Proceedings of the 2014 ACM Symposium on Theory of Computing

Bandits with Switching Costs: $T^{2/3}$ Regret

Ofer Dekel
Microsoft Research
oferd@microsoft.com

Jian Ding[*]
University of Chicago
jianding@galton.uchicago.edu

Tomer Koren[*]
Technion
tomerk@technion.ac.il

Yuval Peres
Microsoft Research
peres@microsoft.com

ABSTRACT

We study the adversarial multi-armed bandit problem in a setting where the player incurs a unit cost each time he switches actions. We prove that the player's T-round minimax regret in this setting is $\widetilde{\Theta}(T^{2/3})$, thereby closing a fundamental gap in our understanding of learning with bandit feedback. In the corresponding full-information version of the problem, the minimax regret is known to grow at a much slower rate of $\Theta(\sqrt{T})$. The difference between these two rates provides the *first* indication that learning with bandit feedback can be significantly harder than learning with full-information feedback (previous results only showed a different dependence on the number of actions, but not on T.)

In addition to characterizing the inherent difficulty of the multi-armed bandit problem with switching costs, our results also resolve several other open problems in online learning. One direct implication is that learning with bandit feedback against bounded-memory adaptive adversaries has a minimax regret of $\widetilde{\Theta}(T^{2/3})$. Another implication is that the minimax regret of online learning in adversarial Markov decision processes (MDPs) is $\widetilde{\Theta}(T^{2/3})$. The key to all of our results is a new randomized construction of a multi-scale random walk, which is of independent interest and likely to prove useful in additional settings.

Categories and Subject Descriptors

F.2 [**Theory of Computation**]: Analysis of Algorithms and Problem Complexity

General Terms

Algorithms, Theory

[*] Parts of this work were done while the author was at Microsoft Research, Redmond.

STOC '14, May 31 – June 03 2014, New York, NY, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2710-7/14/05 ...$15.00.
http://dx.doi.org/10.1145/2591796.2591868

Keywords

Online learning, Multi-armed Bandit, Switching costs, Lower bounds

1. INTRODUCTION

Online learning with a finite set of actions is a fundamental problem in machine learning, with two important special cases: the *Adversarial (Non-Stochastic) Multi-Armed Bandit* [3] and *Predicting with Expert Advice* [5, 10]. This problem is often presented as a T-round repeated game between a player and an adversary: on each round of the game, the player chooses an *action*[1] from the set $[k] = \{1, \dots, k\}$ and incurs a loss in $[0, 1]$ for that action. The player is allowed to randomize, i.e., on each round he selects a distribution over actions and draws an action from that distribution. The loss corresponding to each action on each round is set in advance by the adversary, and in particular, the loss of each action can vary from round to round. The player's goal is to minimize the total loss accumulated over the course of the game.

The bandit problem and the experts problem differ in the feedback received by the player after each round. In the bandit problem, the player only observes his loss (a single number) on each round; this is called *bandit feedback*. In the experts problem, the player observes the loss assigned to each possible action (for a total of k real numbers in each round); this is called *full feedback* or *full information*. A player that receives bandit feedback must balance an exploration/exploitation trade-off, while a player that receives full feedback is only concerned with exploitation.

For example, say that we manage an investment portfolio, we receive daily advice from k financial experts, and on each day we must follow the advice of one expert. The loss associated with each expert on each day reflects the amount of money we would lose by following that expert's advice on that day. If we know the advice given by each expert, the problem is said to provide full feedback. Alternatively, if we purchase advice from a single expert on each day, and the advice of the other $k - 1$ experts remains unknown, the problem is said to provide bandit feedback.

In the problem just described, the player is allowed to switch freely between actions. An equally interesting setting is one where each switch incurs a *switching cost*: In addition to the losses chosen by the adversary, the player

[1] In the bandit problem, each action is called an *arm*; in the experts problem, each action is called an *expert*.

pays a penalty each time his action differs from the one he played on the previous round. In the motivating example described above, switching our primary financial consultant may require terminating a contract with the previous expert and negotiating contract with the new one, or it may just cost us the fees and commissions that result from a significant change in investment strategy. Switching costs arise naturally in a variety of other applications: In online web applications, switching the content of a website too frequently can be annoying to users; in industrial applications, switching actions might entail reconfiguring a production line. Moreover, Geulen et al. [11] reduced a family of online buffering problems to switching cost problems; similarly, Gyorgy and Neu [12] used the switching cost setting to solve the limited-delay universal lossy source coding problem.

We focus on analyzing the inherent difficulty of online learning with switching costs, using the game-theoretic notion of *minimax regret.* To define this notion, we must first specify the setting formally. Before the game begins, the adversary chooses a loss functions $\ell_1, \ldots, \ell_T$, where each ℓ_t maps the action set $[k]$ to $[0, 1]$. Since the entire sequence is chosen in advance, we say that the adversary is *oblivious* (to the player's actions). On round t, the player selects a distribution over the set of actions and draws an action X_t from that distribution. The player then incurs the loss $\ell_t(X_t) + \mathbb{1}_{X_t \neq X_{t-1}}$, which includes the adversarially chosen loss $\ell_t(X_t)$ and the switching cost. To make the loss on the first round well-defined, we set $X_0 = 0$ (so the first action always counts as a switch). The player's cumulative loss at the end of the game equals $\sum_{t=1}^{T} \big(\ell_t(X_t) + \mathbb{1}_{X_t \neq X_{t-1}}\big)$.

Since the loss functions are adversarial, the cumulative loss is only meaningful when compared to an adequate baseline. Therefore, we compare the player's cumulative loss to the loss of the best fixed policy (in hindsight), which is a policy that chooses the same action on all T rounds. Formally, we define the player's *regret* at the end of the game as

$$R \;=\; \sum_{t=1}^{T} \big(\ell_t(X_t) + \mathbb{1}_{X_t \neq X_{t-1}}\big) \;-\; \min_{x \in [k]} \sum_{t=1}^{T} \ell_t(x) \quad . \qquad (1)$$

While regret measures the player's performance on a given instance of the game, the inherent difficulty of the game itself is measured by *minimax expected regret* (or just *minimax regret* for brevity). Intuitively, minimax regret is the expected regret when both the adversary and the player behave optimally. Formally, minimax regret is the minimum over all randomized player strategies, of the maximum over all loss sequences, of $\mathbb{E}[R]$. In this paper, our primary focus is to determine the asymptotic growth rate of the minimax regret as a function of the number of rounds T and the number of actions k.

Minimax regret rates are already well understood in several of the settings discussed above. Without switching costs, the minimax regret of the adversarial multi-armed bandit problem is $\Theta(\sqrt{Tk})$ (see Auer et al. [3], Audibert et al. [2]) and the minimax regret of the experts problem is $\Theta(\sqrt{T \log k})$ (see Littlestone and Warmuth [14], Freund and Schapire [10], Cesa-Bianchi and Lugosi [4]). This implies that when no switching costs are added, the bandit problem is not substantially more difficult than the experts problem (at least when the number of actions is constant), despite the added burden of exploration.

