JMLR: Workshop and Conference Proceedings vol 35:1–18, 2014

Online Learning with Composite Loss Functions

Ofer Dekel OFERD@MICROSOFT.COM
Microsoft Research

Jian Ding JIANDING@GALTON.UCHICAGO.EDU
University of Chicago

Tomer Koren TOMERK@TECHNION.AC.IL
Technion

Yuval Peres PERES@MICROSOFT.COM
Microsoft Research

Abstract

We study a new class of online learning problems where each of the online algorithm's actions is assigned an adversarial value, and the loss of the algorithm at each step is a known and deterministic function of the values assigned to its recent actions. This class includes problems where the algorithm's loss is the *minimum* over the recent adversarial values, the *maximum* over the recent values, or a *linear combination* of the recent values. We analyze the minimax regret of this class of problems when the algorithm receives bandit feedback, and prove that when the *minimum* or *maximum* functions are used, the minimax regret is $\widetilde{\Omega}(T^{2/3})$ (so called *hard* online learning problems), and when a linear function is used, the minimax regret is $\widetilde{O}(\sqrt{T})$ (so called *easy* learning problems). Previously, the only online learning problem that was known to be provably hard was the multi-armed bandit with switching costs.

1. Introduction

Online learning is often described as a T-round repeated game between a randomized player and an adversary. On each round of the game, the player and the adversary play simultaneously: the player (randomly) chooses an action from an action set $\mathcal{X}$ while the adversary assigns a loss value to each action in $\mathcal{X}$. The player then incurs the loss assigned to the action he chose. At the end of each round, the adversary sees the player's action and possibly adapts his strategy. This type of adversary is called an *adaptive* adversary (sometimes also called *reactive* or *non-oblivious*). In this paper, we focus on the simplest online learning setting, where $\mathcal{X}$ is assumed to be the finite set $\{1, \ldots, k\}$.

The adversary has unlimited computational power and therefore, without loss of generality, he can prepare his entire strategy in advance by enumerating over all possible action sequences and predetermining his response to each one. More formally, we assume that the adversary starts the game by choosing a sequence of T history-dependent loss functions, $f_1, \ldots, f_T$, where each $f_t : \mathcal{X}^t \mapsto [0, 1]$ (note that f_t depends on the player's entire history of t actions). With this, the adversary concludes his role in the game and only the player actively participates in the T rounds. On round t, the player (randomly) chooses an action X_t from the action set $\mathcal{X}$ and incurs the loss $f_t(X_{1:t})$ (where $X_{1:t}$ is our shorthand for the sequence $(X_1, \ldots, X_t)$). The player's goal is to accumulate a small total loss, $\sum_{t=1}^{T} f_t(X_{1:t})$.

At the end of each round, the player receives some feedback, which he uses to inform his choices on future rounds. We distinguish between different feedback models. The least informative feedback model we consider is *bandit feedback*, where the player observes his loss on each round, $f_t(X_{1:t})$, but nothing else. In other words, after choosing his action, the player receives a single real number. The prediction game with bandit feedback is commonly known as the *adversarial multi-armed bandit* problem (Auer et al., 2002) and the actions in $\mathcal{X}$ are called *arms*. A more informative feedback model is *full feedback* (also called *full information* feedback), where the player also observes the loss he would have incurred had he played a different action on the current round. In other words, the player receives $f_t(X_{1:(t-1)}, x)$ for each $x \in \mathcal{X}$, for a total of $|\mathcal{X}|$ real numbers on each round. The prediction game with full information is often called *prediction with expert advice* (Cesa-Bianchi et al., 1997) and each action is called an *expert*.

A third feedback model, the most informative of the three, is *counterfactual feedback*. In this model, at the end of round t, the player receives the complete definition of the loss function f_t. In other words, he receives the value of $f_t(x_1, \ldots, x_t)$ for all $(x_1 \ldots, x_t) \in \mathcal{X}^t$ (for a total of $|\mathcal{X}|^t$ real numbers). This form of feedback allows the player to answer questions of the form "how would the adversary have acted today had I played differently in the past?" This form of feedback is neglected in the literature, primarily because most of the existing literature focuses on *oblivious* adversaries (who do not adapt according to the player's past actions), for which counterfactual feedback is equivalent to full feedback.

Since the loss functions are adversarial, their values are only meaningful when compared to an adequate baseline. Therefore, we evaluate the player using the notion of *policy regret* (Arora et al., 2012), abbreviated simply as *regret*, and defined as

$$R = \sum_{t=1}^{T} f_t(X_1, \ldots, X_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x, \ldots, x) . \tag{1}$$

Policy regret compares the player's cumulative loss to the loss of the best policy in hindsight that repeats a single action on all T rounds. The player's goal is to minimize his regret against a worst-case sequence of loss functions. We note that a different definition of regret, which we call *standard regret*, is popular in the literature. However, Arora et al. (2012) showed that standard regret is completely inadequate for analyzing the performance of online learning algorithms against adaptive adversaries, so we stick the definition of regret in Eq. (1)

While regret measures a specific player's performance against a specific sequence of loss functions, the inherent difficulty of the game itself is measured by *minimax regret*. Intuitively, minimax regret is the expected regret of an optimal player, when he faces an optimal adversary. More formally, minimax regret is the minimum over all randomized player strategies, of the maximum over all loss sequences, of $\mathbb{E}[R]$. If the minimax regret grows sublinearly with T, it implies that the per-round regret rate, $R(T)/T$, must diminish with the length of the game T. In this case, we say that the game is *learnable*. Arora et al. (2012) showed that without additional constraints, online learning against an adaptive adversary has a minimax regret of $\Theta(T)$, and is therefore unlearnable. This motivates us to weaken the adaptive adversary and study the minimax regret when we restrict the sequence of loss functions in different ways.

Easy online learning problems. For many years, the standard practice in online learning research was to find online learning settings for which the minimax regret is $\widetilde{\Theta}(\sqrt{T})$. Following Antos et al. (2012), we call problems for which the minimax regret is $\widetilde{\Theta}(\sqrt{T})$ *easy* problems. Initially, minimax regret bounds focused on loss functions that are generated by an *oblivious* adversary. An oblivious adversary does not adapt his loss values to the player's past actions. More formally, this type of adversary first defines a sequence of single-input functions, $\ell_1, \ldots, \ell_T$, where each $\ell_t : \mathcal{X} \mapsto [0,1]$, and then sets

$$\forall\, t \quad f_t(x_1, \ldots, x_t) \;=\; \ell_t(x_t) \enspace .$$

When the adversary is oblivious, the definition of regret used in this paper (Eq. (1)) and the aforementioned *standard regret* are equivalent, so all previous work on oblivious adversaries is relevant in our setting. In the full feedback model, the *Hedge* algorithm (Littlestone and Warmuth, 1994; Freund and Schapire, 1997) and the *Follow the Perturbed Leader* algorithm (Kalai and Vempala, 2005) both guarantee a regret of $\widetilde{O}(\sqrt{T})$ on any oblivious loss sequence (where $\widetilde{O}$ ignores logarithmic terms). A matching lower bound of $\Omega(\sqrt{T})$ appears in Cesa-Bianchi and Lugosi (2006), and allows us to conclude that the minimax regret in this setting is $\widetilde{\Theta}(\sqrt{T})$. In the bandit feedback model, the *Exp3* algorithm (Auer et al., 2002) guarantees a regret of $\widetilde{O}(\sqrt{T})$ against any oblivious loss sequence and implies that the minimax regret in this setting is also $\widetilde{\Theta}(\sqrt{T})$.

An adversary that is slightly more powerful than an oblivious adversary is the *switching cost* adversary, who penalizes the player each time his action is different than the action he chose on the previous round. Formally, the switching cost adversary starts by defining a sequence of single-input functions $\ell_1, \ldots, \ell_T$, where $\ell_t : \mathcal{X} \mapsto [0,1]$, and uses them to set

$$\forall\, t \quad f_t(x, x') \;=\; \ell_t(x') + \mathbb{1}_{x' \neq x} \enspace . \tag{2}$$

Note that the range of f_t is $[0,2]$ instead of $[0,1]$; if this is a problem, it can be easily resolved by replacing $f_t \leftarrow f_t/2$ throughout the analysis. In the full feedback model, the *Follow the Lazy Leader* algorithm (Kalai and Vempala, 2005) and the more recent *Shrinking Dartboard* algorithm (Geulen et al., 2010) both guarantee a regret of $\widetilde{O}(\sqrt{T})$ against any oblivious sequence with a switching cost. The $\Omega(\sqrt{T})$ lower bound against oblivious adversaries holds in this case, and the minimax regret is therefore $\widetilde{\Theta}(\sqrt{T})$.

