An Algorithm for Stochastic and Adversarial Bandits with Switching Costs

Chloé Rouyer [1] **Yevgeny Seldin** [1] **Nicolò Cesa-Bianchi** [2]

Abstract

We propose an algorithm for stochastic and adversarial multiarmed bandits with switching costs, where the algorithm pays a price λ every time it switches the arm being played. Our algorithm is based on adaptation of the Tsallis-INF algorithm of Zimmert & Seldin (2021) and requires no prior knowledge of the regime or time horizon. In the oblivious adversarial setting it achieves the minimax optimal regret bound of $\mathcal{O}\big((\lambda K)^{1/3}T^{2/3} + \sqrt{KT}\big)$, where T is the time horizon and K is the number of arms. In the stochastically constrained adversarial regime, which includes the stochastic regime as a special case, it achieves a regret bound of $\mathcal{O}\left(\big((\lambda K)^{2/3}T^{1/3} + \ln T\big)\sum_{i \neq i^*} \Delta_i^{-1}\right)$, where Δ_i are the suboptimality gaps and i^* is a unique optimal arm. In the special case of $\lambda = 0$ (no switching costs), this bound is also minimax optimal within constants. We also explore variants of the problem, where switching cost is allowed to change over time. We provide experimental evaluation showing competitiveness of our algorithm with the relevant baselines in the stochastic, stochastically constrained adversarial, and adversarial regimes with fixed switching cost.

1. Introduction

Multiarmed bandits are the reference framework for the study of a wide range of sequential decision-making problems, including recommendation, dynamic content optimization, digital auctions, clinical trials, and more. In this framework the algorithm repeatedly picks actions, a.k.a. arms, and, after each selection, observes the loss or reward of the corresponding action. In many application domains, algorithms have to pay a penalty $\lambda > 0$ each time they play an arm different from the one played in the previous round. Such switching cost may occur in the form of a transaction cost in financial trading, or a reconfiguration cost in industrial environments.

So far, the problem of bandits with switching costs has been studied using algorithms whose optimality depends on the nature of the source of losses (or, equivalently, rewards) for the K arms. In the oblivious adversarial case, when losses are generated by an arbitrary deterministic source, Dekel et al. (2012) used a simple variant of the Exp3 algorithm to prove an upper bound of $\mathcal{O}\big((K\ln K)^{1/3}T^{2/3}\big)$ for $\lambda = 1$ (i.e., unit switching cost) — see also (Blum & Mansour, 2007) for an earlier, slightly weaker result. A result by Dekel et al. (2013) implies a lower bound of $\Omega\big((\lambda K)^{1/3}T^{2/3} + \sqrt{KT}\big)$ for all $\lambda \geq 0$. Note the phase transition: if $\lambda > 0$, then the regret asymptotically grows as $T^{2/3}$, as opposed to $\sqrt{T}$ when there is no switching cost.

In the stochastic case, where losses of each arm are generated by an i.i.d. process, Gao et al. (2019) and Esfandiari et al. (2021) used arm elimination algorithms to prove that $\mathcal{O}(\ln T)$ switches are sufficient to achieve the optimal distribution-dependent regret of $\mathcal{O}\big((\ln T)\sum_{i:\Delta_i>0}\Delta_i^{-1}\big)$, where Δ_i is the suboptimality gap of arm i. Hence, in the stochastic case the introduction of switching costs does not lead to a qualitative change of the minimax regret rate.

In practical applications, it is desirable to have algorithms that require no prior knowledge about the nature of the loss generation process and maintain robustness in the adversarial regime simultaneously with the ability to achieve lower regret in the stochastic case. A number of such algorithms have been developed for the standard multiarmed bandits (Bubeck & Slivkins, 2012; Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017; Wei & Luo, 2018; Zimmert & Seldin, 2019; 2021; Masoudian & Seldin, 2021) and the ideas have been extended to several other domains, including combinatorial bandits (Zimmert et al., 2019), decoupled exploration and exploitation (Rouyer & Seldin, 2020), and episodic MDPs (Jin & Luo, 2020). We aim at designing algorithms with similar properties for bandits with switching costs.

[1] Department of Computer Science, University of Copenhagen, Denmark [2] DSRC & Dept. of Computer Science, Università degli Studi di Milano, Milano, Italy. Correspondence to: Chloé Rouyer <chloe@di.ku.dk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Main contributions

Our starting point is the Tsallis-INF algorithm of Zimmert & Seldin (2021), which was shown to achieve minimax regret rates in both stochastic and adversarial regimes for standard bandits. We introduce a modification of this algorithm, which we call Tsallis-Switch, to take care of the switching costs. In the adversarial regime, the regret bound of Tsallis-Switch matches (within constants) the minimax optimal regret bound $\Theta\big((\lambda K)^{1/3}T^{2/3} + \sqrt{KT}\big)$ for any value of $\lambda \geq 0$. In the stochastically constrained adversarial regime, which includes the stochastic regime as a special case, we prove a bound $\mathcal{O}\left(\left((\lambda K)^{2/3}T^{1/3} + \ln T\right) \sum_{i \neq i^*} \Delta_i^{-1}\right)$, where i^* is a unique optimal arm. Note that, in the special case of $\lambda = 0$ (no switching costs), we recover (up to constant factors) the minimax optimal bounds of Tsallis-INF for both regimes. Similarly to Tsallis-INF, our algorithm is fully oblivious to both the regime and the time horizon T.

Tsallis-Switch, which runs Tsallis-INF as a subroutine, uses the standard tool to control the frequency of arm switching: game rounds are grouped into consecutive blocks $B_1, B_2, \ldots$, and Tsallis-Switch runs Tsallis-INF over the blocks, preventing it from switching arms within each block. The number of switches is thus bounded by the number of blocks. Since T is unknown, we use block sizes of increasing length. As a new arm is drawn only at the beginning of each block, the effective range of the losses experienced by Tsallis-INF grows with time. Therefore, we modify the analysis of Tsallis-INF to accommodate losses of varying range. This extension may potentially be of independent interest.