When switching costs are added, the previous literature does not provide a full characterization of minimax regret. Clearly, the lower bound without switching costs still apply with switching costs are added. In the full feedback setting with switching costs, the *Follow the Lazy Leader* algorithm [13] and the *Shrinking Dartboard* algorithm [11] both guarantee a matching upper bound of $O(\sqrt{T \log k})$, so the minimax regret is $\Theta(\sqrt{T \log k})$. However, the minimax regret of the bandit problem with switching costs was not well understood. Arora et al. [1] presented a simple algorithm with a guaranteed regret of $O(k^{1/3}T^{2/3})$, but a matching lower bound was not known.

Recently, Cesa-Bianchi et al. [6] addressed this gap, but fell short of resolving it. Specifically, they modified the game by allowing the loss per round to drift out of the interval $[0, 1]$ and to possibly grow in magnitude to be as large as $\Theta(\sqrt{T})$. In this setting, they proved that the minimax regret (with a constant number of actions k) grows at a rate of $\widetilde{\Theta}(T^{2/3})$. However, allowing unbounded loss per round is quite uncommon and not very natural. Also, it isn't clear what implications their results have on the original problem (i.e., with bounded losses), and whether their $\widetilde{\Theta}(T^{2/3})$ rate is merely an artifact of the unbounded loss functions they allow.

1.1 Our Results

Our main result is a new $\widetilde{\Omega}(T^{2/3})$ lower bound on the regret of the multi-armed bandit problem with switching costs (in the standard setting, with losses bounded in $[0, 1]$).

THEOREM 1. *For any randomized player strategy that relies on bandit feedback, there exists a sequence of loss functions $\ell_1, \ldots, \ell_T$ (where $\ell_t : [k] \mapsto [0, 1]$) that incurs a regret of $R = \widetilde{\Omega}(k^{1/3}T^{2/3})$, provided that $k \le T$.*

When combined with the upper bound in Arora et al. [1], our result implies that the minimax regret of the multi-armed bandit problem with switching costs is $\widetilde{\Theta}(k^{1/3}T^{2/3})$. A direct consequence of this result is that the bandit problem with switching costs is substantially more difficult than the corresponding experts problem. To the best of our knowledge, this is the first example that exhibits (even for constant k) a clear gap between the asymptotic difficulty of online learning with bandit and full feedback, as T grows.

To prove Theorem 1, we apply (the easy direction of) Yao's minimax principle [16], which states that the regret of a randomized player against the worst-case loss sequence is at least the minimax regret of the optimal *deterministic* player against a *stochastic* loss sequence. In other words, as an intermediate step toward proving Theorem 1, we construct a sequence of stochastic loss functions, each from $[k]$ to $[0, 1]$, and prove that the expected regret of any deterministic player (where expectation is now taken with respect to the randomness of the loss sequence) is $\widetilde{\Omega}(k^{1/3}T^{2/3})$.

We then generalize our lower bound to the case where each switch incurs a cost of $c > 0$, where c does not necessarily equal 1 (e.g., set $c = T^q$, for some $q \in [-1, 1]$). A corollary of this result is that any algorithm for the multi-armed bandit problem (without switching costs) that guarantees a regret of $O(\sqrt{T})$ (e.g., the EXP3 algorithm [3]) can be forced to make $\widetilde{\Omega}(T)$ switches. Finally, we observe that our problem is a special case of an online Markov decision process (MDP) learning problem with adversarial rewards and bandit feed-

```
Input: time horizon $T > 0$, number of actions $k \geq 2$
1: Set $\epsilon = k^{1/3}T^{-1/3}/(9\log_2 T)$ and $\sigma = 1/(9\log_2 T)$.
2: Choose $\chi \in [k]$ uniformly at random.
3: Draw $T$ independent zero-mean $\sigma^2$-variance Gaussians $\xi_{1:T}$.
4: Define $W_{0:T}$ recursively by
       $W_0 = 0$ ,
       $\forall\, t \in [T] \quad W_t = W_{\rho(t)} + \xi_t$
   with
       $\rho(t) = t - 2^{\delta(t)}$ ,
   where $\delta(t) = \max\{i \geq 0 : 2^i \text{ divides } t\}$ .
5: For all $t \in [T]$ and $x \in [k]$, set
       $L_t(x) = W_t + \frac{1}{2} - \epsilon \cdot \mathbb{1}_{\chi = x}$ .
Output: loss functions $L_{1:T}$.
```

Figure 1: The adversary's randomized algorithm for generating a loss sequence $L_{1:T}$, which is bounded in $[0, 1]$ with constant positive probability, and ensures an expected regret of $\widetilde{\Omega}(k^{1/3}T^{2/3})$ against any deterministic player.

back, and therefore the minimax regret of that problem is also $\widetilde{\Omega}(T^{2/3})$.

The paper is organized as follows: in Sec. 2 we describe the general construction of the stochastic loss sequence and in Sec. 3 we present the stochastic process that underlies our construction. We then prove our lower bound on regret in Sec. 4 and present extensions and implications in Sec. 5.

2. CONSTRUCTING THE LOSS SEQUENCE

In this section we construct a sequence[2] of stochastic loss functions, $L_{1:T}$, where each L_t randomly maps the action set $[k]$ to the real line. Our construction allows the loss functions to take values outside of the interval $[0, 1]$, but with constant positive probability all of their values are bounded in $[0, 1]$. In other words, we say that an instantiation of this sequence is *admissible* if the range of each function in the sequence is contained in the interval $[0, 1]$, and we show that $L_{1:T}$ is admissible with constant positive probability. We also prove that the expected regret of any deterministic player, when the loss sequence is admissible, is $\widetilde{\Omega}(k^{1/3}T^{2/3})$. The adversary's randomized algorithm for generating the sequence $L_{1:T}$ is given in Fig. 1. The key to this algorithm is the stochastic process $W_{1:T}$, defined on lines 3–4 of Fig. 1. The adversary draws a concrete sequence from this process and uses it to define the loss values of all k actions. First, the adversary picks an action $\chi \in [k]$ uniformly at random, to serve as the best action (whose loss is always smaller than the loss of the other actions). The loss of all actions $x \neq \chi$ is simply set to $L_t(x) = W_t + \frac{1}{2}$. The loss of the best action χ is set to $L_t(\chi) = W_t + \frac{1}{2} - \epsilon$, where ϵ is a predefined gap parameter, and is therefore consistently better than the losses of the other actions.