The switching cost adversary is a special case of a 1-*memory* adversary, who is constrained to choose loss functions that depend only on the player's last two actions (his current action and the previous action). More generally, the m-*memory* adversary chooses loss functions that depend on the player's last $m+1$ actions (the current action plus m previous actions), where m is a parameter. In the counterfactual feedback model, the work of Gyorgy and Neu (2011) implies that the minimax regret against an m-*memory* adversary is $\widetilde{\Theta}(\sqrt{T})$.

Hard online learning problems. Recently, Cesa-Bianchi et al. (2013); Dekel et al. (2013) showed that online learning against a switching cost adversary with bandit feedback (more popularly known as the *multi-armed bandit with switching costs*) has a minimax regret of $\widetilde{\Theta}(T^{2/3})$. This result proves that there exists a natural[1] online learning problem that is learnable, but at a rate that is substantially slower then $\widetilde{\Theta}(\sqrt{T})$. Again following Antos et al. (2012), we say that an online problem is *hard* if its minimax regret is $\widetilde{\Theta}(T^{2/3})$.

Is the multi-armed bandit with switching costs a one-off example, or are there other natural hard online learning problems? In this paper, we answer this question by presenting another hard online learning setting, which is entirely different than the multi-armed bandit with switching costs.

1. By *natural*, we mean that the problem setting can be described succinctly, and that the parameters that define the problem are all independent of T. An example of an unnatural problem with a minimax regret of $\Theta(T^{2/3})$ is the multi-armed bandit problem with $k = T^{1/3}$ arms.

Composite loss functions. We define a family of adversaries that generate *composite loss functions*. An adversary in this class is defined by a memory size $m \geq 0$ and a *loss combining function* $g : [0,1]^{m+1} \mapsto [0,1]$, both of which are fixed and known to the player. The adversary starts by defining a sequence of oblivious functions $\ell_1, \ldots, \ell_T$, where each $\ell_t : \mathcal{X} \mapsto [0,1]$. Then, he uses g and $\ell_{1:T}$ to define the composite loss functions

$$\forall\, t \quad f_t(x_{1:t}) \;=\; g\big(\ell_{t-m}(x_{t-m}), \ldots, \ell_t(x_t)\big) \enspace .$$

For completeness, we assume that $\ell_t \equiv 0$ for $t \leq 0$. The adversary defined above is a special case of a m-memory adversary.

For example, we could set $m = 1$ and choose the *max* function as our loss combining function. This choice define a 1-memory adversary, with loss functions given by

$$\forall\, t \quad f_t(x_{1:t}) \;=\; \max\big(\ell_{t-1}(x_{t-1}), \ell_t(x_t)\big) \enspace .$$

In words, the player's action on each round is given an oblivious value and the loss at time t is the maximum of the current oblivious value and previous one. For brevity, we call this adversary the *max-adversary*. The max-adversary can be used to represent online decision-making scenarios where the player's actions have a prolonged effect, and a poor choice on round t incurs a penalty on round t and again on round $t+1$. Similarly, setting $m = 1$ and choosing *min* as the combining function gives the *min adversary*. This type of adversary models scenarios where the environment forgives poor action choices whenever the previous choice was good. Finally, one can also consider choosing a linear function g. Examples of linear combining functions are

$$f_t(x_{1:t}) \;=\; \frac{1}{2}\big(\ell_{t-1}(x_{t-1}) + \ell_t(x_t)\big) \quad \text{and} \quad f_t(x_{1:t}) \;=\; \ell_{t-1}(x_{t-1}) \enspace .$$

The main technical contribution of this paper is a $\widetilde{\Omega}(T^{2/3})$ lower bound on the minimax regret against the *max* and *min* adversaries, showing that each of them induces a *hard* online learning problem when the player receives bandit feedback. In contrast, we show that any linear combining function induces an *easy* bandit learning problem, with a minimax regret of $\widetilde{\Theta}(\sqrt{T})$. Characterizing the set of combining functions that induce hard bandit learning problems remains an open problem.

Recall that in the bandit feedback model, the player only receives one number as feedback on each round, namely, the value of $f_t(X_{1:t})$. If the loss is a composite loss, we could also consider a setting where the feedback consists of the single number $\ell_t(X_t)$. Since the combining function g is known to the player, he could use the observed values $\ell_1(X_1), \ldots, \ell_t(X_t)$ to calculate the value of $F_t(X_{1:t})$; this implies that this alternative feedback model gives the player more information than the strict bandit feedback model. However, it turns out that the $\widetilde{\Omega}(T^{2/3})$ lower bound holds even in this alternative feedback model, so our analysis below assumes that the player observes $\ell_t(X_t)$ on each round.

Organization. This paper is organized as follows. In Sec. 2, we recall the analysis in Dekel et al. (2013) of the minimax regret of the multi-armed bandit with switching costs. Components of this analysis play a central role in the lower bounds against the composite loss adversary. In Sec. 3 we prove a lower bound on the minimax regret against the min-adversary in the bandit feedback setting, and in Sec. 3.3 we comment on how to prove the same for the max-adversary. A proof that linear combining functions induce easy online learning problems is given in Sec. 4. We conclude in Sec. 5.

2. The Multi-Armed Bandit with Switching Costs

In this section, we recall the analysis in Dekel et al. (2013), which proves a $\widetilde{\Omega}(T^{2/3})$ lower bound on the minimax regret of the multi-armed bandit problem with switching costs. The new results in the sections that follow build upon the constructions and lemmas in Dekel et al. (2013). For simplicity, we focus on the 2-armed bandit with switching costs, namely, we assume that $\mathcal{X} = \{0, 1\}$ (see Dekel et al. (2013) for the analysis with arbitrary k).

First, like many other lower bounds in online learning, we apply (the easy direction of) Yao's minimax principle (Yao, 1977), which states that the regret of a randomized player against the worst-case loss sequence is greater or equal to the minimax regret of an optimal *deterministic* player against a *stochastic* loss sequence. In other words, moving the randomness from the player to the adversary can only make the problem easier for the player. Therefore, it suffices to construct a *stochastic* sequence of loss functions[2], $F_{1:T}$, where each F_t is a random oblivious loss function with a switching cost (as defined in Eq. (2)), such that

$$\mathbb{E}\left[\sum_{t=1}^{T} F_t(X_{1:t}) \;-\; \min_{x \in \mathcal{X}} \sum_{t=1}^{T} F_t(x, \ldots, x)\right] \;=\; \widetilde{\Omega}(T^{2/3}) \;, \tag{3}$$

for any deterministic player strategy.

We begin be defining a stochastic process $W_{0:T}$. Let $\xi_{1:T}$ be T independent zero-mean Gaussian random variables with variance σ^2, where σ is specified below. Let $\rho : [T] \mapsto \{0\} \cup [T]$ be a function that assigns each $t \in [T]$ with a *parent* $\rho(t)$. For now, we allow ρ to be any function that satisfies $\rho(t) < t$ for all t. Using $\xi_{1:T}$ and ρ, we define

$$\begin{aligned} W_0 \;&=\; 0 \;, \\ \forall\, t \in [T] \quad W_t \;&=\; W_{\rho(t)} + \xi_t \;\; . \end{aligned} \tag{4}$$

Note that the constraint $\rho(t) < t$ guarantees that a recursive application of ρ always leads back to zero. The definition of the parent function ρ determines the behavior of the stochastic processes. For example, setting $\rho(t) = 0$ implies that $W_t = \xi_t$ for all t, so the stochastic process is simply a sequence of i.i.d. Gaussians. On the other hand, setting $\rho(t) = t - 1$ results in a Gaussian random walk. Other definitions of ρ can create interesting dependencies between the variables. The specific setting of ρ that satisfies our needs is defined below.

