Better Best of Both Worlds Bounds for Bandits with Switching Costs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study best-of-both-worlds algorithms for bandits with switching cost, recently addressed by Rouyer, Seldin, and Cesa-Bianchi [12]. We introduce a surprisingly simple and effective algorithm that simultaneously achieves minimax optimal regret bound of $\mathcal{O}(T^{2/3})$ in the oblivious adversarial setting and a bound of $\mathcal{O}(\min\{\log(T)/\Delta^2, T^{2/3}\})$ in the stochastically-constrained regime, both with (unit) switching costs, where Δ is the gap between the arms. In the stochastically constrained case, our bound improves over previous results due to [12], that achieved regret of $\mathcal{O}(T^{1/3}/\Delta)$. We accompany our results with a lower bound showing that, in general, $\tilde{\Omega}(\min\{1/\Delta^2, \Delta T\})$ regret is unavoidable in the stochastically-constrained case.

1 Introduction

Multi Armed Bandit (MAB) is one of the most fundamental problems in online learning and sequential decision making. This problem is often framed as a sequential game between a player and an environment played over T rounds. In each round, the player chooses an action from a finite set $[K] = \{1, \dots, K\}$ and incurs a loss in $[0, 1]$ for that action. The environment then only reveals the loss of the chosen action—this is referred to as bandit feedback. The goal of the player is to minimize the *regret*, which measures the difference between the cumulative loss of the player and that of the best arm in hindsight.

Two common regimes often studied in the MAB literature are the *adversarial* setting [6] and the so-called *stochastically-constrained* setting [16] which is a generalization of the more classical stochastic setting. In the former regime, losses are generated arbitrarily and possibly by an adversary; in the latter, losses are assumed to be generated in a way that one arm performs better (in expectation) than any other arm, by a margin of $\Delta > 0$. Both regimes have witnessed a flurry of research [1, 3, 5, 6, 11, 15] leading to optimal regret bounds in each of the settings.

Recently, significant effort has been dedicated for designing *best-of-both-worlds* MAB algorithms, where one does not have a-priori knowledge on the underlying environment but still wishes to enjoy the optimal regret in both regimes simultaneously [4, 7, 13, 14, 16, 17]. Most notably, Zimmert and Seldin [17] analyzed the *Tsallis-INF* algorithm and established that it achieves optimal regret bounds in both stochastically-constrained and adversarial environments, matching the corresponding lower bounds asymptotically.

Another well-studied variant of the MAB setting is that of Bandits with switching cost [2, 8–10], where the learner suffers not only regret but also a penalty for switching actions. As shown by Dekel et al. [8] adding switching cost to the regret incurs a lower bound of $\tilde{\Omega}\left((\lambda K)^{1/3} T^{2/3}\right)$, in contrast to $\Omega(\sqrt{KT})$ in the standard setting, highlighting the difficulty of this setup. Recently Rouyer et al. [12] asked the question of how best-of-both-worlds algorithms can be obtained for MAB where switching

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

costs are considered. Rouyer et al. [12] managed to show a best-of-both-worlds type algorithm that, for constant K and cost-per-switch, achieves optimal regret bound of $\mathcal{O}(T^{2/3})$ in the oblivious adversarial regime, whereas in the stochastically-constrained setup their upper bound is of order $\mathcal{O}(T^{1/3}/\Delta)$, which was unknown to be optimal.

In this work we tighten the above gap. We provide a lower regret bound, and show that for sufficiently large Δ, no algorithm can obtain regret smaller than $\tilde{\Omega}(\min\{1/\Delta^2, \Delta T\})$. Additionally, we introduce a new algorithm that improves the bound of [12] and achieves the above best-of-both-worlds regret up to logarithmic factors. For the more general case of $K > 2$ arms, our algorithm still improves over [12], however in that case our lower bounds and upper bound do not fully coincide; we leave this as an open question for future study.

2 Setup and Background

In the classic Multi Armed Bandit problem with K arms, a game is played consecutively over $T > K$ rounds. At each round $t \leq T$, an adversary (also called the environment) generates a loss vector $\ell_t \in [0,1]^K$. The player (referred to as learner) selects an arm $I_t \in [K]$ according to some distribution $p_t \in \Delta^K$ where $\Delta^K := \{p \in [0,1]^K : \sum_{i\in[K]} p_i = 1\}$, and observes ℓ_{t,I_t}, which is also defined to be its loss at round t. Notice, that the learner never has access to the entire loss vector $\ell_t \in [0,1]^K$.

The performance of the learner is measured in terms of the *regret*. The regret of the learner is defined as

$$\mathcal{R}_T := \sum_{t\in[T]} \ell_{t,I_t} - \min_{i\in[K]} \sum_{t\in[T]} \ell_{t,i}.$$

Another common performance measure we care about is the *pseudo-regret* of the algorithm:

$$\overline{\mathcal{R}}_T := \sum_{t\in[T]} \ell_{t,I_t} - \min_{i\in[K]} \sum_{t\in[T]} \mathbb{E}\big[\ell_{t,i}\big].$$

We next describe two common variants of the problem, which differ in the way the losses are generated.