When faced with the loss sequence $L_{1:T}$, the player attempts to identify which of the k actions has the smaller loss (or equivalently, to reveal the value of χ). Although the loss values of the best action are deterministically separated from those of the other actions by a constant gap, the player only observes one loss value on each round and never knows if his chosen action incurred the higher loss or the lower loss. Our analysis shows that the player's ability to uncover information about the identity of the best action depends on the characteristics of the stochastic process $W_{1:T}$. For example, if this process were an i.i.d. sequence, it is easy to see that the player could identify the best action by estimating the expected loss of each action to within $\epsilon/2$ (for example, using Hoeffding's bound), requiring only $O(\sigma^2/\epsilon^2)$ samples of each action and at most $k-1$ switches between actions. This example already implies that the dependency structure in our construction of $W_{1:T}$ plays a central role. We show that a careful choice of the stochastic process $W_{1:T}$ ensures that the amount of information uncovered by the player during the game is tightly controlled by the number of switches he performs. Therefore, to detect the best action, the player must switch actions frequently and pay the associated switching costs.

3. THE STOCHASTIC PROCESS

The key to our analysis is a careful choice of the stochastic process $W_{1:T}$ that underlies the definition of $L_{1:T}$. In this section we describe a stochastic processes with a controllable dependence structure, which includes i.i.d. Gaussian sequences and simple Gaussian random walks as special cases.

Let $\xi_{1:T}$ be a sequence of independent zero-mean Gaussian random variables with variance σ^2. Let $\rho : [T] \mapsto \{0\} \cup [T]$ be a function that assigns each $t \in [T]$ with a *parent* $\rho(t)$. We allow ρ to be any function that satisfies $\rho(t) < t$ for all t. Now define

$$W_0 = 0 \ ,$$
$$\forall\, t \in [T] \quad W_t = W_{\rho(t)} + \xi_t \ .$$

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 simple Gaussian random walk. Other definitions of ρ can create interesting dependencies between the variables of the stochastic process.

3.1 Depth and Width

We highlight two properties of the parent function ρ (and consequently, of the induced stochastic process) that are essential to our analysis.

Definition 1 (*ancestors, depth*). *Given a parent function ρ, the set of ancestors of t is denoted by $\rho^*(t)$ and defined as the set of positive indices that are encountered when ρ is applied recursively to t. Formally, $\rho^*(t)$ is defined recursively as*

$$\rho^*(0) = \{\}$$
$$\forall\, t \in [T] \quad \rho^*(t) = \rho^*\big(\rho(t)\big) \cup \{\rho(t)\} \ . \qquad (2)$$

The depth of ρ is then defined as $d(\rho) = \max_{t \in [T]} |\rho^*(t)|$.

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

Using this definition, we can write $W_t = \xi_t + \sum_{s \in \rho^*(t)} \xi_s$, where $\xi_0 = 0$. Thus, if $d(\rho) = d$, the induced stochastic process includes sums of at most d independent Gaussians, each with variance σ^2. This implies the following bound.

Lemma 1. *Let $W_{1:T}$ be the stochastic process defined by the parent function ρ. Then*

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

Proof. For any $t \in [T]$, W_t is normally distributed with zero mean and variance bounded by $d(\rho)\sigma^2$. Since a standard Gaussian variable Z satisfies $\mathbb{P}(|Z| \ge z) \le \exp(-\frac{1}{2}z^2)$ for any $z \ge 0$, we infer that

$$\mathbb{P}\left(|W_t| \ge \sigma\sqrt{2d(\rho)\log\tfrac{T}{\delta}} \right) \;\le\; \exp\left(-\log\tfrac{T}{\delta}\right) \;=\; \frac{\delta}{T} \ .$$

The above holds for each $t \in [T]$ and the lemma follows from the union bound. □

Lemma 1 implies that the depth of ρ and the variance σ^2 determine how far the process $W_{1:T}$ will drift. Since we require a process that is bounded with high probability, we need to minimize the depth of ρ. (We could counter the effect of a deep ρ by setting σ to be small, but if we do so, the resulting process would not be able to mask the ϵ gap between the losses of the different actions.) This consideration rules out the simple Gaussian random walk, whose depth is T.

Definition 2 (*cut, width*). *Given a parent function ρ, define*

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

the set of rounds that are separated from their parent by t. The width of ρ is then defined[3] *as $w(\rho) = \max_{t \in [T]} |\mathrm{cut}(t)|$.*

Note that the cut size for any $s \in [T]$ is an integer between 1 and T. One extreme is the simple Gaussian random walk ($\rho(t) = t - 1$), whose cuts are of size 1. The other extreme is the sequence of i.i.d. Gaussians ($\rho(t) = 0$), for which $|\mathrm{cut}(s)| = s$, and therefore $w(\rho) = T$.

Our analysis in Sec. 4.1 shows that any information that the player uncovers about the identity of the best action can be attributed to a switch performed on the current round or on a past round (where the first round is always considered to be a switch). Moreover, we prove that the amount of information that can be extracted from a switch at time t is controlled by the size of $\mathrm{cut}(t)$. Therefore, a process with a small width forces the player to perform many switches. This rules out the sequence of i.i.d. Gaussians, as it is too wide and reveals too much information to a player that selects the same action repeatedly.

3.2 The Multi-scale Random Walk

To prove our lower bound, we require a stochastic process that is neither too deep nor too wide. We present such a process, called the *Multi-scale Random Walk* (MRW), whose depth and width are both *logarithmic* in T. The MRW process is formed by the parent function given by

$$\rho(t) \;=\; t - 2^{\delta(t)} \ , \tag{3}$$

where

$$\delta(t) \;=\; \max\left\{ i \ge 0 : 2^i \text{ divides } t \right\} \ .$$

Put another way, $\rho(t)$ is obtained by taking the binary representation of t, identifying the lowest order 1, and flipping it to 0. For example if $t = 10110\mathbf{1}00$ (which equals the decimal number 180) then $\rho(t) = 10110\mathbf{0}00$ (which equals the decimal number 176).

Fig. 2 depicts the MRW process for $T = 7$. Notice that the process takes steps on *multiple scales*, each of which corresponds to a different power of two. An alternative description of the same process can be obtained by considering a binary tree with leaves corresponding to the random variables $W_{1:T}$, as depicted in Fig. 2. In this description, we associate the *right* edges of the tree, enumerated in a DFS traversal order, with the Gaussian variables $\xi_{1:T}$. Then, each W_t is defined as the sum of the ξ_j's encountered along the path from the root to the leaf corresponding to W_t.

We conclude the section with the following lemma, which summarizes the properties of the MRW process used in our analysis.

Lemma 2. *The depth and width of the MRW are both upper-bounded by $\lfloor \log_2 T \rfloor + 1$.*

Proof. Let $n = \lfloor \log_2 T \rfloor + 1$ and note that any integer $t \in [T]$ can be written using n bits. We shall prove that, for all $t \in [T]$, the number $|\rho^*(t)|$ is bounded by (in fact, is equal to) the number of 1's in the n-digit binary representation of t, while $|\mathrm{cut}(t)|$ is bounded by the number of 0's in that representation plus one. This would immediately imply the lemma, as $|\rho^*(t)|$ and $|\mathrm{cut}(t)|$ are both positive and their sum is at most $n + 1$.