Next, we explain how the stochastic process $W_{1:T}$ defines the stochastic loss functions $F_{1:T}$. First, we randomly choose one of the two actions to be the *better action* by drawing an unbiased Bernoulli χ ($\mathbb{P}(\chi = 0) = \mathbb{P}(\chi = 1)$). Then we let ϵ be a positive *gap* parameter, whose value is specified below, and we set

$$\forall t \quad Z_t(x) \;=\; W_t + \frac{1}{2} - \epsilon \mathbb{1}_{x=\chi} \;\; . \tag{5}$$

Note that $Z_t(\chi)$ is always smaller than $Z_t(1 - \chi)$ by a constant gap of ϵ. Each function in the sequence $Z_{1:T}$ can take values on the entire real line, whereas we require bounded loss functions. To resolve this, we confine the values of $Z_{1:T}$ to the interval $[0, 1]$ by applying a clipping operation,

$$\forall\, t \quad L_t(x) \;=\; \mathrm{clip}(Z_t(x)) \;\; , \quad \text{where } \mathrm{clip}(\alpha) \;=\; \min\{\max\{\alpha, 0\}, 1\} \;\; . \tag{6}$$

2. We use the notation $U_{i:j}$ as shorthand for the sequence $U_i, \ldots, U_j$ throughout.

The sequence $L_{1:T}$ should be thought of as a stochastic oblivious loss sequence. Finally, as in Eq. (2), we add a switching cost and define the sequence of loss functions

$$F_t(x_{1:T}) \;=\; L_t(x_t) + \mathbb{1}_{x' \neq x} \;\; .$$

It remains to specify the parent function ρ, the standard deviation σ, and the gap ϵ. With the right settings, we can prove that $F_{1:T}$ is a stochastic loss sequence that satisfies Eq. (3).

We take a closer look at the parent function ρ. First, we define the *ancestors* of round t, denoted by $\rho^*(t)$, to be the set of positive indices that are encountered when ρ is applied recursively to t. Formally, $\rho^*(t)$ is defined recursively as

$$\begin{aligned} \rho^*(0) \;&=\; \{\} \\ \forall\, t \quad \rho^*(t) \;&=\; \rho^*\big(\rho(t)\big) \;\cup\; \{\rho(t)\} \;\; . \end{aligned} \tag{7}$$

Using this definition, the *depth* of ρ is defined as the size of the largest set of ancestors, $d(\rho) = \max_{t\in[T]} |\rho^*(t)|$. The depth is a key property of ρ and the value of $d(\rho)$ characterizes the extremal values of $W_{1:T}$: by definition, there exists a round t such that W_t is the sum of $d(\rho)$ independent Gaussians, so the typical value of $|W_t|$ is bounded by $\sigma\sqrt{d(\rho)}$. More precisely, Lemma 1 in Dekel et al. (2013) states that

$$\forall\, \delta \in (0,1) \qquad \mathbb{P}\left(\max_{t\in[T]} |W_t| \le \sigma\sqrt{2d(\rho)\log\tfrac{T}{\delta}}\right) \;\ge\; 1-\delta\;. \tag{8}$$

The clipping operation defined in Eq. (6) ensures that the loss is bounded, but the analysis requires that the unclipped sequence $Z_{1:t}$ already be bounded in $[0,1]$ with high probability. This implies that we should choose

$$\sigma \;\sim\; \Big(d(\rho)\log\big(\tfrac{T}{\delta}\big)\Big)^{-1/2} \;\; . \tag{9}$$

Another important property of ρ is its *width*. First, define the *cut* on round t as

$$\mathrm{cut}(t) \;=\; \{s \in [T] \;:\; \rho(s) < t \le s\} \;\; .$$

In words, the cut on round t is the set of rounds that are separated from their parent by t. The *width* of ρ is then defined as the size of the largest cut, $w(\rho) = \max_{t\in[T]} |\mathrm{cut}(t)|$.

The analysis in Dekel et al. (2013) characterizes the player's ability to statistically estimate the value of χ (namely, to uncover the identity of the better action) as a function of the number of switches he performs. Each time the player switches actions, he has an opportunity to collect statistical information on the identity of χ. The amount of information revealed to the player with each switch is controlled by the depth and width of ρ and the values of ϵ and σ. Formally, define the conditional probability measures

$$\mathcal{Q}_0(\cdot) \;=\; \mathbb{P}(\cdot \mid \chi = 0) \qquad \text{and} \qquad \mathcal{Q}_1(\cdot) \;=\; \mathbb{P}(\cdot \mid \chi = 1) \;\; . \tag{10}$$

In words, $\mathcal{Q}_0$ is the conditional probability when action 0 is better and Q_1 is the conditional probability when action 1 is better. Also, let $\mathcal{F}$ be the σ-algebra generated by the player's observations throughout the game, $L_1(X_1), \ldots, L_T(X_T)$. Since the player's actions are a deterministic function of the loss values that he observes, his sequence of actions is measurable by $\mathcal{F}$. The *total variation* distance between $\mathcal{Q}_0$ and $\mathcal{Q}_1$ on $\mathcal{F}$ is defined as

$$d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, Q_1) \;=\; \sup_{A\in\mathcal{F}} \big|\mathcal{Q}_0(A) - \mathcal{Q}_1(A)\big| \;\; .$$

Dekel et al. (2013) proves the following bound on $d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, Q_1)$.

Lemma 1 *Let $F_{1:T}$ be the stochastic loss sequence defined above by the parent function ρ, with variance σ^2 and gap ϵ. Fix a deterministic player and let M be the number of switches he performs as he plays the online game. Then,*

$$d_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, Q_1) \;\le\; \frac{\epsilon}{\sigma}\sqrt{w(\rho)\,\mathbb{E}[M]}\;\;,$$

where Q_0 and Q_1 are as defined in Eq. (10).

Intuitively, the lemma states that if $\mathbb{E}[M]$ is asymptotically smaller than $\frac{\sigma^2}{\epsilon^2 w(\rho)}$ then any $\mathcal{F}$-measurable event (e.g., the event that $X_{10} = 1$ or the event that the player switches actions on round 20) is almost equally likely to occur, whether $\chi = 0$ or $\chi = 1$. In other words, if the player doesn't switch often enough, then he certainly cannot identify the better arm.

Our goal is to build a stochastic loss sequence that forces the player to perform many switches, and Lemma 1 tells us that we must choose a parent function ρ that has a small width. Additionally, setting the variance σ^2 according to Eq. (9) also implies that we want ρ to have a small depth. Dekel et al. (2013) defines the parent function $\rho(t) = t - \gcd(t, 2^T)$ (where $\gcd(\alpha, \beta)$ is the greatest common divisor of α and β). Put another way, ρ takes the number t, finds its binary representation, identifies the least significant bit that equals 1, and flips that bit to zero. It them proves that $d(\rho) = \Theta(\log T)$ and $w(\rho) = \Theta(\log T)$.

The lower bound on the minimax regret of the multi-armed bandit with switching costs is obtained by setting $\epsilon = \Theta(T^{-1/3}/\log T)$. If the expected number of switches is small, namely $\mathbb{E}[M] \le T^{2/3}/\log^2 T$, then Lemma 1 implies that the player cannot identify the better action. From there, it is straightforward to show that the player has a positive probability of choosing the worse action on each round, resulting in a regret of $R = \Theta(\epsilon T)$. Plugging in our choice of ϵ proves that $R = \widetilde{\Omega}(T^{2/3})$. On the other hand, if the number of switches is large, namely, $\mathbb{E}[M] > T^{2/3}/\log^2 T$, then the regret is $\Omega(T^{2/3})$ directly due to the switching cost.

Many of the key constructions and ideas behind this proof are reused below.

3. The Min Adversary is Hard

In this section, we lower bound the minimax regret against the min-adversary in the bandit feedback model where the player only observes a single number, $\ell_t(X_t)$, at the end of round t. The full proof is rather technical, so we begin with a high level proof sketch. As in Sec. 2, Yao's minimax principle once again reduces our problem to one of finding a stochastic loss sequence $L_1, \ldots, L_T$ that forces all deterministic algorithms to incur a regret of $\widetilde{\Omega}(T^{2/3})$. The main idea is to repeat the construction presented in Sec. 2 by simulating a switching cost using the min combining function.