Adversarial (oblivious) regime: In the *oblivious adversarial* regime, at the beginning of the game the environment chooses the loss vectors $\ell_1, \ldots, \ell_T$, and they may be entirely arbitrary. In general, the objective of the learner is to minimize its expected regret for the adversarial regime. One can observe that the *expected pseudo-regret* coincides with the *expected regret*, in this setting. More generally, it can be seen that the expected regret upper bounds the expected pseudo-regret. Namely, $\mathbb{E}[\overline{\mathcal{R}}_T] \leq \mathbb{E}[\mathcal{R}_T]$.

Stochastically-constrained adversarial regime: We also consider the *stochastically-constrained adversarial* regime [16]. In this setting we assume that the loss vectors are drawn from distributions such that there exists some $i^\star \in [K]$,

$$\forall i \neq i^\star : \ \mathbb{E}[\ell_{t,i} - \ell_{t,i^\star}] = \Delta_i, \tag{1}$$

independently of t[1]. That is, the gap between arms remains constant throughout the game, while the losses $\{\ell_{t,i}\}_{t\in[T]}$ of any arm i are drawn from distributions that are allowed to change over time and may depend on the learner's past actions $I_1, \ldots, I_{t-1}$. It is well known that the stochastically-constrained adversarial regime generalizes the well-studied stochastic regime that assumes the losses are generated in an i.i.d. manner.

We denote the best arm at round t to be $i_t^\star = \arg\min_{i\in[K]} \mathbb{E}[\ell_{t,i}]$. Note that since the gap between arms is constant we have that $\forall t \in [T] : i_t^\star = i^\star$ where $i^\star$ is the optimal arm. We consider the case where there is a *unique* best arm. Also, we denote the gap between arm i and $i^\star$ to be $\Delta_i := \mathbb{E}[\ell_{t,i} - \ell_{t,i^\star}]$ where $\Delta_{\min} = \min_{i\neq i^\star} \Delta_i$.

[1]This definition is equivalent to the standard definition of $\forall i, j : \ \mathbb{E}[\ell_{t,i} - \ell_{t,i^\star}] = \Delta_{i,j}$.

We note that in the stochastically-constrained case, the pseudo-regret is often expressed by the sub-optimality gaps Δ_i, and it is given by:

$$\mathbb{E}\left[\overline{\mathcal{R}}_T\right] := \sum_{t\in[T]}\sum_{i\neq i^\star}\mathbb{P}(I_t = i)\Delta_i.$$

2.1 Multi-Armed Bandits with Switching Cost

In the problem described above, there is no limitation on the number of times the player is allowed to switch arms between consecutive rounds. In this work, we consider a setup where the regret is accompanied by a switching cost, denoted by $\lambda \geq 0$ as suggested by Arora et al. [2]. Therefore, we measure our performance by the switching cost regret, denoted by:

$$\overline{\mathcal{R}}_T^\lambda := \overline{\mathcal{R}}_T + \lambda\mathcal{S}_T.$$

Where $\mathcal{S}_T := \sum_{t\in[T]}\mathbb{1}\{I_t \neq I_{t-1}\}$.

Best-of-both-worlds with switching cost. Rouyer et al. [12] considered the setting of switching cost in the framework of best-of-both-worlds analysis, we state their main algorithmic result:

Theorem ([12, Thm 1])**.** *There exists an algorithm, Tsallis-Switch, for which in the adversarial regime the pseudo-regret of Tsallis-Switch for $\lambda \in \Omega(\sqrt{K/T})$:*

$$\mathbb{E}[\overline{\mathcal{R}}_T] \leq \mathcal{O}\left((\lambda K)^{1/3}T^{2/3}\right), \tag{2}$$

and in the stochastically constrained setting:

$$\mathbb{E}[\overline{\mathcal{R}}_T] \leq \mathcal{O}\Big(\sum_{i\neq i^\star}\frac{(\lambda K)^{2/3}T^{1/3} + \log T}{\Delta_i}\Big) \tag{3}$$

3 Main results

Theorem 1. *Provided that $\lambda \geq \sqrt{K/T}$, the expected pseudo-regret with switching cost of "Switch Tsallis, Switch!" (Algorithm 1) satisfies the following simultaneously:*

- *In the adversarial regime,*

$$\mathbb{E}\left[\overline{\mathcal{R}}_T^\lambda\right] = \mathcal{O}\big((\lambda K)^{1/3}T^{2/3}\big). \tag{4}$$

- *In the stochastically constrained adversarial regime,*

$$\mathbb{E}\left[\overline{\mathcal{R}}_T^\lambda\right] = \mathcal{O}\Big(\min\Big\{\Big(\frac{\lambda\log T}{\Delta_{min}} + \log T\Big)\sum_{i\neq i^\star}\frac{1}{\Delta_i}, (\lambda K)^{1/3}T^{2/3}\Big\}\Big). \tag{5}$$

