Tsallis-INF for Decoupled Exploration and Exploitation in Multi-armed Bandits

Chloé Rouyer CHLOE@DI.KU.DK
Yevgeny Seldin SELDIN@DI.KU.DK
University of Copenhagen

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We consider a variation of the multi-armed bandit problem, introduced by Avner et al. (2012), in which the forecaster is allowed to choose one arm to explore and one arm to exploit at every round. The loss of the exploited arm is blindly suffered by the forecaster, while the loss of the explored arm is observed without being suffered. The goal of the learner is to minimize the regret. We derive a new algorithm using regularization by Tsallis entropy to achieve best of both worlds guarantees. In the adversarial setting we show that the algorithm achieves the minimax optimal $O(\sqrt{KT})$ regret bound, slightly improving on the result of Avner et al.. In the stochastic regime the algorithm achieves a time-independent regret bound, significantly improving on the result of Avner et al.. The algorithm also achieves the same time-independent regret bound in the more general stochastically constrained adversarial regime introduced by Wei and Luo (2018).

Keywords: Multi-armed Bandits, Decoupling Exploration and Exploitation in Multi-armed Bandits, Tsallis Entropy

1. Introduction

The multi-armed bandit problem is a central and most basic framework for studying the exploration-exploitation trade-off (Thompson, 1933; Robbins, 1952; Lai and Robbins, 1985; Auer et al., 2002a,b; Slivkins, 2019; Lattimore and Svepesvári, 2020). In the multi-armed bandit game a player repeatedly chooses actions (also called arms) from a set of K actions and observes and suffers the loss of the selected action. This can be contrasted with the full information setting, where after selecting an action the player observes losses of all actions, not just the selected one (Cesa-Bianchi and Lugosi, 2006). The losses may be generated adversarially or stochastically, depending on problem setup. The goal of the learner is to find an action selection strategy minimizing the regret, which is the difference between the cumulative loss of the player and of the best fixed action in hindsight.

We focus on a variation of the multi-armed bandit problem introduced by Avner et al. (2012), in which at each round the learner is allowed to choose one action to play blindly and one action to observe without suffering its loss. The two actions are allowed, but not required to be different. Thus, exploration is decoupled from exploitation. Practical settings having this structure are full information problems with restricted data access, where in principle the loss of any action could be accessed, but each observation, including the one of the selected action, is associated with a cost and the player can only afford one observation per round.

The decoupled setting takes an important place in the space of online learning problems. On the one hand, it is a bridge between full information and bandit setups. In particular, as we discuss below, in the adversarial regime the problem is as hard as a bandit problem, but in the stochastic regime the regret scaling is time-independent, as in full information problems. Seldin et al. (2014) expand this bridge further by introducing multi-armed bandits with paid observations, where a learner can make an arbitrary number of observations at corresponding costs, which provides a continuous interpolation between full information and bandits. On the other hand, the decoupled setting is a bridge between exploration-exploitation and pure exploration problems (Even-Dar et al., 2006; Mannor and Tsitsiklis, 2004; Bubeck et al., 2011). In particular, one could think about applying an algorithm developed for best-arm identification for selecting the exploration arm, although this is not an optimal strategy for the decoupled setting.

Avner et al. (2012) have shown that in the adversarial regime there is a lower bound of $\Omega(\sqrt{KT})$ for the regret in the decoupled setting. Thus, in the worst case the adversary can make the regret as large as in the standard multi-armed bandits. However, they have also shown that in some situations, in particular when one arm dominates all other arms, the regret can be reduced. More specifically, they have proposed an EXP3-style algorithm with the same exploitation strategy as EXP3, but modified exploration strategy, which achieves an $O(\sqrt{KT \ln K})$ regret bound in the worst case adversarial regime and an improved $O(\sqrt{T \ln K})$ regret bound in an adversarial regime with one dominating arm. A similar improvement in dependence on the number of arms was also shown by Seldin et al. (2014) for bandits with paid observations. Avner et al. have also analyzed their algorithm in the stochastic setting, showing that in a configuration with a single best arm (which would thus be dominating) the regret grows as $O(\sqrt{T \ln K})$. However, the analysis required a different tuning of the learning rate for the stochastic setting than for the adversarial one and, therefore, prior knowledge of the regime was essential. The stochastic regret bound was also highly suboptimal, since a simple approach of playing Follow the Leader for exploitation and uniform distribution for exploration leads to a time-independent expected regret bound of $O(\sum_{i:\Delta_i>0} \frac{K}{\Delta_i})$ in the stochastic setting.

Traditionally algorithms for multi-armed bandits and their variations, including the algorithm of Avner et al., were relying on prior knowledge of the nature of the environment, but following the work of Bubeck and Slivkins (2012) there has been a growing interest in algorithms that perform well in both settings without this knowledge (Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Wei and Luo, 2018). Zimmert and Seldin (2019) have eventually used Tsallis entropy regularizer with power $\frac{1}{2}$ to derive an algorithm that achieves the optimal regret bounds for multi-armed bandits in both settings with no prior knowledge of the regime. (Interestingly, the same kind of regularizer has been earlier used by Audibert and Bubeck (2009) to achieve the optimal regret bound for adversarial bandits.) We follow this line of work and propose an algorithm for multi-armed bandits with decoupled exploration and exploitation that achieves refined regret guarantees in both adversarial and stochastic regimes and requires no prior knowledge of the regime. Specifically, we make the following contributions:

- We propose a new algorithm for decoupled exploration and exploitation in multi-armed bandits based on Follow the Regularized Leader framework with regularization by Tsallis entropy.