We start with a stochastic process that is defined by a parent function ρ, similar to the sequence $W_{1:T}$ defined in Sec. 2 (although we require a different parent function than the one defined there). Again, we draw a Bernoulli χ that determines the better of the two possible actions, we choose a *gap* parameter ϵ, and we define the sequence of functions $Z_{1:T}$, as in Eq. (5). This sequence has the important property that, in the bandit feedback model, it reveals information on the value of χ only when the player switches actions.

Next, we identify triplets of rounds, $(t-1, t, t+1)$, where $|W_{t-1} - W_t| \le \tau$ (τ is a *tolerance* parameter, chosen so that $\tau \gg \epsilon$) and some other technical properties hold. Then, we simulate a switching cost on round t by adding a pair of *spikes* to the loss values of the two actions, one on rounds $t-1$ and one on round t. We choose a *spike size* η (such that $\eta \gg \tau$), we draw an unbiased Bernoulli Λ_t, and we set

$$L_{t-1}(x) \;=\; \mathrm{clip}\big(Z_{t-1}(x) + \eta \mathbb{1}_{x \neq \Lambda_t}\big) \qquad \text{and} \qquad L_t(x) \;=\; \mathrm{clip}\big(Z_t(x) + \eta \mathbb{1}_{x = \Lambda_t}\big) \;\; ,$$

where clip() is defined in Eq. (6). In words, with probability $\frac{1}{2}$ we add a spike of size η to the loss of action 0 on round $t-1$ and to the loss of action 1 on round t, and with probability $\frac{1}{2}$ we do the opposite.

Finally, we define the loss on round t using the min combining function

$$F_t(x_{1:t}) \;=\; \min\big(L_{t-1}(x_{t-1}), L_t(x_t)\big) \;\; . \tag{11}$$

We can now demonstrate how the added spikes simulate a switching cost on the order of η. Say that the player switches actions on round t, namely, $X_t \neq X_{t-1}$. Since Λ_t is an independent unbiased Bernoulli, it holds that $X_t = \Lambda_t$ with probability $\frac{1}{2}$. If $X_t = \Lambda_t$, then the player encounters both of the spikes: $L_t(X_t) = Z_t(X_t) + \eta$ and $L_{t-1}(X_{t-1}) = Z_{t-1}(X_{t-1}) + \eta$. Recall that $|Z_{t-1}(0) - Z_{t-1}(1)| \leq \epsilon$ and $|Z_{t-1}(x) - Z_t(x)| \leq \tau$, so

$$F_t(X_{1:t}) \;\in\; \big[Z_t(0) + \eta - (\epsilon + \tau),\; Z_t(0) + \eta + (\epsilon + \tau)\big] \;\; . \tag{12}$$

On the other hand, if the player does not switch actions on round t, his loss then satisfies

$$F_t(X_{1:t}) \;\in\; \big[Z_t(0) - (\epsilon + \tau),\; Z_t(0) + (\epsilon + \tau)\big] \;\; . \tag{13}$$

Comparing the intervals in Eq. (12) and Eq. (13), and recalling that $\eta \gg (\epsilon + \tau)$, we conclude that, with probability $\frac{1}{2}$, the switch caused the player's loss to increase by η. This is the general scheme by which we simulate a switching cost using the min combining function.

There are a several delicate issues that were overlooked in the simplistic proof sketch, and we deal with then below.

3.1. The Stochastic Loss Sequence

We formally describe the stochastic loss sequence used to prove our lower bound. In Sec. 2, we required a deterministic parent function ρ with depth $d(\rho)$ and width $w(\rho)$ that scale logarithmically with T. To lower-bound the minimax regret against the min adversary, we need a *random* parent function for which $d(\rho)$ and $w(\rho)$ are both logarithmic with high probability, and such that $\rho(t) = t-1$ with probability at least $\frac{1}{2}$ for all t. The following lemma proves that such a random parent function exists. The proof is deferred to Appendix A.

Lemma 2 *For any time horizon T, there exists a random function $\rho : [T] \mapsto \{0\} \cup [T]$ with $\rho(t) < t$ for all $t \in [T]$ such that*

- $\forall t \quad \mathbb{P}\big(\rho(t) = t-1 \mid \rho(1), \ldots, \rho(t-1)\big) \geq \frac{1}{2}$*;*
- $w(\rho) \leq \log T + 1$ *with probability* 1*;*
- $d(\rho) = O(\log T)$ *with probability* $1 - O(T^{-1})$.

Let ρ be a random parent function, as described above, and use this ρ to define the loss sequence $Z_{1:T}$, as outlined in Sec. 2. Namely, draw independent zero-mean Gaussians $\xi_{1:T}$ with variance σ^2. Using ρ and $\xi_{1:T}$, define the stochastic process $W_{1:T}$ as specified in Eq. (4). Finally, choose the better arm by drawing an unbiased Bernoulli χ, set a gap parameter ϵ, and use $W_{1:T}$, χ, and ϵ to define the loss sequence $Z_{1:T}$, as in Eq. (5).

Next, we augment the loss sequence $Z_{1:T}$ in a way that simulates a switching cost. For all $2 \leq t \leq T-2$, let E_t be the following event:

$$E_t = \left\{ |W_{t-1} - W_t| \leq \tau \quad \text{and} \quad W_{t+1} < W_t - \tau \quad \text{and} \quad W_{t+2} < W_{t+1} - \tau \right\}, \tag{14}$$

where τ is a tolerance parameter defined below. In other words, E_t occurs if the stochastic process $W_{1:T}$ remains rather flat between rounds $t-1$ and t, and then drops on rounds $t+1$ and $t+2$. We simulate a switching cost on round t if and only if E_t occurs.

We simulate the switching cost by adding pairs of *spikes*, one to the loss of each action, one on round $t-1$ and one on round t. Each spike has an orientation: it either penalizes a switch from action 0 to action 1, or a switch from action 1 to action 0. The orientation of each spike is chosen randomly, as follows. We draw independent unbiased Bernoullis $\Lambda_{2:T-1}$; if a spike is added on round t, it penalizes a switch from action $X_{t-1} = 1 - \Lambda_t$ to action $X_t = \Lambda_t$. Formally, define

$$S_t(x) \;=\; \begin{cases} \eta & \text{if } (E_t \;\wedge\; x = \Lambda_t) \vee (E_{t+1} \;\wedge\; x \neq \Lambda_{t+1}) \\ 0 & \text{otherwise} \end{cases} ,$$

where η is a *spike size* parameter (defined below). Finally, define $L_t(x) = \text{clip}\big(Z_t(x) + S_t(x)\big)$. This defines the sequence of oblivious functions. The min adversary uses these functions to define the loss functions $F_{1:T}$, as in Eq. (11).

In the rest of the section we prove that the regret of any deterministic player against the loss sequence $F_{1:T}$ is $\widetilde{\Omega}(T^{2/3})$. Formally, we prove the following theorem.

Theorem 3 *Let $F_{1:T}$ be the stochastic sequence of loss functions defined above. Then, the expected regret (as defined in Eq. (1)) of any deterministic player against this sequence is $\widetilde{\Omega}(T^{2/3})$.*

3.2. Analysis

For simplicity, we allow ourselves to neglect the clipping operator used in the definition of the loss sequence, an we simply assume that $L_t(x) = Z_t(x) + S_t(x)$. The additional steps required to reintroduce the clipping operator are irrelevant to the current analysis and can be copied from Dekel et al. (2013).

Fix a deterministic algorithm and let $X_1, \ldots, X_T$ denote the random sequence of actions it chooses upon the stochastic loss functions $F_{1:T}$. We define the algorithm's instantaneous (per-round) regret as

$$\forall t \quad R_t \;=\; \min\big(L_{t-1}(X_{t-1}), L_t(X_t)\big) - \min\big(L_{t-1}(\chi), L_t(\chi)\big) \;, \tag{15}$$

and note that our goal is to lower-bound $\mathbb{E}[R] = \sum_{t=1}^{T} \mathbb{E}[R_t]$.