We next compare the bound of "Switch Tsallis, Switch!" and Rouyer et al. [12]. We first observe that for small switching cost, $\lambda \leq O\big(\sqrt{K/T}\big)$, both algorithms basically ignore the switching cost and run standard Tsallis without any type of change hence the algorithms actually coincide, so we only care for the case $\lambda \geq \sqrt{\frac{K}{T}}$. Also, notice that in the adversarial regime Eq. (2) and Eq. (4) are equivalent and both algorithms obtain the minimax optimal regret. In the stochastically constrained regime comparing Eqs. (3) and (5), note that for $\Delta_{\min} \leq (\lambda K)^{1/3}T^{-1/3}\log T$, we have that

$$\sum_{i\neq i^\star}\frac{(\lambda K)^{2/3}T^{1/3}}{\Delta_i} \geq \frac{(\lambda K)^{2/3}T^{1/3}}{\Delta_{\min}} = \tilde{\Omega}\left((\lambda K)^{1/3}T^{2/3}\right),$$

which is comparable to our bound up to logarithmic factors. On the other hand, if $\Delta_{\min} \geq (\lambda K)^{1/3}T^{-1/3}\log T$

$$\sum_{i\neq i^\star}\frac{(\lambda K)^{2/3}T^{1/3}}{\Delta_i} \geq \left(\frac{\lambda K\log T}{\Delta_{\min}}\right)\sum\frac{1}{\Delta_i} = \Omega\left(\frac{\lambda\log T}{\Delta_{\min}}\right),$$

which is comparable to Eq. (5). It can also be observed that when $\Delta_{\min}$ is large enough (larger than say $T^{-1/3}$) our bound improves over Eq. (3) by a factor of $O(T^{1/3}\Delta_{\min})$.

Next, we describe a lower bound, which demonstrate that our bounds are tight for $K = 2$.

Theorem 2. *For any randomized player strategy in a multi armed bandit game of K arms played over T rounds, there exists a stochastically-constrained sequence of deterministic losses $\ell_1, \ldots, \ell_T$ with minimal gap parameter Δ, that incurs $\overline{\mathcal{R}}_T + \mathfrak{S}_T = \tilde{\Omega}\big(\min\{1/\Delta^2, \Delta T\}\big)$.*

4 Algorithm

Our algorithm, "Switch Tsallis, Switch!" (see Algorithm 1), is a simple modification of Tsallif-INF. We start by playing the original Tsallis-INF algorithm introduced by [17], and after a certain amount of switches we switch to a second phase that plays a standard block no-regret algorithm.

Algorithm 1 Switch Tsallis, Switch!

Input: time horizon T, switching cost λ.

Initialize: $S = 0$, $\hat{\ell}_0 = \mathbf{0}_K$, $\eta_t = 2/\sqrt{t}$

for $t = 1, \ldots, T$ **do**

  Compute:

$$p_t = \arg\min_{p \in \Delta^K} \left\{ \sum_{r=0}^{t-1} \hat{\ell}_r \cdot p - \frac{1}{\eta_t} \sum_{i \in [K]} 4\sqrt{p_i} \right\}.$$

  Sample $I_t \sim p_t$, play I_t and observe the loss ℓ_{t,I_t}.

  Update:

$$\forall i \in [K] : \ \hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i}} \mathbb{1}\{I_t = i\}$$

$$S = S + \mathbb{1}\{I_t \neq I_{t-1}\}$$

  if $S \geq K^{1/3}(\frac{T}{\lambda})^{2/3} + \frac{1}{\lambda}$ **then**

    break

if $t < T$ **then** for remaining rounds

  Run Tsallis-INF over blocks of size $\lambda^{2/3} K^{-1/3} T^{1/3}$.

The idea is motivated by our observation that under the stochastically constrained setting, there is a natural bound on the number of switches which is of order $O(\overline{\mathcal{R}}_T/\Delta_{\min})$, so as long as this number doesn't exceed the worst case bound of the adversarial setting we have no reason to perform blocks. In other words, we start by playing under the assumption that we are in the stochastically-constrained regime and if the number of switches is larger than expected, we break and move to an algorithm that handles only the oblivious adversarial case.

Best-of-both-worlds algorithms that start under stochasticity assumption and break are natural to consider. Indeed, in the standard setting, without switching they were studied and suggested [4, 7]. However, while successful at the stochastic case, they suffer from a logarithmic factor in the adversarial regime. Moreover, it is generally quite involve to verify and decide when and how to switch between the regimes. In contrast, what we observe here, is that once switching are involved, the criteria to switch between the regimes becomes quite straightforward which allows us to design such a simple algorithm. Indeed, unlike regret which is hard to estimate, the cost of switching is apparent to the learner, hence we can verify when our switching loss exceeds what we expect in the adversarial regime and decide to switch to enjoy both worlds.