- We show that in the adversarial regime the algorithm achieves $O(\sqrt{KT})$ regret upper bound, improving by a multiplicative factor of $\sqrt{\ln K}$ on the worst-case upper bound of Avner et al. and matching their worst-case lower bound within constants.

- We show that the same algorithm achieves a time independent $O(\max_{i:\Delta_i>0} \frac{K}{\Delta_i})$ regret bound in the stochastic regime, considerably improving on the result of Avner et al.. (The result holds under a technical assumption that the best arm is unique.)

- The same regret bound is achieved in a more general stochastically constrained adversarial regime introduced by Wei and Luo (2018) (also under the assumption on uniqueness of the best arm).

- The algorithm requires no prior knowledge of the nature of the environment.

- Interestingly, the results are achieved with Tsallis entropy regularizer with power $\alpha = \frac{2}{3}$, whereas the optimal power for standard multi-armed bandits is $\alpha = \frac{1}{2}$. In our analysis the power $\alpha = \frac{1}{2}$ does not achieve time-independent stochastic regret bounds in the decoupled setting and, therefore, inferior to $\alpha = \frac{2}{3}$.

The assumption on uniqueness of the best arm in the stochastic and stochastically constrained adversarial regimes underlies the prior work of Zimmert and Seldin (2019). We conjecture that it can be eliminated.

The paper is structured in the following way. In Section 3 we introduce the Follow the Regularized Leader framework and the approach used to decouple exploration and exploitation. The algorithm and main results are presented in Section 4 and their proofs can be found in Section 5. We conclude with a discussion in Section 6.

2. Problem Setting and Notation

We consider a repeated game with K arms. At each round $t = 1, 2, \dots$ of the game the environment picks a loss vector $\ell_t \in [0,1]^K$ and the learner picks an action A_t to exploit and an action B_t to explore. The two actions are allowed to be different, but may also be identical. Then the learner blindly suffers ℓ_{t,A_t} and observes ℓ_{t,B_t} without suffering its loss.

In the oblivious adversarial setting, the environment chooses ℓ_t arbitrarily prior to the beginning of the game.

In the stochastic setting the losses are drawn from distributions with fixed means, i.e., for all i we have $\mathbb{E}\left[\ell_{t,i}\right] = \mu_i$ independently of t.

We also consider a more general stochastically constrained adversarial setting (Wei and Luo, 2018; Zimmert and Seldin, 2019). In this setting the losses are drawn from distributions with fixed gaps, while the baseline means are allowed to fluctuate, i.e., for all i, j we have $\mathbb{E}\left[\ell_{t,i} - \ell_{t,j}\right] = \Delta_{i,j}$ independently of t. The stochastic setting is a special case of the stochastically constrained adversary with $\Delta_{i,j} = \mu_i - \mu_j$. All results in the paper are presented for stochastically constrained adversaries and extend to stochastic environments as a special case.

We measure the performance of an algorithm in terms of pseudo-regret:

$$\mathcal{R}_T := \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \left(\ell_{t,A_t} - \ell_{t,i_T^*}\right)\right],$$

where $i_T^* = \arg\min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right]$ is the best action in hindsight. In the oblivious adversarial regime the losses are independent of the player's actions and the pseudo-regret coincides with the notion of expected regret (Bubeck and Cesa-Bianchi, 2012), defined as

$$\mathcal{R}_T := \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} - \min_i \sum_{t=1}^{T} \ell_{t,i}\right].$$

In the stochastically constrained adversarial setting we let $i^* = \arg\min_i \Delta_{i,1}$ denote an optimal arm (we can take any arm j as the second argument of $\Delta_{i,j}$ in the definition of i^*). Then we have $i_T^* = i^*$ for all T. We define $\Delta_i = \Delta_{i,i^*}$ to be the gaps to the best arm and rewrite the pseudo-regret in the stochastically constrained adversarial setting as

$$\mathcal{R}_T = \sum_{t=1}^{T} \sum_{i \neq i^*} \mathbb{E}\left[p_{t,i}\right] \Delta_i, \tag{1}$$

where $p_{t,i}$ is the probability that arm i is played at round t.