The main technical difficulty of our analysis is getting a handle on the player's ability to identify the occurrence of E_t. If the player could confidently identify E_t on round $t-1$, he could avoid switching on round t. If the player could identify E_t on round t or $t+1$, he could safely switch on round $t+1$ or $t+2$, as E_t cannot co-occur with either E_{t+1} or E_{t+2}. To this end, we define the following sequence of random variables,

$$\forall t \quad \tilde{R}_t \;=\; \min\big(L_{t-1}(X_{t-1}), L_t(X_{t-1})\big) - \min\big(L_{t-1}(\chi), L_t(\chi)\big) \; . \tag{16}$$

The variable $\tilde{R}_t$ is similar to R_t, except that L_t is evaluated on the previous action X_{t-1} rather than the current action X_t. We think of $\tilde{R}_t$ as the instantaneous regret of a player that decides beforehand (before observing the value of $L_{t-1}(X_{t-1})$) not to switch on round t. It turns out that $\tilde{R}_t$ is much easier to analyze, since the player's decision to switch becomes independent of the occurrence of E_t. Specifically, we use $\tilde{R}_t$ to decompose the expected regret as

$$\mathbb{E}[R_t] \;=\; \mathbb{E}[R_t - \tilde{R}_t] \;+\; \mathbb{E}[\tilde{R}_t] \; .$$

We begin the analysis by clarifying the requirement that the event E_t only occurs if $W_{t+1} \leq W_t - \tau$. This requirement serves two separate roles: first, it prevents E_t and E_{t+1} from co-occurring and thus prevents overlapping spikes; second, this requirement prevents E_{t-1} from contributing to the player's loss on round t. This latter property is used throughout our analysis and is formalized in the following lemma.

Lemma 4 *If E_{t-1} occurs then $\tilde{R}_t = Z_t(X_{t-1}) - Z_t(\chi)$ and $R_t - \tilde{R}_t = Z_t(X_t) - Z_t(X_{t-1})$.*

In particular, the lemma shows that the occurrence of E_{t-1} cannot make $F_t(X_{1:t})$ be less than $F_t(\chi, \ldots, \chi)$. This may not be obvious at first glance: the occurrence of E_{t-1} contributes a spike on round t and if that spike is added to χ (the better action), one might imagine that this spike could contribute to $F_t(\chi, \ldots, \chi)$.

It is convenient to modify the algorithm and fix $X_s = X_t$ for all $s > t$ if $|L_t(X_t) - L_{\rho(t)}(X_{\rho(t)})| \geq 4\sigma\sqrt{\log T}$. Note that this event has probability $O(T^{-4})$ for each t, so the modification has a negligible effect on the regret. Recall that $\mathbb{E}[R_t] = \mathbb{E}[R_t - \tilde{R}_t] + \mathbb{E}[\tilde{R}_t]$; we first claim that $\mathbb{E}[\tilde{R}_t]$ is non-negative.

Lemma 5 *For any $1 < t < T$, let $\tilde{R}_t$ be as defined in Eq. (16). Then, it holds that $\mathbb{E}[\tilde{R}_t \mid X_{t-1} = \chi] = 0$ and $\mathbb{E}[\tilde{R}_t \mid X_{t-1} \neq \chi] = \epsilon$.*

Next, we turn to lower bounding $\mathbb{E}[R_t - \tilde{R}_t]$.

Lemma 6 *For any $1 < t < T$, let R_t be the player's instantaneous regret, as defined in Eq. (15), and let $\tilde{R}_t$ be as defined in Eq. (16). Then $\mathbb{E}[R_t - \tilde{R}_t] = \mathbb{P}(X_t \neq X_{t-1}) \cdot \Omega(\eta\tau/\sigma)$, provided that $\tau = o(\eta)$ and that $\epsilon = o(\eta\tau/\sigma)$.*

The proofs of the above lemmas are deferred to Appendix A. We can now prove our main theorem.

Proof of Theorem 3 We prove the theorem by distinguishing between two cases, based on the expected number of switches performed by the player. More specifically, let M be the number of switches performed by the player throughout the game.

First, assume that $\mathbb{E}[M] \geq T^{2/3}/\log^2 T$. Summing the lower-bounds in Lemma 5 and Lemma 6 over all t gives

$$\mathbb{E}[R] \;\geq\; \sum_{t=1}^{T} \mathbb{P}(X_t \neq X_{t-1}) \cdot \Omega(\eta\tau/\sigma) \;=\; \Omega(\eta\tau/\sigma) \cdot \mathbb{E}[M] \; .$$

Setting $\eta = \log^{-2} T$, $\sigma = \log^{-1} T$, $\tau = \log^{-5} T$, $\epsilon = T^{-1/3}/\log T$ (note that all of the constraints on these values specified in Lemma 6 are met) and plugging in our assumption that $\mathbb{E}[M] \geq T^{2/3}/\log^2 T$ gives the lower bound $\mathbb{E}[R] = \Omega\left(T^{2/3}/\log^6 T\right)$.

Next, we assume that $\mathbb{E}[M] < T^{2/3}/\log^2 T$. For any concrete instance of ρ, define the conditional probability measures

$$\mathcal{Q}_0^\rho(\cdot) \;=\; \mathbb{P}(\cdot \mid \chi = 0, \rho) \qquad \text{and} \qquad \mathcal{Q}_1^\rho(\cdot) \;=\; \mathbb{P}(\cdot \mid \chi = 1, \rho) \;\; .$$

We can apply Lemma 1 for any concrete instance of ρ and get $d_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0^\rho, Q_1^\rho) \leq (\epsilon/\sigma)\sqrt{w(\rho)\,\mathbb{E}[M \mid \rho]}$. Taking expectation on both sides of the latter inequality, we get

$$d_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, Q_1) \;\leq\; \mathbb{E}\big[d_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0^\rho, Q_1^\rho)\big] \;\leq\; \frac{\epsilon}{\sigma}\mathbb{E}\big[\sqrt{w(\rho)\,\mathbb{E}[M \mid \rho]}\big] \;\leq\; \frac{\epsilon}{\sigma}\sqrt{\mathbb{E}[w(\rho)]\,\mathbb{E}[M]} \;\; ,$$

where the inequality on the left is due to Jensen's inequality, and the inequality on the right is due to an application of the Cauchy-Schwartz inequality. Plugging in $\epsilon = T^{-1/3}/\log T$, $\sigma = \log^{-1} T$, $\mathbb{E}[w(\rho)] = \Theta(\log T)$ and $\mathbb{E}[M] = O(T^{2/3}/\log^2 T)$, we conclude that

$$d_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, Q_1) \;=\; o(1) \;\; . \tag{17}$$

Again, we decompose $\mathbb{E}[R_t] = \mathbb{E}[\tilde{R}_t] + \mathbb{E}[R_t - \tilde{R}_t]$, but this time we use the fact that Lemma 6 implies $\mathbb{E}[R_t - \tilde{R}_t] \geq 0$, and we focus on lower-bounding $\mathbb{E}[\tilde{R}_t]$. We decompose

$$\mathbb{E}[\tilde{R}_t] \;=\; \mathbb{P}(X_{t=1} = \chi)\,\mathbb{E}[\tilde{R}_t \mid X_{t=1} = \chi] \;+\; \mathbb{P}(X_{t=1} \neq \chi)\,\mathbb{E}[\tilde{R}_t \mid X_{t=1} \neq \chi] \;\; . \tag{18}$$

The first summand on the right-hand side above trivially equals zero. Lemma 5 proves that $\mathbb{E}[\tilde{R}_t \mid X_{t=1} \neq \chi] = \epsilon$. We use Eq. (17) to bound

$$\begin{aligned}
\mathbb{P}(X_{t-1} \neq \chi) \;&=\; \frac{1}{2}\,\mathbb{P}(X_{t-1} = 0 \mid \chi = 1) + \frac{1}{2}\,\mathbb{P}(X_{t-1} = 1 \mid \chi = 0) \\
&\geq\; \frac{1}{2}\,\mathbb{P}(X_{t-1} = 0 \mid \chi = 1) + \frac{1}{2}\,\big(\mathbb{P}(X_{t-1} = 1 \mid \chi = 1) - o(1)\big) \;=\; \frac{1}{2} - o(1) \;\; .
\end{aligned}$$