5 Proofs

5.1 Technical Preliminaries

Theorem 3 ([18, Thm 1])**.** *The expected pseudo-regret of Tsallis-INF under the adversarial regime satisfies:*

$$\mathbb{E}[\overline{\mathcal{R}}_T] \leq 4\sqrt{KT} + 1, \tag{6}$$

and in the stochastically constrained setting:

$$\mathbb{E}[\overline{\mathcal{R}}_T] \leq \mathcal{O}\left(\sum_{i \neq i^\star} \frac{\log T}{\Delta_i} \right) \tag{7}$$

Arora et al. [2] showed how, given a no-regret algorithm against an oblivious adversary, one can convert the algorithm to be played over mini-batches of size B and obtain a regret of $B\,\mathbb{E}\left[\mathcal{R}_{T/B}\right]$. In turn, the regret with switching cost is bounded by $\mathbb{E}\left[\mathcal{R}_T^{\lambda}\right] = B\,\mathbb{E}\left[\mathcal{R}_{T/B}\right] + \lambda\frac{T}{B}$. Using the above bound for the special case of Tsallis-INF, as observed by [12], we obtain the following guarantee:

Corollary 4. *There exists an algorithm (in particular Tsallis-INF with constant blocks of size $B = O\left(\lambda^{2/3}K^{-1/3}T^{1/3}\right)$), for which the expected regret with switching cost, satisfies*

$$\mathbb{E}[\mathcal{R}_T^{\lambda}] \leq 4(\lambda K)^{1/3}T^{2/3} + 1$$

Indeed, using the above observation,

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_T^{\lambda}] &\leq B\sqrt{4KT/B} + 1 + \lambda T/B. && \text{(Eq. (6))}\\
&\leq 4(\lambda K)^{1/3}T^{2/3} + 1 && (B := \lambda^{2/3}K^{-1/3}T^{1/3})
\end{aligned}$$

5.2 Proof of Theorem 1

"Switch Tsallis, Switch!" consists of two parts. We will denote the regret and pseudo-regret attained after the first part by $\mathcal{R}^{(1)}$ and $\overline{\mathcal{R}}^{(1)}$ respectively, while the regret and pseudo-regret achieved in Line 9 are denoted by $\mathcal{R}^{(2)}$ and $\overline{\mathcal{R}}^{(2)}$. Similarly, $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ express the number of switches of each segment. The proof of the adversarial case, i.e. Eq. (4) is straightforward and follows by explicitly bounding the regret in each of these phases:

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_T^{\lambda}\right] &\leq \mathbb{E}\left[\mathcal{R}^{(1)} + \lambda\mathcal{S}^{(1)}\right] + \mathbb{E}\left[\mathcal{R}^{(2)} + \lambda\mathcal{S}^{(2)}\right]\\
&\leq \mathbb{E}\left[\mathcal{R}^{(1)} + \lambda\mathcal{S}^{(1)}\right] + 4(\lambda K)^{1/3}T^{2/3} + 1 && \text{(Corollary 4)}\\
&\leq \mathbb{E}\left[\mathcal{R}^{(1)}\right] + (\lambda K)^{1/3}T^{2/3} + 2 + \lambda + 4(\lambda K)^{1/3}T^{2/3}\\
&\leq 4\sqrt{KT} + 3 + \lambda + 5(\lambda K)^{1/3}T^{2/3} && \text{(Eq. (6))}\\
&\leq \mathcal{O}\left((\lambda K)^{1/3}T^{2/3}\right). && (\lambda \geq \sqrt{K/T})
\end{aligned}$$

We thus continue to the stochastically constrained case. For the proof we rely on the following two Lemmas:

Lemma 5. $\mathbb{E}\left[\overline{\mathcal{R}}^{(2)} + \lambda\mathcal{S}^{(2)}\right] \leq 4\lambda\,\mathbb{E}\left[\mathcal{S}^{(1)}\right]$.

Proof. The algorithm suggested satisfies Corollary 4, hence:

$$\begin{aligned}
\mathbb{E}\left[\overline{\mathcal{R}}^{(2)} + \lambda\mathcal{S}^{(2)}\middle|\mathcal{S}^{(1)}\right] &\leq \mathbb{E}\left[\mathcal{R}^{(2)} + \lambda\mathcal{S}^{(2)}\middle|\mathcal{S}^{(1)}\right]\\
&\leq \begin{cases} 4(\lambda K)^{1/3}T^{2/3} + 1, & \text{if } S^{(1)} \geq K^{1/3}(\frac{T}{\lambda})^{2/3} + \frac{1}{\lambda}\\ 0, & \text{o.w.}\end{cases} && \text{(Corollary 4)}\\
&\leq \begin{cases} 4(\lambda K)^{1/3}T^{2/3} + 4, & \text{if } \mathcal{S}^{(1)} \geq K^{1/3}(\frac{T}{\lambda})^{2/3} + \frac{1}{\lambda}\\ 4\lambda\mathcal{S}^{(1)}, & \text{o.w.}\end{cases}\\
&\leq 4\lambda\mathcal{S}^{(1)}. && (8)
\end{aligned}$$