First, observe that the number of 1's in the representation of the parent $\rho(t)$ is one less than the number of 1's in the representation of t, and $|\rho^*(0)| = 0$. Hence, $|\rho^*(t)|$ equals the number of 1's in the binary representation of t.

Moving on to the width, choose any $t \in [T]$ and consider the cut it defines. We show that each $s \in \mathrm{cut}(t) \setminus \{t\}$ corresponds to a distinct zero in the n-bit binary representation of t. Let $s \in \mathrm{cut}(t) \setminus \{t\}$ and denote $j = \delta(s)$. Note that $\rho(s) = s - 2^j$ is a multiple of 2^{j+1}, so we can write $s - 2^j = a \cdot 2^{j+1}$ for some integer a. By the definition of the cut and since $s \ne t$, we have $a \cdot 2^{j+1} < t < a \cdot 2^{j+1} + 2^j$. Consequently, $s = 2^{j+1} \cdot \lfloor t/2^{j+1} \rfloor + 2^j$ and the coefficient of 2^j in the binary representation of t is zero. Together with the fact that $t \in \mathrm{cut}(t)$, we have shown that the size of the cut defined by t is at most the number of zero bits in its binary representation plus one. □

4. ANALYSIS

In this section, we prove our main result: a $\widetilde{\Omega}(k^{1/3}T^{2/3})$ lower bound on the expected regret of the multi-armed bandit with switching costs, when the loss functions are stochastic and the player is deterministic. Our result is stated formally in the following theorem. Recall that an instantiation of the random functions $L_{1:T}$ is admissible if the range of each loss function lies in the interval $[0,1]$.

Theorem 2. *Let $L_{1:T}$ be the sequence of stochastic loss functions defined in Fig. 1. Then for $T \ge \max\{k, 6\}$, $L_{1:T}$ is admissible with constant positive probability and the expected regret (as defined in Eq. (1)) of any deterministic player,*

[3] The width of ρ coincides with the cut-width of the numbered graph it determines, see [7].

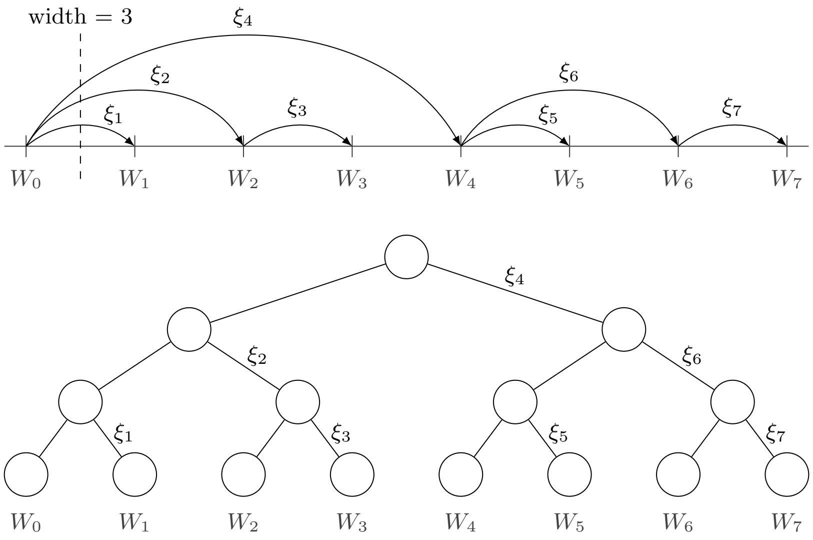

Figure 2: An illustration of the MRW process for $T = 7$. (Top) The MRW with a directed edge from $\rho(t)$ to t, for each $t \in [T]$. (Bottom) The MRW can be equivalently described as the values at the leaves of a binary tree, where the value at each leaf is obtained by summing the i.i.d. Gaussian variables ξ_t's on the (right) edges along the path from the root.

conditioned on the event that $L_{1:T}$ is admissible, is at least $k^{1/3}T^{2/3}/(100\log_2 T)$.

Before we begin with the analysis, we note a delicate point regarding the definition of the random variables $X_{1:T}$. Recall that the player uses a deterministic strategy, one that deterministically maps his past observations of $L_{1:t-1}$ to his next action, X_t. By design, this deterministic mapping is defined for inputs in $[0, 1]$. If the instantiation of the loss sequence is not admissible then the player's next action is undefined. However, notice that the statement of the above theorem is only concerned with the actions of the player's algorithm when the stochastic loss functions are all admissible, and it is independent of the behavior of the player when this is not the case. Therefore, in our analysis we may assume that the player's algorithm can handle any sequence of real loss values as feedback and produces a sequence of actions $X_{1:T}$ regardless of whether the underlying loss functions are admissible or not. For example, we could assume that the player reinterprets his observed values as $\max\big(0, \min(1, L_t(X_t))\big)$ and then applies his standard strategy for bounded losses. Our analysis merely requires that the player has a deterministic response to any sequence of real loss values he observes.

Our analysis requires some new notation. First, let $M = \sum_{t=1}^{T} \mathbb{1}_{X_t \neq X_{t-1}}$ be the number of switches in the action sequence $X_{1:T}$ (recall that we arbitrarily set $X_0 = 1$). Also, for all $t \in [T]$, let $Z_t = L_t(X_t)$ be the loss observed by the player on round t. Recall our assumption that X_t, the player's action on round t, is a deterministic function of his past observations $Z_{1:(t-1)}$.

4.1 Distinguishability Requires Switching

We begin the analysis with a key lemma that relates the player's ability to identify the best action to the number of switches he performs. This lemma also highlights the importance of finding a stochastic process with a small $w(\rho)$. The lemma bounds the distance between each one of the conditional probability measures

$$\mathcal{Q}_i(\cdot) \;=\; \mathbb{P}(\cdot \,|\, \chi = i) \;, \qquad i = 1, 2, \ldots, k \;,$$

and the probability measure $\mathcal{Q}_0$ that corresponds to an (imaginary) adversary that uses $\chi = 0$. Thus $\mathcal{Q}_0(\cdot)$ is the probability when all actions incur the same loss. Let $\mathcal{F}$ be the σ-algebra generated by the player's observations $Z_{1:T}$. Then the *total variation* distance between $\mathcal{Q}_0$ and $\mathcal{Q}_i$ on $\mathcal{F}$ is defined as

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

This distance captures the player's ability to identify whether action i is better than or equivalent to the other actions based on the loss values he observes. The following lemma upper-bounds this distance in terms of the number of switches to action i or from action i, denoted by the random variable M_i, and in terms of the width $w(\rho)$ of the underlying stochastic process. Here we use the notation $\mathbb{E}_{\mathcal{Q}_j}$ to refer to the expectation with respect to the distribution $\mathcal{Q}_j$, for any $j = 0, 1, \ldots, k$.