Plugging everything back into Eq. (18) gives $\mathbb{E}[\tilde{R}_t] = \Theta(\epsilon)$. We conclude that

$$\mathbb{E}[R] \;\geq\; \sum_{t=1}^{T} \mathbb{E}[\tilde{R}_t] \;=\; \Theta(T\epsilon) \;\; .$$

Recalling that $\epsilon = T^{-1/3}/\log T$ concludes the analysis. ■

3.3. The Max Adversary

In the previous section, we proved that the minimax regret, with bandit feedback, against the min adversary is $\widetilde{\Omega}(T^{2/3})$. The same can be proved for the max adversary, using an almost identical proof technique, namely, by using the max combining function to simulate a switching cost. The construction of the loss process $Z_{1:T}$ remains as defined above. The event E_t changes, and requires $|W_{t-1} - W_t| \leq \tau$ and $W_{t+1} > W_t + \eta$. The spikes also change: we set

$$S_{t-1}(\Lambda_t) = 1, \quad S_t(\Lambda_t) = 1, \quad S_{t-1}(1 - \Lambda_t) = 0, \quad S_t(1 - \Lambda_t) = 0 \;\; .$$

The formal proof is omitted.

4. Linear Composite Functions are Easy

In this section, we consider composite functions that are linear in the oblivious function $\ell_{t-m:t}$. Namely, the adversary chooses a memory size $m \geq 1$ and defines

$$\forall t \quad f_t(x_{1:t}) \;=\; a_m \ell_{t-m}(x_{t-m}) + \cdots + a_0 \ell_t(x_t) \;\;, \tag{19}$$

where $a_0, a_1, \ldots, a_m$ are fixed, bounded, and known coefficients, at least one of which is non-zero (otherwise the regret is trivially zero). In order to ensure that $f_t(x_{1:t}) \in [0,1]$ for all t, we assume that $\sum_{i=0}^m a_i \leq 1$. We can also assume, without loss of generality, that in fact $\sum_{i=0}^m a_i = 1$, since scaling all of the loss functions by a constant scales the regret by the same constant. Recall that, for completeness, we assumed that $\ell_t \equiv 0$ for $t \leq 0$.

Algorithm 1 STRATEGY FOR LINEAR COMPOSITE FUNCTIONS

```
set d = min{i ≥ 0 : a_i ≠ 0}
initialize d + 1 independent instances A_0, ..., A_d of EXP3.
initialize z_0 = z_{-1} = ... = z_{-m+1} = 0
for t = 1 to T do
   set j = t mod (d + 1)
   draw x_t ~ A_j, play x_t and observe feedback f_t(x_{1:t})
   set z_t ← (1/a_d) ( f_t(x_{1:t}) − Σ_{i=d+1}^{m} a_i z_{t−i} )
   feed A_j with feedback z_t (for action x_t)
end for
```

We show that an adversary that chooses a linear composite loss induces an *easy* bandit learning problem. More specifically, we present a strategy, given in Algorithm 1, that achieves $\widetilde{O}(\sqrt{T})$ regret against any loss function sequence of this type. This strategy uses the EXP3 algorithm (Auer et al., 2002) as a black box, and relies on the guarantee that EXP3 attains a regret of $O(\sqrt{Tk \log T})$ against any oblivious loss sequence, with bandit feedback.

Theorem 7 *For any sequence of loss functions $f_{1:T}$ of the form in Eq. (19), the expected regret of Algorithm 1 satisfies $R = O(\sqrt{mTk \log k})$.*

The proof of Theorem 7 is deferred to Appendix B.

5. Conclusion

Cesa-Bianchi et al. (2013); Dekel et al. (2013) were the first to show that a finite-horizon online bandit problem with a finite set of actions can be *hard*. They achieved this by proving that the minimax regret of the multi-armed bandit with switching costs has a rate of $\widetilde{\Theta}(T^{2/3})$. In this paper, we defined the class of online learning problems that define their loss values using a composite loss function, and proved that two non-linear instances of this problem are also hard. Although we reused some technical components from the analysis in Dekel et al. (2013), the composite loss function setting is quite distinct from the multi-armed bandit with switching costs, as it does not explicitly penalize switching. Our result reinforces the idea that the class of hard online learning problems may be a rich class, which contains many different natural settings. To confirm this, we must discover additional online learning settings that are provably hard.

We also proved that linear composite functions induce easy bandit learning problems. Characterizing the set of combining functions that induce hard problems remains an open problem.

References

A. Antos, G. Bartók, D. Pál, and C. Szepesvári. Toward a classification of finite partial-monitoring games. *Theoretical Computer Science*, 2012.

R. Arora, O. Dekel, and A. Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning*, 2012.

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

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, May 1997.

N. Cesa-Bianchi, O. Dekel, and O. Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems 26*, 2013.

O. Dekel, J. Ding, T. Koren, and Y. Peres. Bandits with switching costs: $T^{2/3}$ regret. *arXiv preprint arXiv:1310.2997. (STOC 2014, to appear)*, 2013.

Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and System Sciences*, 55(1):119–139, 1997.

S. Geulen, B. Vöcking, and M. Winkler. Regret minimization for online buffering problems using the weighted majority algorithm. In *Proceedings of the 23rd International Conference on Learning Theory*, pages 132–143, 2010.

A. Gyorgy and G. Neu. Near-optimal rates for limited-delay universal lossy source coding. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2218–2222. IEEE, 2011.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71:291–307, 2005.

N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

A. Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of the 18th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 222–227, 1977.

Appendix A. Technical Proofs

We first prove Lemma 2.

Proof of Lemma 2 We begin with the deterministic parent function used in Sec. 2, denoted here by $\tilde{\rho}$, and defined as $\tilde{\rho}(t) = t - \gcd(t, 2^T)$. Additionally, draw independent unbiased Bernoullis $B_{1:T}$. We now define the random function ρ. If $B_t = 0$ then set $\rho(t) = t - 1$. To define the remaining values of ρ, rename the ordered sequence $(t : B_t = 1)$ as $(U_1, U_2, \ldots)$ and also set $U_0 = 0$. If $B_t = 1$, let k be such that $t = U_k$ and set $\rho(t) = U_{\tilde{\rho}(k)}$. This concludes our construction, and we move on to prove that it satisfies the desired properties.

The probability that $\rho(t) = t - 1$ is at least $\frac{1}{2}$, since this occurs whenever the unbiased bit B_t equals zero. Dekel et al. (2013) proves that the width of $\tilde{\rho}$ is bounded by $\log T + 1$, and the width of ρ never exceeds this bound. Dekel et al. (2013) also proves that the depth of $\tilde{\rho}$ is bounded by $\log T + 1$. The depth difference $d(\rho) - d(\tilde{\rho})$ is at most $\max_k (U_k - U_{k-1})$ by construction. A union bound implies that the probability that this maximum exceeds $\ell = 2 \log T$ is at most $T \cdot 2^{-\ell} = T^{-1}$. Thus $\mathbb{P}\big(d(\rho) \geq 4 \log T\big) \leq T^{-1}$. ■

We next provide the proofs of the technical lemmas of Sec. 3. We begin with Lemma 4.