Where we used that by the general regret definition we have that for any algorithm $\mathbb{E}\left[\overline{\mathcal{R}}_T\right] \leq \mathbb{E}\left[\mathcal{R}_T\right]$. Taking expectation on both sides of Eq. (8) concludes the proof. ■

Lemma 6.

$$\lambda\,\mathbb{E}[S^{(1)}] \leq \mathcal{O}\left(\min\left\{\left(\frac{\lambda}{\Delta_{min}} + 1\right)\mathbb{E}\left[\overline{\mathcal{R}}^{(1)}\right], (\lambda K)^{1/3}T^{2/3}\right\}\right).$$

Proof. Consider some arbitrary arm $i \in [K]$. When a switch occurs at round t, either I_{t-1} or I_t are different from i. Hence, using the union bound one can bound the expected number of switches $\mathbb{E}[\mathcal{S}]$, regardless to the environment regime (either adversarial or stochastically-constrained adversarial) as follows:

$$\mathbb{E}[\mathcal{S}] \leq 1 + 2\sum_{t\in[T]}\sum_{i\neq i^{\star}}\mathbb{E}[p_{t,i}].$$

Where we have used the fact that $\mathbb{P}(I_t = i) = \mathbb{E}[p_{t,i}]$. Additionally, in the stochastically-constrained regime we also have that:

$$\sum_{t\in[T]}\sum_{i\neq i^\star}\mathbb{E}[p_{t,i}] \leq \sum_{t\in[T]}\sum_{i\neq i^\star}\frac{\mathbb{E}[p_{t,i}]\Delta_i}{\Delta_{\min}} = \frac{\mathbb{E}[\overline{\mathcal{R}}^{(1)}]}{\Delta_{\min}}.$$

Utilizing the stopping criterion of "Switch Tsallis, Switch!," we can bound the expected number of switches $\mathbb{E}[\mathcal{S}^{(1)}]$, regardless to the environment regime (either adversarial or stochastically-constrained adversarial) and obtain:

$$\lambda\,\mathbb{E}[\mathcal{S}^{(1)}] \leq \min\left\{\lambda + 2\lambda\frac{\mathbb{E}[\overline{\mathcal{R}}^{(1)}]}{\Delta_{\min}}, \lambda + (\lambda K)^{1/3}T^{2/3} + 1\right\}.$$

■

The total expected pseudo-regret is upper bounded by the summation of the pseudo-regret attained for each part of the algorithm. We will consider two cases.

$$\begin{aligned}
\mathbb{E}\left[\overline{\mathcal{R}}_T^{\lambda}\right] &\leq \mathbb{E}\left[\overline{\mathcal{R}}^{(1)} + \lambda\mathcal{S}^{(1)} + \overline{\mathcal{R}}^{(2)} + \lambda\mathcal{S}^{(2)}\right] \\
&\leq \mathbb{E}\left[\overline{\mathcal{R}}^{(1)}\right] + 5\lambda\,\mathbb{E}\left[\mathcal{S}^{(1)}\right] && \text{(Lemma 5)} \\
&\leq \mathbb{E}\left[\overline{\mathcal{R}}^{(1)}\right] + \min\left\{5\lambda + \frac{10\lambda}{\Delta_{\min}}\,\mathbb{E}\left[\overline{\mathcal{R}}^{(1)}\right], 5\lambda + 5(\lambda K)^{1/3}T^{2/3} + 5\right\} && \text{(Lemma 6)} \\
&\leq \min\left\{5\lambda + \left(\frac{10\lambda}{\Delta_{\min}} + 1\right)\mathbb{E}\left[\overline{\mathcal{R}}^{(1)}\right], 5\lambda + 5(\lambda K)^{1/3}T^{2/3} + 5 + \mathbb{E}\left[\overline{\mathcal{R}}^{(1)}\right]\right\} \\
&\leq \mathcal{O}\left(\min\left\{\left(\frac{\lambda}{\Delta_{\min}} + 1\right)\mathbb{E}\left[\overline{\mathcal{R}}^{(1)}\right], (\lambda K)^{1/3}T^{2/3} + \mathbb{E}\left[\overline{\mathcal{R}}^{(1)}\right]\right\}\right) \\
&\leq \mathcal{O}\left(\min\left\{\left(\frac{\lambda}{\Delta_{\min}} + 1\right)\mathcal{O}\left(\sum_{i\neq i^\star}\frac{\log T}{\Delta_i}\right), (\lambda K)^{1/3}T^{2/3} + \mathbb{E}\left[\mathcal{R}^{(1)}\right]\right\}\right) && \text{(Eq. (7))} \\
&\leq \mathcal{O}\left(\min\left\{\left(\frac{\lambda}{\Delta_{\min}} + 1\right)\mathcal{O}\left(\sum_{i\neq i^\star}\frac{\log T}{\Delta_i}\right), (\lambda K)^{1/3}T^{2/3} + \sqrt{KT}\right\}\right) && \text{(Eq. (6))}
\end{aligned}$$