LEMMA 3. *For all $i \in [k]$, it holds that*

$$d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i) \;\le\; (\epsilon/2\sigma)\sqrt{w(\rho)\,\mathbb{E}_{\mathcal{Q}_0}[M_i]}$$

and

$$d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i) \;\le\; (\epsilon/2\sigma)\sqrt{w(\rho)\,\mathbb{E}_{\mathcal{Q}_i}[M_i]} \;.$$

To see the significance of this lemma, consider first the case $k = 2$, where $M_1 = M_2 = M$ by definition. By the triangle inequality, $d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_1, \mathcal{Q}_2) \le d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_1) + d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_2)$.

Concavity of square root yields

$$\begin{aligned}\sqrt{\mathbb{E}_{\mathcal{Q}_1}[M]} + \sqrt{\mathbb{E}_{\mathcal{Q}_2}[M]} \;&\le\; \sqrt{2\left(\mathbb{E}_{\mathcal{Q}_1}[M] + \mathbb{E}_{\mathcal{Q}_2}[M]\right)} \\ &=\; 2\sqrt{\mathbb{E}[M]}\ .\end{aligned}$$

The second claim of Lemma 3 for $k = 2$ now implies that $d_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_1, \mathcal{Q}_2) \le (\epsilon/\sigma)\sqrt{w(\rho)\,\mathbb{E}[M]}$. This inequality clarifies the dilemma facing the player: If he switches actions frequently so that $\mathbb{E}[M] = \Omega(T^{2/3}/\log(T))$, the switching costs guarantee the desired lower bound on regret. Otherwise, $\mathbb{E}[M] = o(T^{2/3}/\log(T))$; since $\epsilon/\sigma = \Theta(T^{-1/3})$ and $w(\rho) = \Theta(\log(T))$, the distance $d_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_1, \mathcal{Q}_2)$ tends to zero with T, the player is unable to distinguish between the two actions, and he suffers an expected regret of order $\Theta(\epsilon T) = \Theta(T^{2/3}/\log(T))$. We do not formalize this argument here, since we prove the lower bound for any k below.

Proof of Lemma 3. Let $Y_0 = \frac{1}{2}$ and $Y_t = L(X_t)$ for all $t \in [T]$. Note that X_t is a deterministic function of $Y_{0:(t-1)}$. Define $Y_S = \{Y_t\}_{t\in S}$ and let $\Delta(Y_S \mid Y_{S'})$ be the relative entropy (i.e., the Kullback-Leibler divergence) between the joint distribution of Y_S, conditioned on $Y_{S'}$, under $\mathcal{Q}_0$ and $\mathcal{Q}_i$. Namely,

$$\Delta(Y_S \mid Y_{S'}) \;=\; \mathbb{E}_{\mathcal{Q}_0}\left[\log \frac{f_0(Y_S \mid Y_{S'})}{f_i(Y_S \mid Y_{S'})}\right]\ , \qquad (4)$$

where $f_j(Y_S \mid Y_{S'})$ denotes the density of Y_S conditioned on $Y_{S'}$, induced by the measure $\mathcal{Q}_j$. For brevity, also define $\Delta(Y_S) = \Delta(Y_S \mid \emptyset)$. Given $Y_{\rho^*(t)}$, Y_t is conditionally independent[4] of Y_s, for all $s \notin \rho^*(t)$. Using the chain rule for relative entropy (see, e.g., Theorem 2.5.3 in [8]), we can decompose the quantity $\Delta(Y_{0:T})$ as

$$\Delta(Y_{0:T}) \;=\; \Delta(Y_0) + \sum_{t=1}^{T} \Delta\big(Y_t \mid Y_{\rho^*(t)}\big) \qquad (5)$$

and deal separately with each term in the sum. First note that $\Delta(Y_0) = 0$, because Y_0 is a constant. The value of $\Delta\big(Y_t \mid Y_{\rho^*(t)}\big)$ is computed by considering three separate cases. If $X_t = X_{\rho(t)}$ (i.e., the player chooses the same action on rounds t and $\rho(t)$) then the distribution of Y_t conditioned on $Y_{\rho^*(t)}$ is $N(Y_{\rho(t)}, \sigma^2)$ under both $\mathcal{Q}_0$ and $\mathcal{Q}_i$, where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. If $X_t = i$ and $X_{\rho(t)} \ne i$ then the distribution of Y_t conditioned on $Y_{\rho^*(t)}$ is $N(Y_{\rho(t)}, \sigma^2)$ under $\mathcal{Q}_0$ and $N(Y_{\rho(t)} - \epsilon, \sigma^2)$ under $\mathcal{Q}_i$. Finally, if $X_t \ne i$ and $X_{\rho(t)} = i$ then the distribution of Y_t conditioned on $Y_{\rho^*(t)}$ is $N(Y_{\rho(t)}, \sigma^2)$ under $\mathcal{Q}_0$ and $N(Y_{\rho(t)} + \epsilon, \sigma^2)$ under $\mathcal{Q}_i$. Overall,

$$\begin{aligned}\Delta\big(Y_t \mid Y_{\rho^*(t)}\big) \;&=\; \mathcal{Q}_0\left(X_t = i, X_{\rho(t)} \ne i\right) \\ &\qquad \cdot d_{\mathrm{KL}}\left(N(0,\sigma^2) \,\big\|\, N(-\epsilon,\sigma^2)\right) \\ &\quad + \mathcal{Q}_0\left(X_t \ne i, X_{\rho(t)} = i\right) \\ &\qquad \cdot d_{\mathrm{KL}}\left(N(0,\sigma^2) \,\big\|\, N(\epsilon,\sigma^2)\right) \\ &=\; \frac{\epsilon^2}{2\sigma^2}\,\mathcal{Q}_0(A_t)\ , \end{aligned} \qquad (6)$$

where $A_t = \big\{X_t = i, X_{\rho(t)} \ne i \,\vee\, X_t \ne i, X_{\rho(t)} = i\big\}$ is the event that the player switches an odd number of times (and in particular, at least once) to or from action i between rounds $\rho(t)$ and t. Substituting Eq. (6) into Eq. (5) gives

$$\Delta(Y_{0:T}) \;=\; \frac{\epsilon^2}{2\sigma^2}\sum_{t=1}^{T}\mathcal{Q}_0(A_t) \;=\; \frac{\epsilon^2}{2\sigma^2}\,\mathbb{E}_{\mathcal{Q}_0}\left[\sum_{t=1}^{T}\mathbb{1}_{A_t}\right]\ . \quad (7)$$

The event A_t implies that there exists at least one time s of switch to or from action i, such that $t \in \mathrm{cut}(s)$. Therefore, if we let $S_{1:M_i}$ denote the random sequence of times of such switches (in the action sequence $X_{1:T}$), then

$$\sum_{t=1}^{T}\mathbb{1}_{A_t} \;\le\; \sum_{r=1}^{M_i}\sum_{t\in\mathrm{cut}(S_r)}\mathbb{1}_{A_t} \;\le\; \sum_{r=1}^{M_i}|\mathrm{cut}(S_r)| \;\le\; w(\rho)\,M_i\ .$$

Plugging this inequality back into Eq. (7) gives

$$\Delta(Y_{0:T}) \;\le\; \frac{\epsilon^2 w(\rho)}{2\sigma^2}\,\mathbb{E}_{\mathcal{Q}_0}[M_i]\ .$$

Pinsker's inequality (Lemma 11.6.1 in [8]) now implies that

$$\sup_{A\in\mathcal{F}'}\big(\mathcal{Q}_0(A) - \mathcal{Q}_i(A)\big) \;\le\; \frac{\epsilon}{2\sigma}\sqrt{w(\rho)\,\mathbb{E}_{\mathcal{Q}_0}[M_i]}\ ,$$

where $\mathcal{F}'$ is the σ-algebra generated by $Y_{0:T}$. We can replace $\mathcal{F}'$ with $\mathcal{F}$ above to obtain $d_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_i)$ in the left-hand side, simply because $Z_{1:T}$ is a deterministic function of $Y_{0:T}$ and therefore $\mathcal{F} \subset \mathcal{F}'$.