2. Problem Setting and Notations

We consider a repeated game with K arms and a switching cost $\lambda \geq 0$. At each round $t = 1, 2, \ldots$ of the game, the environment picks a loss vector $\ell_t \in [0,1]^K$, and the algorithm chooses an arm $J_t \in [K]$ to play. The learner then incurs the loss ℓ_{t,J_t}, which is observed. If $J_t \neq J_{t-1}$, then the learner also suffers an extra penalty of λ. The penalty λ is known to the learner. We use the same setting as Dekel et al. (2013), and assume that $J_0 = 0$, which means that there is always a switch at the first round.

We consider two regimes for the losses. In the oblivious adversarial regime, the loss vectors ℓ_t are arbitrarily generated by the environment and do not depend on the actions taken by the learner. We also work in the stochastically constrained adversarial regime. This setting, introduced by Wei & Luo (2018), generalizes the widely studied stochastic regime by allowing losses to be drawn from distributions with fixed gaps. It means that at for all i, $\mathbb{E}\left[\ell_{t,i}\right]$ can fluctuate with t, but $\mathbb{E}\left[\ell_{t,i} - \ell_{t,j}\right] = \Delta_{i,j}$ remains constant over time for all pairs i, j,. The suboptimality gaps are then defined as $\Delta_i = \Delta_{i,1} - \min_j \Delta_{j,1}$.

We define the pseudo-regret with switching costs as follows,

$$\begin{aligned}\mathrm{RS}(T,\lambda) &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,J_t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right] \\ &\quad + \lambda \sum_{t=1}^{T} \mathbb{P}(J_{t-1} \neq J_t) \\ &= R_T + \lambda\, S_T. \qquad (1)\end{aligned}$$

We recognize that $R_T = \mathrm{RS}(T,0)$ is the classical definition of the pseudo regret (without switching costs), while S_T counts the expected number of switches. Furthermore, we recall that in the stochastically constrained adversarial regime, the pseudo-regret can be rewritten in terms of the sub-optimality gaps, as:

$$\mathrm{R}_T = \sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}\left[p_{t,i}\right] \Delta_i, \qquad (2)$$

where $p_{t,i}$ is the probability of playing action i at round t.

3. Using Blocks to Control Switching Frequency

In order to control S_T, we limit the number of action switches that the algorithm makes by dividing the game rounds into blocks and forcing the algorithm to play the same action for all the rounds within a block. Given a sequence of blocks $(B_n)_{n\geq 1}$ of lengths $|B_n|$, and a time horizon T, we define N as the smallest integer, such that $\sum_{n=1}^{N} |B_n| \geq T$, and we truncate the last block, such that the cumulative length of the first N blocks sum up to T.

As $S_T \leq N$, we bound N and the pseudo-regret R_T (without the switching costs) over the N blocks. Let $c_{n,i} = \sum_{s \in B_n} \ell_{s,i}$ be the cumulative loss of action i in block n. Since $\ell_{t,i} \in [0,1]$, we have $c_{n,i} \in \big[0, |B_n|\big]$. We use I_n to refer to the action played by the algorithm in block n. Then, for all $t \in B_n$, we have $J_t = I_n$ and

$$\mathrm{R}_T = \mathbb{E}\left[\sum_{n=1}^{N} c_{n,I_n}\right] - \min_j \mathbb{E}\left[\sum_{n=1}^{N} c_{n,j}\right].$$

4. The Algorithm

Our Tsallis-Switch algorithm (see Algorithm 1) calls Tsallis-INF at the beginning of each block to obtain an action, plays the proposed action in each round within the block, and then feeds back to Tsallis-INF the total loss suffered by the action over the block. As blocks have varying lengths, we adapt the Tsallis-INF algorithm and its analysis to losses of varying range.

Algorithm 1 Tsallis-Switch

```
    **Input:** Learning rates $\eta_1 \geq \eta_2 \geq \cdots > 0$.
    Block lengths $|B_1|, |B_2|, \ldots$.
    **Initialize:** $\tilde{C}_0 = \mathbf{0}_K$
    **for** $n = 1, 2, \ldots$ **do**
        $p_n = \arg\min_{p \in \Delta^{K-1}} \left\{ \langle p, \tilde{C}_{n-1} \rangle - \sum_{i=1}^{K} \frac{4\sqrt{p_i} - 2p_i}{\eta_n} \right\}$.
        Sample $I_n \sim p_n$ and play it for all rounds $t \in B_n$.
        Observe and suffer $c_{n,I_n} = \sum_{t \in B_n} \ell_{t,I_n}$.
        $\forall i \in [K] : \tilde{c}_{n,i} = \begin{cases} \frac{c_{n,i}}{p_{n,i}}, & \text{if } I_n = i, \\ 0, & \text{otherwise.} \end{cases}$
        $\forall\, i \in [K] : \quad \tilde{C}_n(i) = \tilde{C}_{n-1}(i) + \tilde{c}_{n,i}$.
    **end for**
```

5. Main Results

We start by considering the case where the switching cost λ is a fixed parameter given to the algorithm. Since λ is known in advance, it can be used to tune the block lengths.

Theorem 1. *Let $\lambda \geq 0$ be the switching cost. Define blocks with lengths $|B_n| = \max\{\lceil a_n \rceil, 1\}$, where $a_n = \frac{3\lambda}{2}\sqrt{\frac{n}{K}}$. The preudo-regret of* Tsallis-Switch *with learning rate $\eta_n = \frac{2}{a_n+1}\sqrt{\frac{2}{n}}$ executed over the blocks in any adversarial environment satisfies:*

$$R(T,\lambda) \leq 5.25(\lambda K)^{1/3} T^{2/3} + 6.4\sqrt{KT} + 3\sqrt{2K} + 5.25\lambda + 6.25.$$

Furthermore, in any stochastically constrained adversarial regime with a unique best arm i^*, the pseudo-regret additionally satisfies:

$$R(T,\lambda) \leq \left(66(\lambda K)^{2/3} T^{1/3} + 32 \ln T\right) \sum_{i \neq i^*} \frac{1}{\Delta_i} + \left(160\lambda^{2/3} T^{1/3} K^{1/6} + 160\lambda + 49\lambda^2 + 32\right) \sum_{i \neq i^*} \frac{1}{\Delta_i} + \frac{544\lambda}{\sqrt{K}} + \lambda + 66.$$