5.3 Proof of Theorem 2

Our lower bound builds upon the work of Dekel et al. [8] and we adapt it to the stochastically-constrained adversarial regime. Dekel et al. [8] suggested the following process, depicted in Algorithm 2, to generate an adversarial loss sequence. With correct choice of parameter $\Delta = O(T^{-1/3})$, the process ensures for any deterministic player, a regret of order $\mathbb{E}[\mathcal{R}_T] = \tilde{\Omega}(K^{1/3}T^{2/3})$. For our purposes we need to take care of two things: First we need to generalize the bound to arbitrary Δ. Second, one can see that the loss sequence generated by the adversary in Algorithm 2, is not stochastically constrained (as defined in Section 2). To cope with this, we develop a more fine-grained analysis over a modified loss sequence that assures the stochastically-constrained assumption is met. Towards proving Theorem 2, we present the following Theorem.

Theorem 7. *Let $\{\ell_1, \ldots, \ell_T\}$ be the stochastic sequence of loss functions defined in Algorithm 2. Then for any deterministic player against this sequence it holds*

$$\mathbb{E}[\mathcal{R}_T + \mathcal{S}_T] \geq \min\left\{1/(40^2\Delta^2\log_2^3 T), \Delta T/24\right\}.$$

We deter the proof of Theorem 7 to the end of this section, and we now proceed with proving our desired lower bound in Theorem 2.

Proof of Theorem 2. It can be observed, that the process depicted in Algorithm 2 is almost stochastically constrained, in fact, if at t and i we do not perform clipping to $[0, 1]$, i.e. $\ell_t(i) = \ell'_t(i)$ the sequence is indeed stochastically constrained. Let us define then an event H as follows,

$$H = \left\{\forall t \in [T] :\ X_t + \tfrac{1}{2} \in \left[\tfrac{1}{6}, \tfrac{5}{6}\right]\right\}.$$

Algorithm 2 The adversary's loss generation process proposed by Dekel et al. [8].

Input: time horizon T, the minimal gap Δ.
Output: loss sequence $\left\{\ell_t \in [0,1]^K\right\}_{t\in[T]}$.

1: Set $\sigma = (9\log_2 T)^{-1}$.
2: Draw T independent Gaussian variables - $\{n_t \sim \mathcal{N}(0,\sigma^2)\}_{t\in[T]}$.
3: Define the process $\{X_t\}_{t\in[T]}$ by

$$\forall t \in [T] : X_t = X_{r(t)} + n_t,$$

where $X_0 = 0$, $r(t) = t - 2^{m(t)}$, and $m(t) = \max\{i \geq 0 : 2^i \text{ divides } t\}$.
4: Choose $i^\star \in [2]$ uniformly at random.
5: For all $t \in [T]$ and $i \in [2]$, set

$$\ell'_t(i) = X_t + \tfrac{1}{2} - \Delta \cdot \mathbb{1}\{i^\star = i\}$$

If $0 \leq \ell'_t(i) \leq 1$, set $\ell_t(i) = \ell'_t(i)$, else perform clipping:

$$\ell_t(i) = \min\{\max\{\ell'_t(i), 0\}, 1\}$$

If we restrict $\Delta \leq \frac{1}{6}$, the event H implies that $\ell_t = \ell'_t$ for all $t \in [T]$. By standard Gaussian arguments one can derive the following Lemma.

Lemma 8 ([8, Lemma 1])**.** *For any $\delta \in (0,1)$ it holds*

$$\mathbb{P}\left(\left\{\forall t \in [T] : |X_t| \leq 2\sigma\sqrt{\log_2 T \log(T/\delta)}\right\}\right) \geq 1 - \delta.$$

Setting $\delta = 1/T$ and $\sigma = 1/(9\log_2 T)$ we get,

$$\mathbb{P}(H) = \mathbb{P}\left(\left\{\forall t \in [T] : |X_t| \leq \tfrac{1}{3}\right\}\right) \geq 1 - 1/T.$$

This implies,

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_T + \mathcal{S}_T] &\leq \mathbb{E}[\mathcal{R}_T + \mathcal{S}_T | H^c] \cdot \tfrac{1}{T} + \mathbb{E}[\mathcal{R}_T + \mathcal{S}_T | H] && (\mathcal{R}_T + \mathcal{S}_T \geq 0) \\
&\leq \Delta + 1 + \mathbb{E}[\mathcal{R}_T + \mathcal{S}_T | H] && (\mathcal{R}_T + \mathcal{S}_T \leq (\Delta+1)T)
\end{aligned}$$

Taking this together with Theorem 7 we get that for any deterministic player,

$$\mathbb{E}[\mathcal{R}_T + \mathcal{S}_T | H] = \Omega\left(\min\left\{1/(\Delta^2 \log_2^3 T), \Delta T\right\}\right). \tag{9}$$

Here we used the fact that for any $\Delta > \frac{1}{6}$ there exists a trivial lower bound of 1. In other words, if we let the loss sequence to be the conditional process given that H is fulfilled, we obtain a stochastic process that generates a random sequence that satisfies the conditions of a stochastically-constrained loss sequence, i.e. Eq. (1) is met. Eq. (9), in turn, bounds the expected regret given that our loss sequence is drawn from the above process.