This proves the first claim of the lemma. To prove the second bound, we can simply reverse the roles of $\mathcal{Q}_0$ and $\mathcal{Q}_i$ in our arguments above and obtain the same bound over the total variation distance but in terms of the expectation with respect to the distribution $\mathcal{Q}_i$. □

4.2 Regret Lower Bound

With Lemma 3 in hand, we can prove Theorem 2 and conclude Theorem 1. We begin with a simple corollary of the lemma.

Corollary 1. *It holds that*

$$\frac{1}{k}\sum_{i=1}^{k} d_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_i) \;\le\; \frac{\epsilon}{\sigma\sqrt{2k}}\cdot\sqrt{w(\rho)\,\mathbb{E}_{\mathcal{Q}_0}[M]}\ .$$

Proof. Averaging the inequalities of Lemma 3 over $i = 1, 2, \ldots, k$, using the concavity of the root function and noting that $\sum_{i=1}^{k} M_i = 2M$ (as each switch is counted twice in the sum) yields

$$\begin{aligned}\frac{1}{k}\sum_{i=1}^{k} d_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_0, \mathcal{Q}_i) \;&\le\; \frac{\epsilon}{2\sigma}\cdot\frac{1}{k}\sum_{i=1}^{k}\sqrt{w(\rho)\,\mathbb{E}_{\mathcal{Q}_0}[M_i]} \\ &\le\; \frac{\epsilon}{\sigma\sqrt{2k}}\cdot\sqrt{w(\rho)\,\mathbb{E}_{\mathcal{Q}_0}[M]}\ ,\end{aligned}$$

as claimed. □

We now turn to analyzing the player's expected regret. Using the definitions above, this regret can be written as

$$R \;=\; \sum_{t=1}^{T} L_t(X_t) + M - \min_{x\in[k]}\sum_{t=1}^{T} L_t(x)\ .$$

The next lemma shows that the conditional expectation in the statement of Theorem 2 is not much smaller than the unconditional expected regret, $\mathbb{E}[R]$.

[4] Note that this conditional independence property holds for both distributions $\mathcal{Q}_i$ and $\mathcal{Q}_0$.

LEMMA 4. *Consider the event*

$$B \;=\; \{\forall\, t \in [T], x \in [k] \;:\; L_t(x) \in [0,1]\}$$

under which the loss sequence is admissible. Then

$$\mathbb{E}[R \mid B] \;\geq\; \mathbb{E}[R] - \epsilon T/6$$

provided that $T \geq \max\{k, 6\}$.

PROOF. We first show that $\mathbb{P}(B) \geq 5/6$. As the process $W_{1:T}$ has depth $d \leq \lfloor \log_2 T \rfloor + 1 \leq 2\log_2 T$, Lemma 1 with $\delta = 1/T \leq 1/6$ implies that with probability at least 5/6, we have

$$|W_t| \;\leq\; \sigma\sqrt{2d\log\tfrac{T}{\delta}} \;\leq\; \sigma\sqrt{8\log_2 T \log T} \;\leq\; 3\sigma\log_2 T$$

for all $t \in [T]$. Thus, setting $\sigma = 1/(9\log_2 T)$ we obtain that

$$\mathbb{P}\left(\forall t \in [t] \quad \tfrac{1}{2} + W_t \in \left[\tfrac{1}{6}, \tfrac{5}{6}\right]\right) \;\geq\; \frac{5}{6}\;.$$

For $T \geq \max\{k, 6\}$ we have $\epsilon < 1/6$ and thus $L_t(x) \in [0,1]$ for all $x \in [k]$ whenever $\frac{1}{2} + W_t \in [\frac{1}{6}, \frac{5}{6}]$. This implies that $\mathbb{P}(B) \geq 5/6$.

If B does not occur then the regret can be at most ϵT (since the per-round regret is at most ϵ). Thus $\mathbb{E}[R \mid \neg B] \leq \epsilon T$ and so

$$\begin{aligned}
\mathbb{E}[R \mid B] \;&\geq\; \mathbb{E}[R] - \mathbb{E}[R \mid \neg B]\cdot\mathbb{P}(\neg B)\\
&\geq\; \mathbb{E}[R] - \epsilon T/6\;,
\end{aligned}$$

as required. □

Next, we relate the expected regret to the total variation between $\mathcal{Q}_0$ and the $\mathcal{Q}_i$'s.

LEMMA 5. *The expected regret* $\mathbb{E}[R]$ *is lower bounded in terms of the distributions* $\mathcal{Q}_0, \mathcal{Q}_1, \ldots, \mathcal{Q}_k$ *as*

$$\mathbb{E}[R] \;\geq\; \frac{\epsilon T}{2} - \frac{\epsilon T}{k}\cdot\sum_{i=1}^{k} d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i) + \mathbb{E}[M]\;.$$

PROOF. For $i \in [k]$, let N_i denote the number of times the player picks action i, so we can write $R = \epsilon\,(T - N_\chi) + M$. Consequently,

$$\begin{aligned}
\mathbb{E}[R] \;&=\; \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[\epsilon\,(T - N_i) + M \mid \chi = i]\\
&=\; \epsilon T - \frac{\epsilon}{k}\sum_{i=1}^{k}\mathbb{E}_{\mathcal{Q}_i}[N_i] + \mathbb{E}[M]\;. \qquad (8)
\end{aligned}$$

On the other hand, for all $i \in [k]$ and $t \in [T]$, the event $\{X_t = i\}$ is in the σ-field $\mathcal{F}$, so $\mathcal{Q}_i(X_t = i) - \mathcal{Q}_0(X_t = i) \leq d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i)$. Summing over $t = 1, \ldots, T$ yields $\mathbb{E}_{\mathcal{Q}_i}[N_i] - \mathbb{E}_{\mathcal{Q}_0}[N_i] \leq T \cdot d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i)$, whence

$$\begin{aligned}
\sum_{i=1}^{k}\mathbb{E}_{\mathcal{Q}_i}[N_i] \;&\leq\; T\cdot\sum_{i=1}^{k} d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i) + \sum_{i=1}^{k}\mathbb{E}_{\mathcal{Q}_0}[N_i]\\
&=\; T\cdot\sum_{i=1}^{k} d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i) + T\;.
\end{aligned}$$

Plugging this into Eq. (8) and using $k \geq 2$ gives

$$\begin{aligned}
\mathbb{E}[R] \;&\geq\; \epsilon T - \frac{\epsilon T}{k}\cdot\sum_{i=1}^{k} d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i) - \frac{\epsilon T}{k} + \mathbb{E}[M]\\
&\geq\; \frac{\epsilon T}{2} - \frac{\epsilon T}{k}\cdot\sum_{i=1}^{k} d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i) + \mathbb{E}[M]\;,
\end{aligned}$$

as claimed. □

We are now ready to prove Theorem 2.