A proof is provided in Section 6. For $\lambda = 0$ (no switching costs) both regret bounds match within constants the corresponding bounds of Tsallis-INF for multiarmed bandits with no switching costs. Furthermore, in the adversarial regime the algorithm achieves the optimal regret rate for all values of λ. In the stochastically constrained adversarial regime, for $\lambda > 0$ the regret grows as $T^{1/3}$ rather than logarithmically in T. This is also the case for the stochastic regime, which is a special case. While the algorithm does not achieve the logarithmic regret rate in the stochastic regime, as do the algorithms of Gao et al. (2019) and Esfandiari et al. (2021), it still exploits the simplicity of the regime and reduces the regret rate from $T^{2/3}$ to $T^{1/3}$. Additionally, in contrast to the work of Gao et al. (2019) and Esfandiari et al. (2021), the stochastic regret guarantee holds simultaneously with the adversarial regret guarantee, and the algorithm requires no knowledge of the time horizon. We also note that we are unaware of specialized lower bounds for the more general stochastically constrained adversarial regime with switching costs, and it is unknown whether the corresponding regret guarantee is minimax optimal. Theorem 1 is based on the following generalized analysis of the Tsallis-INF algorithm that accommodates losses of varying range. The result may be of independent interest.

Theorem 2. *Consider a multi-armed bandit problem where the loss vector at round t belongs to $[0, b_t]^K$ and b_t is revealed to the algorithm before round t. Then the pseudo-regret of* Tsallis-Switch *in any adversarial environment for any positive and non-decreasing sequence of learning rates $(\eta_t)_{t \geq 1}$ satisfies*

$$R_T \leq \sqrt{K} \left(\sum_{t=1}^{T} \frac{\eta_t}{2} b_t^2 + \frac{4}{\eta_T} \right) + 1. \tag{3}$$

Furthermore, in the stochastically constrained adversarial regime with a unique best arm i^*, the pseudo regret also satisfies

$$R_T \leq \sum_{t=1}^{T} \sum_{i \neq i^*} \frac{\left(\frac{7}{2}\eta_t b_t^2 + 2c\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right)\right)^2}{4\Delta_i b_t} + \sum_{t=1}^{T_0} \eta_t b_t^2 + 2, \tag{4}$$

$$\textit{where } c = \begin{cases} 2, & \textit{if } \forall t : \ \frac{5\eta_t}{4} b_t^2 \geq 2\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right), \\ 4, & \textit{otherwise}. \end{cases}$$

In particular, if $b_t = B$ for all rounds t, we have the following more interpretable result.

Corollary 3. *Consider a multi-armed bandit problem with loss vectors belonging to $[0, B]^K$. Then the pseudo-regret of Tsallis-INF with $\eta_t = \frac{2}{B\sqrt{t}}$ satisfies $R_T \leq 4B\sqrt{KT} + 1$ in any adversarial regime. Furthermore, in the stochastically constrained adversarial regime with a unique best arm i^*, the pseudo regret additionally satisfies*

$$R_T \leq 21B(\ln T + 1) \sum_{i \neq i^*} \frac{1}{\Delta_i} + 8\sqrt{B} + 2.$$

5.1. Varying Switching Cost

Now we consider a setting, where the switching cost may change after each switch. The learner is given the n-th switching cost λ_n right after the $n-1$-th switch is taken, and we allow the length of the block $|B_n|$ to depend on it. In this setting, the cumulative expected switching cost becomes

$$S\left(T, (\lambda_n)_{n\geq 1}\right) = \sum_{n=1}^{N} \lambda_n \mathbb{P}(I_n \neq I_{n-1}),$$

where, as before, N is the smallest number of blocks to cover T rounds. We construct blocks, such that the terms R_T and $S\left(T, (\lambda_n)_{n\geq 1}\right)$ remain balanced.

Theorem 4. *Let $(\lambda_n)_{n\geq 1}$ be a sequence of non-negative switching costs. The pseudo-regret with switching costs of* Tsallis-Switch *executed with block lengths $|B_n| = \max\left\{\left\lceil \frac{\sqrt{\lambda_n}\sqrt{\sum_{s=1}^{n}\lambda_s + \frac{\sqrt{K}}{\sqrt{s}}}}{\sqrt{K}}\right\rceil, 1\right\}$ and $\eta_n = \frac{2\sqrt{2K}}{3a_n}$, where $a_n = \left(\sum_{s=1}^{n}\lambda_s + \sqrt{K/s}\right)$, satisfies:*

$$R(T, \lambda) \leq \sum_{n=1}^{N} 7\lambda_n + 12\sqrt{KN} + 2, \tag{5}$$

where N is the smallest integer such that $\sum_{n=1}^{N}|B_n| \geq T$. Furthermore, in the stochastically constrained adversarial regime with a unique best arm i^*, the pseudo regret additionally satisfies

$$R\left(T, (\lambda_n)_{n\geq 1}\right) \leq \sum_{n=1}^{N}\sum_{i\neq i^*} \frac{\left(11\lambda_n + \lambda_{n+1} + \frac{10\sqrt{2}}{\sqrt{n}}\right)^2}{4\Delta_i |B_n|} + \sum_{n=1}^{N_0}\left(\frac{2\sqrt{2}\lambda_n}{\sqrt{K}}\right) + 4\sqrt{2N_0} + \lambda_1 + 2,$$

where N_0 is the smallest $n \leq N$ such that for all $n \geq N_0$, $\eta_n |B_n| \leq \frac{1}{4}$. If such an integer does not exist, then $N_0 = N$.

A proof is provided in Appendix D. Note that for $\lambda_n = \lambda$, the bound (5) for the adversarial setting is of the same order as the corresponding bound in Theorem 1.

If λ_n is not monotone, then controlling the first term in the regret bound for the stochastically constrained adversarial regime is challenging, because the block length $|B_n|$ in the denominator does not depend on λ_{n+1} in the numerator. Below, we provide a specialization of the regret bound assuming that the switching costs increase as $\lambda_n = n^\alpha$ for some $\alpha > 0$.