Next, since any randomized player can be implemented by a random combination of deterministic players we conclude that Eq. (9) holds for any randomized player. This then immediately implies that there exists a deterministic loss sequence $\{\ell_1, \ldots, \ell_T\}$, that is stochastically-constrained, for which $\mathcal{R}_T + \mathcal{S}_T$ is lower bounded by the RHS of Eq. (9). Clearly, a simple derivation shows that the lower bound holds for any $K \geq 2$, as we can always extend the loss sequence for $K > 2$ by setting $\ell_{t,i} = 1$ for any $i > 2$. Lastly, we argue that $\mathbb{E}[\mathcal{R}_T] = \mathbb{E}[\overline{\mathcal{R}}_T]$. This is a direct implication of the loss sequence construction, as for all t we have $i^\star = \arg\min_{i\in[K]} \ell_{t,i}$. Therefore, the acclaimed bounds are achieved also with respect to the pseudo-regret, which concludes the proof. ■

Proof of Theorem 7. The proof of Theorem 7 requires additional notations. Let the conditional probability measures for all $i \in [K]$ be

$$\mathcal{Q}_i(\cdot) = \mathbb{P}(\cdot \mid i^\star = i),$$

and let $\mathcal{F}$ be the σ-algebra generated by the player's observations $\{\ell_{t,I_t}\}_{t\in[T]}$. Denote the *total variation* distance between $\mathcal{Q}_i$ and $\mathcal{Q}_j$ on $\mathcal{F}$ by

$$d_{\mathrm{TV}}^{\mathcal{F}}(\mathcal{Q}_i, \mathcal{Q}_j) = \sup_{E\in\mathcal{F}} \left|\mathcal{Q}_i(E) - \mathcal{Q}_j(E)\right|.$$

We also denote $\mathbb{E}_{\mathcal{Q}_i}$ as the expectation on the conditional distribution Q_i. Lastly, we present the following result from Dekel et al. [8].

Lemma 9 ([8, Lemma 3]). *For $K = 2$ and $T \geq 4$ it holds that*

$$d_{TV}^{\mathcal{F}}(\mathfrak{Q}_1, \mathfrak{Q}_2) \leq (\Delta/\sigma)\sqrt{2\log_2 T\, \mathbb{E}[\mathcal{S}_T]}.$$

Observe that $\mathcal{R}_T \geq 0$ by the construction in Algorithm 2. Then, if $\mathbb{E}[\mathcal{S}_T] \geq 1/(c\Delta^2 \log_2^3 T)$ for $c = 40^2$ we have that $\mathbb{E}[\mathcal{R}_T + \mathcal{S}_T] \geq 1/(c\Delta^2 \log_2^3 T)$, which guarantees the desired lower bound. On the other hand, applying Lemma 9 when $\mathbb{E}[\mathcal{S}_T] \leq 1/(c\Delta^2 \log_2^3 T)$, we get

$$d_{\mathrm{TV}}^{\mathcal{F}}(\mathfrak{Q}_1, \mathfrak{Q}_2) \leq (1/\sigma)\sqrt{2/(c\log_2^2 T)} \leq \tfrac{1}{3}. \tag{10}$$

Let E be the event that arm $i = 1$ is picked at least $T/2$ times, namely

$$E = \left\{ \sum_{t \in [T]} \mathbb{1}\{I_t = 1\} \geq T/2 \right\},$$

and let E^c be its complementary event. If $\mathfrak{Q}_1(E) \leq \frac{1}{2}$ then,

$$\begin{aligned} \mathbb{E}[\mathcal{R}_T] &\geq \mathbb{E}_{\mathfrak{Q}_1}[\mathcal{R}_T | E^c] \cdot \mathfrak{Q}_1(E^c) \cdot \mathbb{P}(i^\star = 1) && (\mathcal{R}_T \geq 0) \\ &\geq \Delta T/8. && (\mathcal{R}_T \geq \Delta T/2 \text{ under the conditional event}) \end{aligned}$$

If $\mathfrak{Q}_1(E) > \frac{1}{2}$ then from Eq. (10) we obtain that $\mathfrak{Q}_2(E) \geq \frac{1}{6}$. This implies,

$$\begin{aligned} \mathbb{E}[\mathcal{R}_T] &\geq \mathbb{E}_{\mathfrak{Q}_2}[\mathcal{R}_T | E] \cdot \mathfrak{Q}_2(E) \cdot \mathbb{P}(i^\star = 2) && (\mathcal{R}_T \geq 0) \\ &\geq \Delta T/24. && (\mathcal{R}_T \geq \Delta T/2 \text{ under the conditional event}) \end{aligned}$$

Since $\mathcal{S}_T \geq 0$ we conclude the proof. ■

6 Discussion

Best-of-both-worlds algorithm is an extremely challenging setting and in particular, the case of regret with switching cost poses interesting challenges. We presented here an algorithm that achieves (up to logarithmic factors) optimal minimax rates for the case of two arms, i.e. $K = 2$. Surprisingly, the result is obtained using a very simple modification of the standarad best-of-both-worlds Tsallis-INF algorithm. Several open problems though seem to arise from our work.