PROOF OF THEOREM 2. We first prove the theorem for deterministic players that make no more than ϵT switches on *any* sequence of loss functions, and relax this assumption toward the end of the proof. For algorithms with this property, we have $\mathcal{Q}_0(M > \epsilon T) = \mathcal{Q}_i(M > \epsilon T) = 0$ for all $i \in [k]$. Since $\{M \geq m\} \in \mathcal{F}$, this implies

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}_0}[M] - \mathbb{E}_{\mathcal{Q}_i}[M] \;&=\; \sum_{m=1}^{\lfloor \epsilon T \rfloor}\left(\mathcal{Q}_0(M \geq m) - \mathcal{Q}_i(M \geq m)\right)\\
&\leq\; \epsilon T \cdot d^{\mathcal{F}}_{\mathrm{TV}}(Q_0, Q_i)
\end{aligned}$$

for all $i \in [k]$, that gives

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}_0}[M] - \mathbb{E}[M] \;&=\; \frac{1}{k}\sum_{i=1}^{k}\left(\mathbb{E}_{\mathcal{Q}_0}[M] - \mathbb{E}_{\mathcal{Q}_i}[M]\right)\\
&\leq\; \frac{\epsilon T}{k}\sum_{i=1}^{k} d^{\mathcal{F}}_{\mathrm{TV}}(Q_0, Q_i)\;.
\end{aligned}$$

Combining this with the results of Lemma 4 and Lemma 5, we obtain

$$\mathbb{E}[R \mid B] \;\geq\; \frac{\epsilon T}{3} - \frac{2\epsilon T}{k}\sum_{i=1}^{k} d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i) + \mathbb{E}_{\mathcal{Q}_0}[M]\;,$$

where B is the event defined in Lemma 4.

On the other hand, recall Lemma 2 that states that the width of the MRW process is bounded by $w(\rho) \leq \lfloor \log_2 T \rfloor + 1 \leq 2\log_2 T$. Corollary 1 together with this bound gives

$$\frac{1}{k}\sum_{i=1}^{k} d^{\mathcal{F}}_{\mathrm{TV}}(\mathcal{Q}_0, \mathcal{Q}_i) \;\leq\; \frac{\epsilon}{\sigma\sqrt{k}}\cdot\sqrt{\mathbb{E}_{\mathcal{Q}_0}[M]\,\log_2 T}\;.$$

Plugging this into the previous inequality and using the notation $m = \sqrt{\mathbb{E}_{\mathcal{Q}_0}[M]}$ results with the lower bound

$$\mathbb{E}[R \mid B] \;\geq\; \frac{\epsilon T}{3} + m\left(m - \frac{2\epsilon^2}{\sigma\sqrt{k}}\,T\sqrt{\log_2 T}\right)\;.$$

The right hand side, which is minimized at

$$m \;=\; (\epsilon^2/\sigma\sqrt{k})\,T\sqrt{\log_2 T}\;,$$

can be further lower bounded by $\epsilon T/3 - (\epsilon^4/\sigma^2 k)\,T^2\log_2 T$. Using our choice of

$$\sigma \;=\; \frac{1}{9\log_2 T} \quad\text{and}\quad \epsilon \;=\; \frac{k^{1/3}T^{-1/3}}{9\log_2 T}$$

gives

$$\mathbb{E}[R \mid B] \;\geq\; \left(\frac{1}{27} - \frac{1}{81}\right)\cdot\frac{k^{1/3}T^{2/3}}{\log_2 T} \;\geq\; \frac{k^{1/3}T^{2/3}}{50\log_2 T}\;. \qquad (9)$$

This proves the theorem for algorithms with the assumed property. In order to relax this assumption, note that we can turn any player algorithm to an algorithm that makes at most ϵT switches, simply by halting the algorithm once it makes $\lfloor \epsilon T \rfloor$ switches and repeating its last action on the remaining rounds. The regret R^* of the modified algorithm equals R unless $M > \epsilon T$ and in the latter case $R^* \leq R + \epsilon T \leq 2R$ as the per-round regret can be at most ϵ, so $\mathbb{E}[R^* \mid B] \leq$

$2\,\mathbb{E}[R \mid B]$. Since $\mathbb{E}[R^* \mid B]$ is lower bounded by the right-hand side of Eq. (9), this implies the claimed lower bound on the expected regret of any deterministic player. □

Finally, we can prove Theorem 1.

Proof of Theorem 1. Recall that any randomized algorithm is equivalent to an a-priori random choice of a deterministic algorithm, for which the statement of Theorem 2 applies. Since the adversary is oblivious to the player's actions, the statement of Theorem 2 for a randomized player (where the expectation is now taken with respect to both the functions $L_{1:T}$ and the player's random bits) follows by taking the expectation over its internal randomization.

The fact that the expected regret with respect to the randomization in $L_{1:T}$, conditioned on the (probable) event that $L_{1:T}$ is admissible, is lower bounded by the stated quantity implies that there exists some admissible realization $\ell_{1:T}$ for which the regret is lower bounded by the same quantity. □

5. EXTENSIONS AND IMPLICATIONS

In this section we present few extensions of our results and discuss several implications.

5.1 Binary losses

In our construction of a randomized adversary, described in Sec. 2, the loss values $L_t(x)$ are all real numbers in the interval $[0, 1]$. One might wonder whether a similar construction exists where each of the loss values is constrained to be either 0 or 1. A simple adaptation of our construction shows that this is indeed the case. To see this, simply set the loss of action x at time t to be the outcome of a biased coin toss with bias $L_t(x)$; in case $L_t(x) \notin [0, 1]$, set this loss value arbitrarily. In this sequence of binary loss functions, action χ is consistently better *in expectation* by an ϵ gap whenever the loss values are all bounded in the $[0, 1]$ interval, which is sufficient in our analysis. Our arguments regarding the player's inability to identify the best action still apply since the feedback he observes is only further obscured by additional random noise.

5.2 Arbitrary Switching Cost

Assume that each switch incurs a cost of c to the player, instead of a unit cost as before. Repeating the proof of Theorem 2, we are able to get an $\widetilde{\Omega}(c^{1/3}k^{1/3}T^{2/3})$ lower bound, which is tight with respect to T, k and c (up to poly-log factors) in light of the upper bound of Arora et al. [1].