Corollary 5. *Assume that for $n \geq 1$, $\lambda_n = n^\alpha$ for some $\alpha > 0$. Then the regret bound for the stochastically constrained adversarial regime with a unique best arm i^* in Theorem 4 satisfies*

$$R\left(T, (\lambda_n)_{n\geq 1}\right) \leq \mathcal{O}\left(\sum_{i\neq i^*} \frac{K^{\frac{2\alpha+2}{2\alpha+3}} T^{\frac{2\alpha+1}{2\alpha+3}} + K^{\frac{2\alpha}{2\alpha+3}} T^{\frac{4\alpha}{2\alpha+3}}}{\Delta_i}\right).$$

A proof is provided in Appendix D. At the limit $\alpha \to 0$, the bound scales as $\mathcal{O}\big(K^{2/3}T^{1/3}\sum_{i\neq i^*}\frac{1}{\Delta_i}\big)$, which matches the pseudo-regret bound in the stochastically constrained adversarial regime in Theorem 1 with $\lambda = 1$. Note also that the bound remains sublinear in T, as long as $\alpha < \frac{3}{2}$. In other words, with a switching cost as high as $\lambda_n = n^{3/2-\varepsilon}$, for any $\varepsilon > 0$, Tsallis-Switch still has sublinear regret.

6. Proofs

We start by introducing some preliminary definitions and results. Recall that the pseudo-regret can be decomposed into a sum of stability and penalty terms (Lattimore & Svepesvári, 2020; Zimmert & Seldin, 2021). Let Φ_n be defined as:

$$\Phi_n(C) = \max_{p\in\Delta^{K-1}}\left\{\langle p, C\rangle + \sum_i \frac{4\sqrt{p_i} - 2p_i}{\eta_n}\right\}.$$

Note that the distribution p_n used by Tsallis-Switch to draw action I_n for block B_n satisfies $p_n = \nabla\Phi_n(-\tilde{C}_{n-1})$. We can write:

$$\begin{aligned}\mathbb{E}\left[\sum_{n=1}^{N} c_{n,I_n}\right] - \min_j \mathbb{E}\left[\sum_{n=1}^{N} c_{n,j}\right] &= \underbrace{\mathbb{E}\left[\sum_{n=1}^{N} c_{n,I_n} + \Phi_n(-\tilde{C}_n) - \Phi_n(-\tilde{C}_{n-1})\right]}_{\text{stability}} \\ &\quad + \underbrace{\mathbb{E}\left[\sum_{n=1}^{N} \Phi_n(-\tilde{C}_{n-1}) - \Phi_n(-\tilde{C}_n) - c_{n,i^*_N}\right]}_{\text{penalty}},\end{aligned} \tag{6}$$

where i^*_N is any arm with smallest cumulative loss over the N blocks (i.e., a best arm in hindsight).

We start by introducing bounds on the stability and the penalty parts of the regret. The results generalize the corresponding results of Zimmert & Seldin (2021) to handle losses that take values in varying ranges and may be larger than 1. The proofs are provided in Appendix B. Note the multiplicative factor b_n^2 in the stability term.

Lemma 6. *For any sequence of positive learning rates $(\eta_n)_{n\geq 1}$ and any sequence of bounds $(b_n)_{n\geq 1}$ on the losses at round n, the* stability *term of the regret bound of* Tsallis-Switch *satisfies:*

$$\mathbb{E}\left[\sum_{n=1}^{N} c_{n,I_n} + \Phi_n(-\tilde{C}_n) - \Phi_n(-\tilde{C}_{n-1})\right] \leq \sum_{n=1}^{N} \frac{\eta_n}{2} b_n^2 \sum_{i=1}^{K} \sqrt{\mathbb{E}\left[p_{n,i}\right]}.$$

Furthermore, if $\eta_n b_n \leq \frac{1}{4}$, then for any fixed j:

$$\mathbb{E}\left[c_{n,I_n} + \Phi_n(-\tilde{C}_n) - \Phi_n(-\tilde{C}_{n-1})\right] \\ \leq \frac{\eta_n}{2} b_n^2 \sum_{i \neq j} \left(\sqrt{\mathbb{E}\left[p_{n,i}\right]} + 2.5\mathbb{E}\left[p_{n,i}\right] \right).$$

In particular, if there exists N_0 such that for all $n \geq N_0$, $\eta_n b_n \leq \frac{1}{4}$, then:

$$\mathbb{E}\left[\sum_{n=1}^{N} c_{n,I_n} + \Phi_n(-\tilde{C}_n) - \Phi_n(-\tilde{C}_{n-1})\right] \\ \leq \sum_{n=1}^{N} \frac{\eta_n}{2} b_n^2 \sum_{i \neq j} \left(\sqrt{\mathbb{E}\left[p_{n,i}\right]} + 2.5\mathbb{E}\left[p_{n,i}\right] \right) + \sum_{n=1}^{N_0} \frac{\eta_n}{2} b_n^2.$$

The penalty term is not affected by the change of the range of the losses.

Lemma 7. *For any non-increasing positive learning rate sequence $(\eta_n)_{n\geq 1}$, the* penalty *term of the regret bound of* Tsallis-Switch *satisfies:*

$$\mathbb{E}\left[\sum_{n=1}^{N} \Phi_n(-\tilde{C}_{n-1}) - \Phi_n(-\tilde{C}_n) - c_{n,i_N^*}\right] \leq \frac{4\sqrt{K}}{\eta_N} + 1.$$

Furthermore, if we define η_0, such that $\eta_0^{-1} = 0$, then

$$\mathbb{E}\left[\sum_{n=1}^{N} \Phi_n(-\tilde{C}_{n-1}) - \Phi_n(-\tilde{C}_n) - c_{n,i_N^*}\right] \\ \leq 4\sum_{n=1}^{N} (\eta_n^{-1} - \eta_{n-1}^{-1}) \sum_{i \neq i_N^*} \left(\sqrt{\mathbb{E}\left[p_{n,i}\right]} - \frac{1}{2}\mathbb{E}\left[p_{n,i}\right] \right) + 1.$$

We also present a bound for the cumulative switching cost, which is the key to obtain refined guarantees in the stochastically constrained adversarial regime.