Proof of Lemma 4 Note that the occurrence of E_{t-1} implies that $W_t \leq W_{t-1} - \tau$ and that $W_{t+1} \leq W_t - \tau$, which means that a spike is not added on round t. Therefore, $L_t(x_t) = Z_t(x_t)$ for any x_t and $\min\big(L_{t-1}(x_{t-1}), L_t(x_t)\big) = Z_t(x_t)$ for any x_{t-1} and x_t. The first claim follows from two applications of this observation: once with $x_{t-1} = x_t = X_{t-1}$ and once with $x_{t-1} = x_t = \chi$. The second claim is obtained by setting $x_{t-1} = X_{t-1}, x_t = X_t$. ■

Proof of Lemma 5 If $X_{t-1} = \chi$ then $\tilde{R}_t = 0$ trivially. Assume henceforth that $X_{t-1} \neq \chi$. If E_{t-1} occurs then Lemma 4 guarantees that $\tilde{R}_t = Z_t(X_{t-1}) - Z_t(\chi)$, which equals ϵ by the definition of Z_t. If $\neg E_{t-1}$ and $\neg E_t$ then

$$\tilde{R}_t \;=\; \min\big(Z_{t-1}(X_{t-1}), Z_t(X_{t-1})\big) - \min\big(Z_{t-1}(\chi), Z_t(\chi)\big) \;,$$

which, again, equals ϵ. If E_t occurs then the loss depends on whether $W_{t-1} \geq W_t$ and on the value of Λ_t. We can first focus on the case where $W_{t-1} \geq W_t$. If $\Lambda_t \neq X_{t-1}$ then the assumption that $\eta \gg \tau$ implies that $\min(L_{t-1}(\chi), L_t(\chi)) = Z_{t-1}(\chi)$ and $\min(L_{t-1}(X_{t-1}), L_t(X_{t-1})) = Z_t(X_{t-1})$, and therefore

$$\tilde{R}_t \;=\; Z_t(X_{t-1}) - Z_{t-1}(\chi) \;=\; \epsilon - |W_{t-1} - W_t| \;, \tag{20}$$

which could be negative. On the other hand, if $\Lambda_t = X_{t-1}$, then $\min(L_{t-1}(\chi), L_t(\chi)) = Z_t(\chi)$ and $\min(L_{t-1}(X_{t-1}), L_t(X_{t-1})) = Z_{t-1}(X_{t-1})$, and therefore

$$\tilde{R}_t \;=\; Z_{t-1}(X_{t-1}) - Z_t(\chi) \;=\; \epsilon + |W_{t-1} - W_t| \;. \tag{21}$$

Now note that Λ_t is an unbiased Bernoulli that is independent of X_{t-1} (this argument would have failed had we directly analyzed R_t instead of $\tilde{R}_t$). Therefore, the possibility of having a negative regret in Eq. (20) is offset by the equally probable possibility of a positive regret in Eq. (21). In other words,

$$\mathbb{E}\big[\tilde{R}_t \,\big|\, X_{t-1} \neq \chi, W_{t-1} \geq W_t, E_t\big] \;=\; \frac{1}{2}\big(\epsilon - |W_{t-1} - W_t|\big) + \frac{1}{2}\big(\epsilon + |W_{t-1} - W_t|\big) \;=\; \epsilon \;.$$

The same calculation applies when $W_{t-1} < W_t$. Overall, we have shown that $\mathbb{E}[\tilde{R}_t \mid X_{t-1} \neq \chi] = \epsilon$. ■

Proof of Lemma 6 Since R_t and $\tilde{R}_t$ only differ when $X_t \neq X_{t-1}$, we have that

$$\mathbb{E}[R_t - \tilde{R}_t] \;=\; \mathbb{P}(X_t \neq X_{t-1})\, \mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}] \;\;,$$

so it remains to prove that $\mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}] = \Omega(\eta\tau/\sigma)$. We deal with two cases, depending on the occurrence of E_t, and write

$$\begin{aligned}\mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}] \;&=\; \mathbb{P}(\neg E_t \mid X_t \neq X_{t-1})\, \mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}, \neg E_t] \\ &\quad+\; \mathbb{P}(E_t \mid X_t \neq X_{t-1})\, \mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}, E_t] \;\;. \end{aligned} \tag{22}$$

We begin by lower-bounding the first case, where $\neg E_t$. If E_{t-1} occurs, then Lemma 4 guarantees that $R_t - \tilde{R}_t = Z_t(X_t) - Z_t(X_{t-1})$, which is at least $-\epsilon$. Otherwise, if neither E_{t-1} or E_t occur, then again $R_t - \tilde{R}_t \geq -\epsilon$. We upper-bound $\mathbb{P}(\neg E_t \mid X_t \neq X_{t-1}) \leq 1$ and get that

$$\mathbb{P}(\neg E_t \mid X_t \neq X_{t-1})\, \mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}, \neg E_t] \;\geq\; -\epsilon \;\;. \tag{23}$$

Next, we lower-bound the second case, where E_t. Lemma 8 below lower-bounds $\mathbb{P}(E_t \mid X_t \neq X_{t-1}) = \Omega(\tau/\sigma)$. Lemma 9 below lower-bounds $\mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}, E_t] \geq \eta/3 - \tau$ for T sufficiently large. Recalling the assumption that $\eta \gg \tau$, we conclude that

$$\mathbb{P}(E_t \mid X_t \neq X_{t-1})\, \mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}, E_t] \;=\; \Omega\left(\frac{\eta\tau}{\sigma}\right) \;\;.$$

Eq. (23) can be neglected since $\eta\tau/\sigma \gg \epsilon$, and this concludes the proof. ■

Lemma 8 *Suppose $\eta, \tau \leq \sigma/\log T$. For all $t > 1$ it holds that $\mathbb{P}(E_t \mid X_t \neq X_{t-1}) \;=\; \Omega(\tau/\sigma)$.*

Proof By our earlier modification of the algorithm, we assume that

$$|L_s(X_s) - L_{\rho(s)}(X_{\rho(s)})| \;\leq\; 4\sigma\sqrt{\log T} \text{ for } s \in \{t-2, t-1\} \tag{24}$$

(which occurs with probability at least $1 - O(T^{-4})$). Otherwise, the event $X_t \neq X_{t-1}$ would never occur due to our modification of the algorithm and the statement is irrelevant.

In order to prove the lemma, we verify a stronger statement that $\mathbb{P}(E_t \mid \mathcal{F}_{t-1}) \;=\; \Omega(\tau/\sigma)$, where $\mathcal{F}_{t-1}$ is the σ-field generated by the player's observations up to round $t-1$ (note that X_t is $\mathcal{F}_{t-1}$-measurable). Let $f_1(\ell_1, \ldots, \ell_{t-1})$ be the conditional density of $(L_1(X_1), \ldots, L_{t-1}(X_{t-1}))$ given E_{t-1}, and let $f_2(\ell_1, \ldots, \ell_{t-1})$ be the conditional density of $(L_1(X_1), \ldots, L_{t-1}(X_{t-1}))$ given E^c_{t-1}. We get that

$$\min \frac{f_2(\ell_1, \ldots, \ell_{t-1})\mathbb{P}(E^c_{t-1})}{f_1(\ell_1, \ldots, \ell_{t-1})\mathbb{P}(E_{t-1})} \;\geq\; \min_{x \leq 4\sigma\sqrt{\log T}} \left(\frac{\mathrm{e}^{x^2/\sigma^2}}{\mathrm{e}^{(x+\eta)^2/\sigma^2}}\right)^2 \;=\; 1 - o(1)\,,$$

where the the first minimum is over all sequences that are compatible with Eq. (24), and the last inequality follows from the assumption that $\eta \leq \sigma/\log T$. Hence, we have

$$\mathbb{P}(E^c_{t-1} \mid \mathcal{F}_{t-1}) \;\geq\; 1/2 + o(1)\,. \tag{25}$$

Further, we see that

$$\begin{aligned}
&\frac{\mathbb{P}(E_{t-1}^c, \rho(t) = t-1, \rho(t+1) = t, \xi_t \geq -\tau \mid \mathcal{F}_{t-1})}{\mathbb{P}(E_{t-1}^c \mid \mathcal{F}_{t-1})} \\
&\geq \mathbb{P}(\rho(t) = t-1, \rho(t+1) = t \mid \rho(1), \ldots, \rho(t-1)) \cdot \mathbb{P}(\xi_t \geq -\tau) \\
&\geq \frac{1-o(1)}{8} \,. \qquad (26)
\end{aligned}$$

Conditioning on the event $E_{t-1}^c \cap \{\rho(t) = t-1, \rho(t+1) = t, \xi_t \geq -\tau\}$, we note that E_t is independent of $\mathcal{F}_{t-1}$. Combined with Eq. (25) and Eq. (26), it follows that

$$\begin{aligned}
\mathbb{P}(E_t \mid \mathcal{F}_{t-1}) &\geq \frac{1-o(1)}{8} \cdot \mathbb{P}(E_t \mid E_{t-1}^c, \rho(t) = t-1, \rho(t+1) = t, \xi_t \geq -\tau) \\
&\geq \frac{1-o(1)}{8} \cdot \mathbb{P}(|\xi_t| \leq \tau, \xi_{t+1} < -\tau \mid \xi_t \geq -\tau) \\
&= \Omega(\tau/\sigma) \,,
\end{aligned}$$

where the last inequality follows from the assumption that $\tau < \sigma / \log T$. ∎

