update of the underlying state variable $\boldsymbol{w}_t$ (which in turn determines the distribution of the corresponding $\tilde{\boldsymbol{w}}_t$) during the rounds from $t-d+1$ to t, the coefficients $\alpha_{t-s}^{(s)}$ are in $[0,1]$ for all s and t, and Base BCO is assumed to be ξ-stable in the sense of Definition 11. Piecing together as in the proof of Theorem 2 proves the claim. ■

Appendix H. Proof of Lemma 14

Proof First recall the standard fact that if $f:\Omega\to[0,1]$ is β-smooth w.r.t. $||\cdot||$, then

$$f(\boldsymbol{w}')\leq f(\boldsymbol{w})+\nabla f(\boldsymbol{w})^\top(\boldsymbol{w}'-\boldsymbol{w})+\frac{\beta}{2}||\boldsymbol{w}'-\boldsymbol{w}||^2\,.$$

Let $\mathbb{E}_t[\cdot]$ denote expectation conditioned on all random events up to time $t-1$. Then, by the convexity of f_{t+1}, we have

$$\mathbb{E}[f_{t+1}(\tilde{\boldsymbol{w}}_t)]=\mathbb{E}\big[\mathbb{E}_t[f_{t+1}(\tilde{\boldsymbol{w}}_t)]\big]\geq\mathbb{E}\big[f_{t+1}(\mathbb{E}_t[\tilde{\boldsymbol{w}}_t])\big]=\mathbb{E}\big[f_{t+1}(\mathbb{E}_t[\boldsymbol{w}_t])\big]=\mathbb{E}[f_{t+1}(\boldsymbol{w}_t)]\,.$$

Moreover, by the β-smoothness of f_{t+1}, we can write

$$\begin{aligned}
\mathbb{E}[f_{t+1}(\tilde{\boldsymbol{w}}_{t+1})] &= \mathbb{E}\big[\mathbb{E}_{t+1}[f_{t+1}(\tilde{\boldsymbol{w}}_{t+1})]\big] \\
&= \mathbb{E}\big[\mathbb{E}_{t+1}[f_{t+1}(\boldsymbol{w}_{t+1} + \delta\, H_{t+1}^{-1/2} s_{t+1})]\big] \\
&\leq \mathbb{E}\left[\mathbb{E}_{t+1}\left[f_{t+1}(\boldsymbol{w}_{t+1}) + \delta\, \nabla f_{t+1}(\boldsymbol{w}_{t+1})^\top H_{t+1}^{-1/2} s_{t+1} + \frac{\beta\delta^2}{2} s_{t+1}^\top H_{t+1}^{-1} s_{t+1}\right]\right] \\
&= \mathbb{E}\left[f_{t+1}(\boldsymbol{w}_{t+1}) + \mathbb{E}_{t+1}\left[\frac{\beta\delta^2}{2} s_{t+1}^\top H_{t+1}^{-1} s_{t+1}\right]\right] \\
&= \mathbb{E}\left[f_{t+1}(\boldsymbol{w}_{t+1})\right] + \frac{\beta\delta^2}{2}\,\mathbb{E}\left[||H_{t+1}^{-1/2} s_{t+1}||^2\right] \\
&\leq \mathbb{E}\left[f_{t+1}(\boldsymbol{w}_{t+1})\right] + \frac{\beta\delta^2\mathcal{D}^2}{2}\,,
\end{aligned}$$

the last inequality following from the properties of the Dikin ellipsoid associated with the self-concordant barrier Ψ, ensuring that $\boldsymbol{w}_{t+1} + H_{t+1}^{-1/2} s_{t+1}$ belongs to Ω, hence bounding $||H_{t+1}^{-1/2} s_{t+1}||$ by the diameter $\mathcal{D}$. Putting together, we have so far obtained

$$\begin{aligned}
\Big[\mathbb{E}\left[f_{t+1}(\tilde{\boldsymbol{w}}_{t+1}) - f_{t+1}(\tilde{\boldsymbol{w}}_t)\right]\Big]_+ &\leq \left[\mathbb{E}\left[f_{t+1}(\boldsymbol{w}_{t+1}) - f_{t+1}(\boldsymbol{w}_t)\right] + \frac{\beta\delta^2\mathcal{D}^2}{2}\right]_+ \\
&\leq \Big[\mathbb{E}\left[f_{t+1}(\boldsymbol{w}_{t+1}) - f_{t+1}(\boldsymbol{w}_t)\right]\Big]_+ + \frac{\beta\delta^2\mathcal{D}^2}{2}\,, \qquad (13)
\end{aligned}$$

where we have further used the fact that $[a]_+$ is nondecreasing in $a \in \mathbb{R}$, and that $[a+b]_+ \leq [a]_+ + [b]_+$ for all $a, b \in \mathbb{R}$.