Lemma 8. *Consider a sequence of switching costs $(\lambda_n)_{n\geq 1}$. Then for any fixed j, the cumulative switching cost satisfies*

$$S\left(T, (\lambda_n)_{n\geq 1}\right) \leq \lambda_1 + \sum_{n=1}^{N} (\lambda_n + \lambda_{n+1}) \sum_{i \neq j} \mathbb{P}(I_n = i).$$

Proof of Lemma 8. By convention, there is always a switch at round 1. For subsequent rounds, when there is a switch at round n at least one of I_{n-1} or I_n is not equal to j. Thus, we have:

$$\mathbb{P}(I_{n-1} \neq I_n) \leq \sum_{i \neq j} \mathbb{P}(I_{n-1} = i) + \mathbb{P}(I_n = i),$$

and the cumulative switching cost satisfies

$$\begin{aligned} S\left(T, (\lambda_n)_{n\geq 1}\right) &= \lambda_1 + \sum_{n=2}^{N} \lambda_n \mathbb{P}(I_{n-1} \neq I_n) \\ &\leq \lambda_1 + \sum_{n=2}^{N} \lambda_n \left(\sum_{i \neq j} \mathbb{P}(I_{n-1} = i) + \mathbb{P}(I_n = i) \right) \\ &\leq \lambda_1 + \sum_{n=1}^{N} \sum_{i \neq j} (\lambda_n + \lambda_{n+1}) \mathbb{P}(I_n = i), \end{aligned}$$

which concludes the proof. □

Armed with these results, we can move on to the proof of Theorem 1.

Proof of Theorem 1. In order to apply our results to blocks, we first calculate an upper bound on the number of blocks N. The length of the n-th block is defined as $|B_n| = \max\left\{\left\lceil \frac{3\lambda\sqrt{n}}{2\sqrt{K}} \right\rceil, 1\right\}$. The sequence $(B_n)_{n\geq 1}$ satisfies $|B_n| \geq b(n)$ for $b(n) = \frac{3\lambda\sqrt{n}}{2\sqrt{K}}$ and is non-decreasing. Let $N^* = K^{1/3}(T/\lambda)^{2/3}$ and observe that:

$$\begin{aligned} \sum_{n=1}^{\lfloor N^* \rfloor + 1} |B_n| &\geq \sum_{n=1}^{\lfloor N^* \rfloor + 1} \frac{3\lambda\sqrt{n}}{2\sqrt{K}} \geq \int_0^{\lfloor N^* \rfloor + 1} \frac{3\lambda\sqrt{n}}{2\sqrt{K}} \\ &\geq \int_0^{N^*} \frac{3\lambda\sqrt{n}}{2\sqrt{K}} = \frac{\lambda}{\sqrt{K}} (N^*)^{3/2} \geq T. \end{aligned}$$

Thus, we can upper bound N by $K^{1/3}(T/\lambda)^{2/3} + 1$.

Proof of the adversarial bound. We start by focusing on the bound in the adversarial regime. To do so, we need to control the stability and penalty terms in (6), and also the number of switches. As we already said, the number of switches is bounded by the number of blocks, $S_T \leq N \leq K^{1/3}(T/\lambda)^{2/3} + 1$, and thus the cumulative switching cost satisfies $\lambda S_T \leq \lambda N \leq K^{1/3} T^{2/3} \lambda^{1/3} + \lambda$.

Next, we bound the quantity $\eta_n |B_n|^2$ for all $n \leq N$:

$$\frac{\eta_n}{2} |B_n|^2 \leq \frac{\sqrt{2}}{\sqrt{n}} \left(\frac{3\lambda\sqrt{n}}{2\sqrt{K}} + 1 \right) \leq \frac{3\lambda}{\sqrt{2K}} + \frac{\sqrt{2}}{\sqrt{n}}. \tag{7}$$

Note that even though the last block B_N may be truncated, we can upper bound its length by the non-truncated length of that block.

Then, we bound the inverse of the learning rate at round N,

$$\frac{1}{\eta_N} \leq \frac{\sqrt{N}}{2\sqrt{2}} \left(\frac{3\lambda\sqrt{N}}{2\sqrt{K}} + 1 \right) \leq \frac{3\sqrt{2}}{8} \frac{\lambda N}{\sqrt{K}} + \frac{\sqrt{2}}{4}\sqrt{N}.$$

In order to bound the pseudo-regret over the N blocks, we apply inequality (3) from Theorem 2. We then add the cumulative switching cost and use the upper bound on N derived earlier,

$$\begin{aligned}R(T,\lambda) &\leq 3\sqrt{2}\lambda N + 3\sqrt{2KN} + \lambda N + 1\\ &= (3\sqrt{2}+1)\lambda N + 3\sqrt{2KN} + 1\\ &\leq 5.25\lambda^{1/3}K^{1/3}T^{2/3} + 3\sqrt{2}\frac{K^{2/3}T^{1/3}}{\lambda^{1/3}}\\ &\quad + 3\sqrt{2K} + 5.25\lambda + 6.25.\end{aligned}$$

For small λ the term $K^{2/3}(T/\lambda)^{1/3}$ dominates the expression. However, when $\lambda \leq \frac{2}{3}\sqrt{\frac{K}{T}}$, then for all $n \leq T$ we have $\frac{3\lambda\sqrt{n}}{2\sqrt{K}} \leq \sqrt{\frac{n}{T}} \leq 1$, which means that $|B_n| = 1$. In this case the algorithm is not using blocks and we have $\lambda S_T \leq \lambda T \leq \frac{2}{3}\sqrt{KT}$. As we also have $a_n \leq 1$, we get $\frac{\sqrt{2}}{\sqrt{n}} \leq \eta_n \leq \frac{2\sqrt{2}}{\sqrt{n}}$. In this case we use Lemmas 6 and 7 to bound the stability and the penalty terms and obtain that stability and penalty are both bounded by $2\sqrt{2KN}$. Thus, overall, for $\lambda \leq \frac{2}{3}\sqrt{\frac{K}{T}}$ we have $R(T,\lambda) \leq 6.4\sqrt{KT}$, and for $\lambda \geq \frac{2}{3}\sqrt{\frac{K}{T}}$ we have $K^{2/3}(T/\lambda)^{1/3} \leq 1.15\sqrt{KT}$.