Open Question 1. *Given arbitrary K, what is the optimal minimax regret rate, in the stochastically constrained setting, of any algorithm that achieves, in the adversarial regime, regret of*

$$\mathbb{E}[\mathcal{R}_T^\lambda] = \tilde{\mathcal{O}}\left((\lambda K)^{1/3} T^{2/3}\right).$$

In particular, in our result, it is interesting to find out if the term $\mathcal{O}\left(\sum_{i \neq i^\star} \frac{\lambda \log T}{\Delta_{\min}\Delta_i}\right)$ can be replaced by $\mathcal{O}\left(\sum \frac{\lambda \log T}{\Delta_i^2}\right)$. Note that this term is obtained by a worst-case analysis of the switching cost that assumes that we obtained the regret by only switching from the optimal arm to the consecutive second-to-best arm. It seems more likely that any reasonable algorithm roughly switches to each arm i, order of $\tilde{\mathcal{O}}(1/\Delta_i^2)$ times, leading to more optimistic rate.

Another natural open problem is to try and generalize the lower bound to the general case. For simplicity we state the next problem for the case all arms have the same gap

Open Question 2. *Suppose $\Delta_1 = \Delta_2 =, \ldots, = \Delta_{min}$, and $\Delta_{min} \geq (\lambda K)^{1/3} T^{-2/3}$.*

Is it possible to construct an algorithm that achives regret, in the adversarial regime of

$$\mathbb{E}[\mathcal{R}_T^\lambda] = \tilde{\mathcal{O}}\left((\lambda K)^{1/3} T^{2/3}\right).$$

and in the stochastically constrained case:

$$\mathbb{E}[\mathcal{R}_T^\lambda] = o\left(\frac{\lambda K \log T}{\Delta_{min}^2}\right).$$

Finally, we would like to stress that our lower bound applies to a *stochastically constrained* setting, where in principle we often care to understand the stochastic case:

Open Question 3. *What is the optimal expected pseudo-regret with switching cost that can be achieved by an algorithm that achieves regret, in the adversarial regime of*

$$\mathbb{E}[\mathcal{R}_T^\lambda] = \tilde{\mathcal{O}}\left((\lambda K)^{1/3} T^{2/3}\right),$$

against an i.i.d sequence $\ell_1, \ldots, \ell_T$ that satisfies Eq. (1)?

Achieving a non-trivial lower bound for the above case seems like a very challenging task. In particular, it is known that, if we don't attempt to achieve best-of-both-worlds rate then an upper bound of $O(\sum_{i \neq i^\star} \frac{\log T}{\Delta_i})$ is achievable [9, 10]. Interestingly, then, our lower bound at Theorem 2 presents a separation between the stochastic and stochastically constrained case, leaving open the possibility that a best-of-both-worlds algorithm between adversarial and stochastically constrained case is possible but note necessarily against stochastic player. Proving the reverse may require new algorithmic techniques. In particular, the current analysis of Tsallis-INF is valid for the stochastically-constrained case as much as to the stochastic case. An improved upper bound for the pure stochastic case, though, cannot improve over the stochastically constrained case as demonstrated by Theorem 2.

References

[1] J. D. Abernethy, C. Lee, and A. Tewari. Fighting bandits with a new kind of smoothness. *Advances in Neural Information Processing Systems*, 28, 2015.

[2] R. Arora, O. Dekel, and A. Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400*, 2012.

[3] J.-Y. Audibert, S. Bubeck, et al. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

[4] P. Auer and C.-K. Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120. PMLR, 2016.

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

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

[7] S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1. JMLR Workshop and Conference Proceedings, 2012.

[8] O. Dekel, J. Ding, T. Koren, and Y. Peres. Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 459–467, 2014.

[9] H. Esfandiari, A. Karbasi, A. Mehrabian, and V. Mirrokni. Regret bounds for batched bandits. *arXiv preprint arXiv:1910.04959*, 2019.

[10] Z. Gao, Y. Han, Z. Ren, and Z. Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] T. L. Lai, H. Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[12] C. Rouyer, Y. Seldin, and N. Cesa-Bianchi. An algorithm for stochastic and adversarial bandits with switching costs. In *International Conference on Machine Learning*, pages 9127–9135. PMLR, 2021.

[13] Y. Seldin and G. Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 1743–1759. PMLR, 2017.

[14] Y. Seldin and A. Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295. PMLR, 2014.

[15] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[16] C.-Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

[17] J. Zimmert and Y. Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *J. Mach. Learn. Res.*, 22:28:1–28:49, 2021. URL http://jmlr.org/papers/v22/19-753.html.

[18] J. Zimmert, H. Luo, and C.-Y. Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692. PMLR, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]