Now, consider the Bregman divergence associated with the (strongly convex) barrier function Ψ:

$$B_\Psi(\boldsymbol{w}, \boldsymbol{w}') = \Psi(\boldsymbol{w}) - \Psi(\boldsymbol{w}') - \nabla\Psi(\boldsymbol{w}')^\top(\boldsymbol{w} - \boldsymbol{w}')\,.$$

Since the sequence $\{\boldsymbol{w}_t\}_{t=1\ldots T}$ is generated by a Follow The Regularized Leader (FTRL) algorithm, we have —see, e.g., (Hazan, 2016, Equation (5.2))

$$B_\Psi(\boldsymbol{w}_t, \boldsymbol{w}_{t+1}) \leq \eta \hat{g}_t^\top(\boldsymbol{w}_t - \boldsymbol{w}_{t+1}) \leq \eta ||\hat{g}_t||_t^*\, ||\boldsymbol{w}_t - \boldsymbol{w}_{t+1}||_t\,, \qquad (14)$$

where $||\cdot||_t$ is the local norm induced by the Hessian of Ψ at $\boldsymbol{w}_t$, i.e., $||\boldsymbol{w}||_t = \left(\boldsymbol{w}^\top \nabla^2\Psi(\boldsymbol{w}_t)\boldsymbol{w}\right)^{1/2}$, and $||\cdot||_t^*$ is its dual, $||\boldsymbol{w}||_t^* = \left(\boldsymbol{w}^\top(\nabla^2\Psi(\boldsymbol{w}_t))^{-1}\boldsymbol{w}\right)^{1/2}$. By the strong convexity of Ψ w.r.t. $||\cdot||$ we have

$$B_\Psi(\boldsymbol{w}_t, \boldsymbol{w}_{t+1}) \geq \frac{\alpha}{2}\,||\boldsymbol{w}_t - \boldsymbol{w}_{t+1}||^2\,,$$

for some constant $\alpha > 0$. Moreover, one can show that $||\hat{g}_t||_t^* \leq n/\delta$ (Saha and Tewari, 2011) and, provided $\eta \leq \frac{\delta}{16n}$, also that $||\boldsymbol{w}_t - \boldsymbol{w}_{t+1}||_t \leq 8\eta\, ||\hat{g}_t||_t^*$ (e.g., Abernethy et al. (2012) or Lemma 6.8 in Hazan (2016) applied to the self-concordant function $\boldsymbol{w} \to \eta \sum_{\tau=1}^{t-1} \boldsymbol{w}^\top \hat{g}_\tau + \Psi(\boldsymbol{w})$), so that putting together as in (14) gives

$$||\boldsymbol{w}_t - \boldsymbol{w}_{t+1}|| = \mathcal{O}\left(\frac{\eta\, n}{\delta}\right)\,, \qquad (15)$$

where the $\mathcal{O}$ notation hides here the inverse dependence on α. Finally, since f_t is [0,1]-bounded and β-smooth on a set of diameter $\mathcal{D}$, it must be that f_t is also Lipschitz with constant $L \leq \frac{2}{\mathcal{D}} + \mathcal{D}\beta$, so that combining with (13) and (15) yields

$$\begin{aligned}\Big[\mathbb{E}\left[f_{t+1}(\tilde{\boldsymbol{w}}_{t+1}) - f_{t+1}(\tilde{\boldsymbol{w}}_t)\right]\Big]_+ &\leq \left(\frac{2}{\mathcal{D}} + \mathcal{D}\beta\right)\mathbb{E}\left[||\boldsymbol{w}_{t+1} - \boldsymbol{w}_t||\right] + \frac{\beta\delta^2\mathcal{D}^2}{2} \\ &= \mathcal{O}\left(\left(\frac{1}{\mathcal{D}} + \mathcal{D}\beta\right)\frac{\eta\, n}{\delta} + \beta\delta^2\mathcal{D}^2\right)\,,\end{aligned}$$

as claimed. ■