Theorem 3. *Let the cost of switch be $c > 0$ and assume that $T > c \cdot \max\{k, 6\}$. For any randomized player strategy that relies on bandit feedback, there exists a sequence of loss functions $\ell_{1:T}$ (where $\ell_t : [k] \mapsto [0, 1]$) that incurs a regret of $R = \widetilde{\Omega}(c^{1/3}k^{1/3}T^{2/3})$.*

Proof. Redefine the gap between the actions in the construction of the functions $L_{1:T}$ to $\epsilon = (ck)^{1/3}T^{-1/3}/(9\log_2 T)$. Using the same notation as in the proof of Theorem 2, we can show that

$$\mathbb{E}[R \mid B] \;\geq\; \frac{\epsilon T}{3} + m\left(cm - \frac{2\epsilon^2}{\sigma\sqrt{k}}\,T\sqrt{\log_2 T}\right) .$$

The right-hand side is minimized at $m = (\epsilon^2/c\sigma\sqrt{k})\,T\sqrt{\log_2 T}$ and is lower bounded by $\epsilon T/3 - (\epsilon^4/\sigma^2 ck)\,T^2\log_2 T$. Setting $\epsilon = (ck)^{1/3}T^{-1/3}/(9\log_2 T)$ and using our choice of $\sigma = 1/(9\log_2 T)$ gives the lower bound

$$\mathbb{E}[R \mid B] \;\geq\; \frac{c^{1/3}k^{1/3}T^{2/3}}{50\log_2 T} .$$

Proceeding as in the proofs of Theorem 2 and Theorem 1, we establish the existence of the admissible sequence $\ell_{1:T}$. □

5.3 Tradeoff between Loss and Switches

As a corollary of Theorem 3, we can quantify the tradeoff between the loss accumulate by a multi-armed bandit algorithm and the number of switches it performs. For simplicity, we treat the number of actions k as a constant and state the result only in terms of T.

Theorem 4. *Let $\mathcal{A}$ be a multi-armed bandit algorithm that guarantees an expected regret (without switching costs) of $\widetilde{O}(T^\alpha)$. Then, there exists a sequence of loss functions that forces $\mathcal{A}$ to make $\widetilde{\Omega}(T^{2(1-\alpha)})$ switches.*

In particular, the popular EXP3 algorithm [3] guarantees a regret of $O(\sqrt{T})$ without switching costs. In this case, Theorem 4 implies that EXP3 can be forced to make $\widetilde{\Omega}(T)$ switches.

Proof of Theorem 4. Assume the contrary, i.e. that $\mathcal{A}$ can guarantee a regret of $\widetilde{O}(T^\alpha)$ (without switching costs) with $\widetilde{O}(T^\beta)$ switches over any sequence of T loss functions, with $\alpha + \beta/2 < 1$. In this case, we can pick a real number γ such that $\alpha < \gamma < 1 - \beta/2$. Consider the performance of this algorithm in a setting where the cost of a switch is $c = T^{3\gamma-2}$. Clearly, the expected regret (including switching costs) of the algorithm in this setting is upper bounded by

$$\widetilde{O}(T^\alpha + T^{3\gamma-2} \cdot T^\beta) \;=\; \widetilde{o}(T^\gamma) \;\;,$$

over any sequence of loss functions, as $\alpha < \gamma$ and $\beta < 2 - 2\gamma$. This contradicts Theorem 3, which guarantees the existence of a loss sequence that incurs a regret (including switching costs) of $\widetilde{\Omega}(T^{(3\gamma-2)/3} \cdot T^{2/3}) = \widetilde{\Omega}(T^\gamma)$. □

5.4 Lower Bound for Online Adversarial Markov Decision Processes

The multi-armed bandit problem with switching costs is a special case of the online adversarial deterministic Markov decision process (ADMDP) with bandit feedback (see Dekel and Hazan [9] for a formal description of this setting). The important aspect of the ADMDP setting is that the player has a state, and that his loss on each round depends both on his action and on his current state. Moreover, the player's action on round t determines his state on round $t + 1$. The k-armed bandit problem with switching costs can be described as a k-state ADMDP, where each state represents the player's previous action. The player incurs the loss associated with the action he chooses and pays an additional cost whenever he changes his state.

As a result, our lower bound applies to the class of ADMDP problems. Dekel and Hazan [9] proves a matching upper bound, which implies that the (undiscounted) minimax regret of the ADMDP problem is $\widetilde{\Theta}(T^{2/3})$. The ADMDP setting belongs to the more general class of adversarial MDPs with bandit feedback [17, 15], where the state transitions are allowed to be stochastic. This implies a $\widetilde{\Omega}(T^{2/3})$ lower bound on the (undiscounted) minimax regret of the general setting.

6. CONCLUSION

In this paper, we proved that the T-round k-action multi-armed bandit problem with switching costs has a minimax regret of $\widetilde{\Theta}(k^{1/3}T^{2/3})$, and is therefore strictly harder than the corresponding experts problem (with full feedback). To the best of our knowledge, this is the first example of a setting in which learning with bandit feedback is significantly harder than learning with full-information feedback (in terms of the dependence on T). Furthermore, we believe our work is the first to demonstrate that online learning may become significantly easier if the loss sequences are known to be i.i.d. over time[5]. Our analysis shows that the difficulty of the problem stems from the player's need to *pay for exploring* the quality of the different actions.

Since our problem is a special case of online learning with bandit feedback against a bounded-memory adaptive adversary, we conclude that the minimax regret of the general setting is also $\widetilde{\Omega}(T^{2/3})$, which matches the upper bounds of Arora et al. [1]. We also showed how our construction resolves several other open problems in online learning. Moreover, we believe that the multi-scale random walk, defined in Sec. 3.2, will prove to be a useful tool in other settings.

References

[1] 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.

[2] J.-Y. Audibert, S. Bubeck, et al. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22th annual conference on learning theory (COLT)*, pages 217–226, 2009.

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

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

[5] 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.

[6] 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.

[7] F. R. K. Chung and P. D. Seymour. Graphs with small bandwidth and cutwidth. *Discrete Mathematics*, 75(1-3):113–119, 1989.

[8] T. Cover and J. Thomas. *Elements of information theory.* John Wiley & Sons, 2006.

[9] O. Dekel and E. Hazan. Better rates for any adversarial deterministic MDP. In *Proceedings of the Thirtieth International Conference on Machine Learning*, 2013.

[10] Y. Freund and R. 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.

[11] 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.

[12] 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.

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

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

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

[16] 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.

[17] J. Y. Yu, S. Mannor, and N. Shimkin. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

[5] When the loss sequences are i.i.d., Cesa-Bianchi et al. [6] show that the multi-armed bandit problem with switching costs has a minimax regret of $\widetilde{\Theta}(\sqrt{T})$.