Piecing together all parts of the bound finishes the proof.

Proof of the stochastically constrained adversarial bound. We now derive refined guarantees in the stochastically constrained adversarial regime with a unique best arm i^*. We start by deriving bounds for the stability and penalty terms in (6).

Let N_0 be a constant, such that for $n \geq N_0$ we have $\eta_n|B_n| \leq \frac{1}{4}$. We note that $\eta_n|B_n| \leq \frac{2\sqrt{2}}{\sqrt{n}}$, so picking $N_0 = 128$ works. For the stability term we use the second part of Lemma 6 with $j = i^*$. Using (7) to bound $\frac{\eta_n}{2}|B_n|^2$ we obtain that the stability term is upper bounded by

$$\sum_{n=1}^{N}\left(\frac{3\sqrt{2}\lambda}{2\sqrt{K}} + \frac{\sqrt{2}}{\sqrt{n}}\right)\sum_{i\neq i^*}\left(\sqrt{\mathbb{E}\left[p_{n,i}\right]} + 2.5\mathbb{E}\left[p_{n,i}\right]\right) + \sum_{n=1}^{N_0}\left(\frac{3\sqrt{2}}{2}\frac{\lambda}{\sqrt{K}} + \frac{\sqrt{2}}{\sqrt{n}}\right).$$

For the penalty term, we first bound the difference between the inverse of two consecutive learning rates.

$$\begin{aligned}&\eta_n^{-1} - \eta_{n-1}^{-1}\\ &\quad= \left(\frac{3\lambda\sqrt{n}}{2\sqrt{K}} + 1\right)\frac{\sqrt{n}}{2\sqrt{2}} - \left(\frac{3\lambda\sqrt{n-1}}{2\sqrt{K}} + 1\right)\frac{\sqrt{n-1}}{2\sqrt{2}}\\ &\quad= \frac{3\sqrt{2}\lambda}{8\sqrt{K}} + \frac{\sqrt{n}-\sqrt{n-1}}{2\sqrt{2}}\\ &\quad\leq \frac{3\sqrt{2}\lambda}{8\sqrt{K}} + \frac{\sqrt{2}}{4\sqrt{n}}.\end{aligned}$$

Now we use the second part of Lemma 7 to bound the penalty term as follows

$$\sum_{n=1}^{N}\left(\frac{3\sqrt{2}\lambda}{2\sqrt{K}} + \frac{\sqrt{2}}{\sqrt{n}}\right)\sum_{i\neq i^*}\left(\sqrt{\mathbb{E}\left[p_{n,i}\right]} - \frac{1}{2}\mathbb{E}\left[p_{n,i}\right]\right) + 1.$$

Summing the two bounds, and using that for all n, i, $\mathbb{E}\left[p_{n,i}\right] \leq \sqrt{\mathbb{E}\left[p_{n,i}\right]}$, we have:

$$\begin{aligned}R_T \leq \sum_{n=1}^{N}&\left(\left(\frac{6\sqrt{2}\lambda}{\sqrt{K}} + \frac{4\sqrt{2}}{\sqrt{n}}\right)\sum_{i\neq i^*}\sqrt{\mathbb{E}\left[p_{n,i}\right]}\right)\\ &+ \frac{3\sqrt{2}\lambda}{2\sqrt{K}}N_0 + 2\sqrt{2N_0} + 1.\end{aligned}$$

Now we use the self-bounding technique (Zimmert & Seldin, 2021), which states that if L and U are such that $L \leq R \leq U$, then $R \leq 2U - L$. For the lower bound L, we use the following identity for the regret

$$R_T = \sum_{n=1}^{N}|B_n|\sum_{i\neq i^*}\Delta_i\mathbb{E}\left[p_{n,i}\right],$$

where B_N is truncated, so that $|B_1| + \cdots + |B_N| = T$. Using the previous expression for the upper bound U, we get:

$$\begin{aligned}R_T \leq &\sum_{n=1}^{N}\left(\frac{12\sqrt{2}\lambda}{\sqrt{K}} + \frac{8\sqrt{2}}{\sqrt{n}}\right)\sum_{i\neq i^*}\sqrt{\mathbb{E}\left[p_{n,i}\right]}\\ &- \sum_{n=1}^{N}|B_n|\sum_{i\neq i^*}\Delta_i\mathbb{E}\left[p_{n,i}\right] + \frac{544\lambda}{\sqrt{K}} + 66.\end{aligned}$$

We bound the cumulative switching cost using Lemma 8:

$$\lambda S_T \leq \lambda + \sum_{n=1}^{N}\sum_{i\neq i^*}2\lambda\mathbb{E}\left[p_{n,i}\right].$$

We add those two bounds together to obtain a bound on the regret with switching costs. Note (again) that $\mathbb{E}\left[p_{n,i}\right] \leq \sqrt{\mathbb{E}\left[p_{n,i}\right]}$ for all n and i, and that $\frac{\sqrt{2}}{\sqrt{K}} \leq 1$. Thus, we can upper bound the pseudo-regret with switching costs as:

$$\begin{aligned}&R(T,\lambda)\\ &\leq \sum_{n=1}^{N}\sum_{i\neq i^*}\left(\left(14\lambda + \frac{8\sqrt{2}}{\sqrt{n}}\right)\sqrt{\mathbb{E}\left[p_{n,i}\right]} - \Delta_i|B_n|\mathbb{E}\left[p_{n,i}\right]\right)\\ &\quad + \frac{544\lambda}{\sqrt{K}} + \lambda + 66.\end{aligned}$$