Lemma 9 *For all $t > 1$ it holds that $\mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}, E_t] \geq \eta/3 - \tau$.*

Proof We rewrite $\mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}, E_t]$ as

$$\begin{aligned}
&\mathbb{P}(\Lambda_t = X_t \mid X_t \neq X_{t-1}, E_t)\, \mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}, E_t, \Lambda_t = X_t] \\
&+ \mathbb{P}(\Lambda_t \neq X_t \mid X_t \neq X_{t-1}, E_t)\, \mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}, E_t, \Lambda_t \neq X_t] \,.
\end{aligned}$$

First consider the case where $\Lambda_t = X_t$, namely, the orientation of the spikes coincides with the direction of the player's switch. In this case,

$$\mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}, E_t, \Lambda_t = X_t] \geq \eta - \tau \,.$$

If $\Lambda_t = X_t$ then the orientation of the spikes does not coincide with the switch direction and

$$\mathbb{E}[R_t - \tilde{R}_t \mid X_t \neq X_{t-1}, E_t, \Lambda_t \neq X_t] \geq -\tau \,.$$

Lemma 10 below implies that

$$\mathbb{P}(\Lambda_t = X_t \mid X_t \neq X_{t-1}, E_t) \geq \frac{1}{3} \,,$$

which concludes the proof. ∎

Lemma 10 *Suppose that $\eta \leq \sigma / \log T$. For a sufficiently large T it holds that*

$$\frac{\mathbb{P}(\Lambda_t = X_t \mid X_t \neq X_{t-1}, E_t)}{\mathbb{P}(\Lambda_t \neq X_t \mid X_t \neq X_{t-1}, E_t)} \geq \frac{1}{2} \,.$$

Proof The ratio on the left can be rewritten, using Bayes' rule, as

$$\frac{\mathbb{P}(X_t \neq X_{t-1} \mid \Lambda_t = X_t, E_t)}{\mathbb{P}(X_t \neq X_{t-1} \mid \Lambda_t \neq X_t, E_t)} \ .$$

To see this is at least $\frac{1}{2}$, condition on the history until time $t-2$ and note that by our earlier modification of the algorithm, we may assume that $|L_{t-1}(X_{t-1}) - L_{\rho(t-1)}(X_{\rho(t-1)})| \leq 4\sigma\sqrt{\log T}$. We let $f_1(x)$ be the conditional density of $L_{t-1}(X_{t-1}) - L_{\rho(t-1)}(X_{\rho(t-1)})$ given $\{X_t = \Lambda_t\} \cap E_t$, and let $f_2(x)$ be the conditional density of $L_{t-1}(X_{t-1}) - L_{\rho(t-1)}(X_{\rho(t-1)})$ given $\{\Lambda_t \neq X_t\} \cap E_t$. Therefore, we see that f_1 is the density function for $\sigma Z + \eta$, and f_2 is the density function for σZ where $Z \sim N(0,1)$. Thus, we have

$$\min_{|x| \leq 4\sigma\sqrt{\log T}} \frac{f_1(x)}{f_2(x)} \ = \ \min_{|x| \leq 4\sigma\sqrt{\log T}} \frac{\mathrm{e}^{-(x-\eta)^2/2\sigma^2}}{\mathrm{e}^{-x^2/2\sigma^2}} \ = \ 1 - o(1) \ , \tag{27}$$

where the last inequality follows from the assumption that $\eta \leq \sigma/\log T$. Now consider two scenarios of the game where the observations are identical up to time $t-2$, and then for the two scenarios we condition on events $\{\Lambda_t = X_t\} \cap E_t$ and $\{\Lambda_t \neq X_t\} \cap E_t$ respectively. Then by Eq. (27) the observation at time $t-1$ is statistically close, and therefore the algorithm will make a decision for X_t that is statistically close in these two scenarios. Formally, we get that

$$\frac{\mathbb{P}(X_t \neq X_{t-1} \mid \Lambda_t = X_t, E_t)}{\mathbb{P}(X_t \neq X_{t-1} \mid \Lambda_t \neq X_t, E_t)} \ \geq \ \min_{|x| \leq 4\sigma\sqrt{\log T}} \frac{f_1(x)}{f_2(x)} \ = \ 1 - o(1) \ ,$$

completing the proof of the lemma. ■

Appendix B. Analysis of Linear Composite Functions

Here we give the proof of Theorem 7.

Proof of Theorem 7 First, observe that $z_t = \ell_{t-d}(x_{t-d})$ for all $t \in [T]$. Indeed, for $t = 1$ this follows directly from the definition of $f_{1:m}$ (and from the fact that $z_t = 0$ for $t \leq 0$), and for $t > 1$ an inductive argument shows that

$$z_t \ = \ \frac{1}{a_d}\left(f_t(x_{1:t}) - \sum_{i=d+1}^{m} a_i z_{t-i}\right) \ = \ \frac{1}{a_d}\left(f_t(x_{1:t}) - \sum_{i=d+1}^{m} a_i \ell_{t-i}(x_{t-i})\right) \ = \ \ell_{t-d}(x_{t-d}) \ .$$

Hence, each algorithm $\mathcal{A}_j$ actually plays a standard bandit game with the subsampled sequence of oblivious loss functions $\ell_j, \ell_{j+(d+1)}, \ell_{j+2(d+1)}, \ldots$. Consequently, for each $j = 0, 1, \ldots, d$ we have

$$\forall\, x \in [k] \ , \quad \mathbb{E}\left[\sum_{t \in S_j} \ell_t(x_t)\right] - \sum_{t \in S_j} \ell_t(x) \ = \ O\left(\sqrt{\frac{Tk\log k}{d}}\right) \ , \tag{28}$$

where $S_j = \{t \in [T] \,:\, t = j \mod (d+1)\}$, and we used the fact that $|S_j| = \Theta(T/d)$. Since the sets $S_0, \ldots, S_d$ are disjoint and their union equals $[T]$, by summing Eq. (28) over $j = 0, 1, \ldots, d$ we obtain

$$\forall\, x \in [k]\,, \quad \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(x_t)\right] - \sum_{t=1}^{T} \ell_t(x) \;=\; O(\sqrt{dTk \log k}) \;=\; O(\sqrt{mTk \log k}) \;. \tag{29}$$

However, notice that the loss of the player satisfies

$$\begin{aligned}\sum_{t=1}^{T} f_t(x_{1:t}) \;&=\; \sum_{t=1}^{T} \big(a_m \ell_{t-m}(x_{t-m}) + \cdots + a_0 \ell_t(x_t)\big) \\ &\le\; a_m \sum_{t=1}^{T} \ell_t(x_t) + \cdots + a_0 \sum_{t=1}^{T} \ell_t(x_t) \\ &=\; \sum_{t=1}^{T} \ell_t(x_t)\;,\end{aligned}$$

where the last equality uses the assumption that $\sum_{i=0}^{m} a_i = 1$. A similar calculation shows that for any fixed $x \in [k]$,

$$\begin{aligned}\sum_{t=1}^{T} f_t(x, \ldots, x) \;&=\; \sum_{t=1}^{T} \big(a_m \ell_{t-m}(x) + \cdots + a_0 \ell_t(x)\big) \\ &\ge\; a_m \left(\sum_{t=1}^{T} \ell_t(x) - m\right) + \cdots + a_0 \left(\sum_{t=1}^{T} \ell_t(x) - m\right) \\ &=\; \sum_{t=1}^{T} \ell_t(x) - m\;.\end{aligned}$$

Putting things together, we obtain that for all $x \in [k]$,

$$\sum_{t=1}^{T} f_t(x_{1:t}) - \sum_{t=1}^{T} f_t(x, \ldots, x) \;\le\; \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(x) + m\;.$$

Finally, taking the expectation of this inequality and combining with Eq. (29) completes the proof. ■