Now we note that each term in the inner sum is an expression of the form $a\sqrt{x} - bx$, which for $x \in [0,\infty]$ is maximized at $x = \frac{a^2}{4b}$. Put attention that the cumulative switching cost is part of the optimization problem. So, for any i and any $n < N$, we have:

$$
\begin{aligned}
&\left(14\lambda + \frac{8\sqrt{2}}{\sqrt{n}}\right)\sqrt{\mathbb{E}\left[p_{n,i}\right]} - \Delta_i |B_n| \mathbb{E}\left[p_{n,i}\right] \\
&\quad\le \frac{\left(14\lambda + \frac{8\sqrt{2}}{\sqrt{n}}\right)^2}{4\Delta_i |B_n|} \\
&\quad\le \frac{(14\lambda)^2}{4\Delta_i\left(\frac{3\lambda\sqrt{n}}{2\sqrt{K}}\right)} + 2\frac{14\lambda\left(\frac{8\sqrt{2}}{\sqrt{n}}\right)}{4\Delta_i} + \frac{\left(\frac{8\sqrt{2}}{\sqrt{n}}\right)^2}{4\Delta_i} \qquad (8)\\
&\quad\le \frac{33\lambda\sqrt{K}}{\Delta_i\sqrt{n}} + \frac{80\lambda}{\Delta_i\sqrt{n}} + \frac{32}{\Delta_i n}, \qquad (9)
\end{aligned}
$$

where in the first term of (8) we have lower bounded $|B_n|$ by b_n and in the last two terms by 1. As the last block may be truncated, for $n = N$ we bound $|B_N|$ in the first term in (9) by 1, leading to

$$
\begin{aligned}
&\left(14\lambda + \frac{8\sqrt{2}}{\sqrt{N}}\right)\sqrt{\mathbb{E}\left[p_{N,i}\right]} - \Delta_i |B_N| \mathbb{E}\left[p_{N,i}\right] \\
&\qquad\qquad\le \frac{49\lambda^2}{\Delta_i} + \frac{80\lambda}{\Delta_i\sqrt{n}} + \frac{32}{\Delta_i n},
\end{aligned}
$$

All that remains is to sum over n. For the first term in (9) we have:

$$
\begin{aligned}
\frac{49\lambda^2}{\Delta_i} + \sum_{n=1}^{N-1}\frac{33\lambda\sqrt{K}}{\Delta_i\sqrt{n}} &\le 66\frac{\lambda\sqrt{K(N-1)}}{\Delta_i} + \frac{49\lambda^2}{\Delta_i} \\
&\le 66\frac{\lambda^{2/3}T^{1/3}K^{2/3}}{\Delta_i} + \frac{49\lambda^2}{\Delta_i}.
\end{aligned}
$$

Similarly, the second term in (9) gives:

$$
\sum_{n=1}^{N}\frac{80\lambda}{\Delta_i\sqrt{n}} \le 160\frac{\lambda\sqrt{N}}{\Delta_i} \le 160\frac{\lambda^{2/3}T^{1/3}K^{1/6}+\lambda}{\Delta_i}.
$$

For the last term in (9), we use the fact that $N \le T$ and we have:

$$
\sum_{n=1}^{N}\frac{32}{\Delta_i n} \le \frac{32\ln T}{\Delta_i} + \frac{32}{\Delta_i}.
$$

Putting everything together finishes the proof. □

7. Experiments

We compare the performance of Tsallis-Switch to different baselines, both in the stochastic and in the stochastically constrained adversarial regime. We compare Tsallis-Switch with block lengths chosen as in Theorem 1 against Tsallis-INF without blocks, and against the BaSE algorithm of Gao et al. (2019), which achieves a regret of $\mathcal{O}\left(\sum_{i\neq i^*}\frac{\log T}{\Delta_i}\right)$ with $\mathcal{O}(\log T)$ switches in the stochastic regime. We use T to tune the parameters of BaSE, and we consider both arithmetic and geometric blocks —see (Gao et al., 2019) for details.

We also include in our baselines the EXP3 algorithm with a time-varying learning rate, and the block version of EXP3, where the blocks have length $\lambda^{2/3}\frac{T^{1/3}}{K^{1/3}}$. Both block length and learning rate are chosen according to the analysis of EXP3 in the adversarial regime.

In the experiments, we fix the number of arms $K = 8$, and set the expected loss of a suboptimal arm to 0.5. We generate binary losses using two sets of parameters: an "easy" setting, where the gaps $\Delta = 0.2$ are large and the switching costs $\lambda = 0.025$ are small. A "hard" setting, where the gaps $\Delta = 0.05$ are small and the switching costs $\lambda = 1$ are large. For each experiment, we plot the pseudo-regret, the number of switches, and the pseudo-regret with switching cost. This allows us to observe the trade-off between the pseudo-regret and the number of switches.

In the first experiment (Figure 2) we use stochastic i.i.d. data with the easy setting ($\Delta = 0.2$ and $\lambda = 0.025$). As the gaps are large, even the methods that do not use blocks are not making many switches, and the best performance is achieved by Tsallis-INF without blocks. In Figure 3 we use the hard setting ($\Delta = 0.05$ and $\lambda = 1$). In this case, we see a trade-off between achieving a small pseudo-regret and limiting the cumulative switching cost. The small value of Δ forces a larger number of switches, and because the cost of switching is now large, the cumulative switching cost dominates the pseudo-regret with switching cost.

In Figure 4, we test a stochastic setting with small gaps and zero switching cost. In this case, we observe that Tsallis-Inf and Tsallis-Switch outperform both EXP3 and the BaSE algorithms. Note that here Tsallis-Switch and Tsallis-Inf have very similar performances, though not identical due to a slight difference in the tuning of learning rates.

We present a wider range of experiments in Appendix E. We show that our algorithm outperforms the BaSE algorithm in the stochastically constrained adversarial regime. Being an elimination-based algorithm, BaSE also fails in the adversarial regime.

8. Discussion

We introduced Tsallis-Switch, the first algorithm for multi-armed bandits with switching costs that provides adversarial pseudo-regret guarantees simultaneously with improved pseudo-regret guarantees in the stochastic regime, as well as the more general stochastically constrained adversarial regime. The adversarial regret bound matches the minimax lower bound within constants, and guarantees $T^{2/3}$ scaling

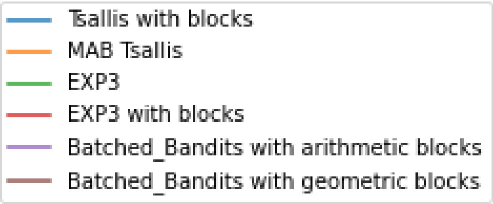

Figure 1. Legend for all plots.

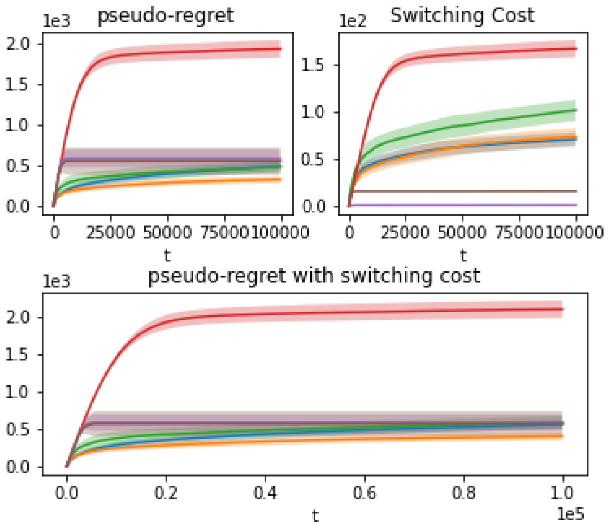

Figure 2. Stochastic losses, $\Delta = 0.2$ and $\lambda = 0.025$ (easy setting).

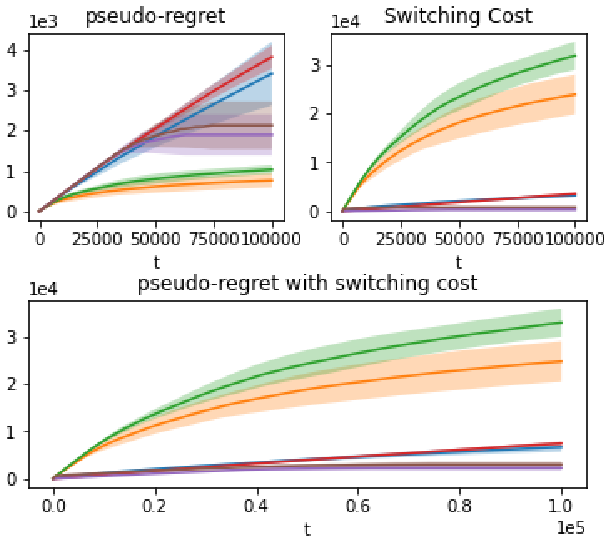

Figure 3. Stochastic losses, $\Delta = 0.05$ and $\lambda = 1$ (hard setting).

of the regret in time. The stochastic and stochastically constrained adversarial bounds reduce the dependence of the regret on time down to $T^{1/3}$. Our experiments demonstrate that Tsallis-Switch is competitive with the relevant benchmarks over a range of settings: in the stochastic setting, it is competitive with state-of-the-art algorithms for stochastic bandits with switching costs, and outperforms state-of-the-art adversarial algorithms. In the adversarial setting, it is competitive with state-of-the-art adversarial algorithms and significantly outperforms the stochastic ones.

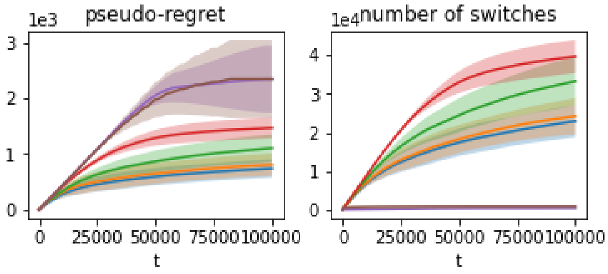

Figure 4. Stochastic losses and no switching cost, $\lambda = 0$ and $\Delta = 0.05$. As the switching costs are 0, the pseudo-regret and the pseudo-regret with switching costs are equal.

Our work opens multiple directions for future research. For example, it is known that in the stochastic setting with switching costs it is possible to achieve logarithmic regret scaling, but it is unknown whether it is achievable simultaneously with the adversarial regret guarantee. It is also unknown whether logarithmic regret scaling is achievable for the more general stochastically constrained adversarial regime with switching costs (even with no simultaneous requirement of an adversarial regret guarantee). Elimination of the assumption on uniqueness of the best arm in the stochastically constrained adversarial regime is another challenging direction to work on. Unfortunately, for now it is unknown how to eliminate this assumption even in the analysis of the Tsallis-INF algorithm for multiarmed bandits without switching costs. But while in the setting without switching costs the assumption has been empirically shown to be an artifact of the analysis having no negative impact on the regret (Zimmert & Seldin, 2021), in the setting with switching costs treating multiple best arms is more challenging, because switching between best arms is costly.

Acknowledgements

CR and YS acknowledge partial support by the Independent Research Fund Denmark, grant number 9040-00361B. NCB is partially supported by the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR) and by the EU Horizon 2020 ICT-48 research and innovation action number 951847, ELISE (European Learning and Intelligent Systems Excellence).

References

Auer, P. and Chiang, C.-K. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2016.

Blum, A. and Mansour, Y. Learning, regret minimization, and equilibria. In Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V. (eds.), *Algorithmic game theory*. Cambridge University Press, 2007.

Bubeck, S. and Slivkins, A. The best of both worlds: stochastic and adversarial bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2012.

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

Dekel, O., Ding, J., Koren, T., and Peres, Y. Bandits with switching costs: $t^{2/3}$ regret. In *Proceedings of the Annual Symposium on the Theory of Computing (STOC)*, 2013.

Esfandiari, H., Karbasi, A., Mehrabian, A., and Mirrokni, V. Regret bounds for batched bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

Jin, T. and Luo, H. Simultaneously learning stochastic and adversarial episodic MDPs with known transition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Lattimore, T. and Svepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.

Masoudian, S. and Seldin, Y. Improved analysis of robustness of the Tsallis-INF algorithm to adversarial corruptions in stochastic multiarmed bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2021.

Rouyer, C. and Seldin, Y. Tsallis-inf for decoupled exploration and exploitation in multi-armed bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2020.

Seldin, Y. and Lugosi, G. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2017.

Seldin, Y. and Slivkins, A. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

Wei, C. and Luo, H. More adaptive algorithms for adversarial bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2018.

Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Zimmert, J. and Seldin, Y. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 2021.

Zimmert, J., Luo, H., and Wei, C. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.