3. Decoupling Exploration and Exploitation in Follow the Regularized Leader Framework

The algorithm that we present is based on follow the regularized leader (FTRL) framework (Shalev-Shwartz, 2012). Following Zimmert and Seldin (2019), we use the regularizer

$$\Psi_t(w) = -\frac{1}{\eta_t} \sum_i \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)}, \tag{2}$$

which is a slight modification of the negative Tsallis entropy with power α defined by $H_\alpha(w) := \frac{1}{1-\alpha}(1 - \sum_i w_i^\alpha)$ (Tsallis, 1988). We focus on $\alpha \in (0,1)$, but one of the interesting properties of the above regularizer is that in the limits $\alpha \to 0$ and $\alpha \to 1$ it recovers the log-barrier and the negative entropy regularizers, respectively (Zimmert and Seldin, 2019). In particular, the EXP3 algorithm with losses (Auer et al., 2002b; Bubeck and Cesa-Bianchi, 2012) can be seen as a limit case of FTRL with regularization by Tsallis entropy with $\alpha \to 1$. As we are in the bandit setting and only observe one element of the loss vector at each round, we construct an unbiased estimate $\tilde{\ell}_t$ of the loss vector ℓ_t by using importance-weighted sampling

$$\forall t \in [T], i \in [K], \qquad \tilde{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{1}(B_t = i)}{q_{t,i}},$$

where q_t is the distribution for sampling the exploratory action B_t and $\mathbb{1}$ is the indicator function.

We define the Decoupled-Tsallis-INF algorithm for an arbitrary exploration distribution q_t in Algorithm 1.

In order to analyse the algorithm, we decompose the pseudo-regret into a stability and penalty components (Lattimore and Svepesvári, 2020; Zimmert and Seldin, 2019),

$$\mathcal{R}_T = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right]}_{\text{stability}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \Phi_t(-\tilde{L}_{t-1}) - \Phi_t(-\tilde{L}_t) - \ell_{t,i_T^*}\right]}_{\text{penalty}}, \tag{3}$$

Algorithm 1: Decoupled-Tsallis-INF

```
Input      : Learning rates $\eta_1 \geq \eta_2 \geq \dots > 0$.
Initialize: $\tilde{L}_0 = \mathbf{0}_K$
for $t = 1, 2, \dots$ do
    $p_t = \arg\min_{p \in \Delta^{K-1}} \left\{ \left\langle p, \tilde{L}_{t-1} \right\rangle - \frac{1}{\eta_t} \sum_{i=1}^{K} \frac{p_i^\alpha - \alpha p_i}{\alpha(1-\alpha)} \right\}$
    Construct exploration distribution $q_t$ (see Lemma 2)
    Sample $A_t$ according to $p_t$, play it and suffer $\ell_{t,A_t}$.
    Sample $B_t$ according to $q_t$ and observe $\ell_{t,B_t}$.
    $\forall\, i \in [K]: \quad \tilde{\ell}_{t,i} = \frac{\ell_{t,i} \mathbb{1}\{B_t = i\}}{q_{t,i}} = \begin{cases} \frac{\ell_{t,i}}{q_{t,i}}, & \text{if } B_t = i, \\ 0, & \text{otherwise.} \end{cases}$
    $\forall\, i \in [K]: \quad \tilde{L}_t(i) = \tilde{L}_{t-1}(i) + \tilde{\ell}_{t,i}$.
end
```

where i_T^* is the best action in hindsight, and the potential function is defined by

$$\Phi_t(-L) = \max_{w \in \Delta^{K-1}} \left\{ \langle w, -L \rangle + \frac{1}{\eta_t} \sum_{i=1}^{K} \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)} \right\}.$$

In order to achieve a tight bound on the pseudo-regret, one has to derive tight bounds on the stability and penalty. Recall that $\tilde{L}_t$ is an unbiased estimate of L_t and observe that the penalty term does not depend on the query distribution q_t. The stability term of the regret of Algorithm 1 satisfies the following lemma.

Lemma 1 *For any $\alpha \in (0,1)$ and any positive learning rate value, the stability term of the regret of Decoupled-Tsallis-INF with an arbitrary exploration distribution q_t satisfies:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1}) \right] \leq \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_t}{2} \frac{(p_{t,i})^{2-\alpha}}{q_{t,i}} \right].$$

A proof of the lemma is provided in Appendix B.2. We can see that the bound depends on the choice of exploration distribution q_t, and that picking $q_t = p_t$ recovers the bound for the stability term of the regret of Tsallis-INF for multi-armed bandits (Zimmert and Seldin, 2019). In the decoupled case we have the freedom of picking $q_t \neq p_t$, so we select the distribution q_t which minimizes the bound on the stability term in Lemma 1.

Lemma 2 *The right hand side of the bound in Lemma 1 is minimized by the distribution q_t defined by*

$$\forall t \in [T], i \in [K], \quad q_{t,i} = \frac{(p_{t,i})^{1-\alpha/2}}{\sum_{j=1}^{K} (p_{t,j})^{1-\alpha/2}}.$$

We provide a proof of this Lemma in Appendix C. In the previous section, we have mentioned that in the limit of $\alpha \to 1$ Tsallis-INF converges to the EXP3 algorithm. By taking $\alpha = 1$ in Lemma 2 we recover the exploration distribution used by Avner et al. (2012): $q_{t,i} = \frac{\sqrt{p_{t,i}}}{\sum_{j=1}^{K} \sqrt{p_{t,j}}}$.

4. Main Results

In the rest of the paper we analyse Decoupled-Tsallis-INF with exploration distribution q_t defined in Lemma 2. The first theorem bounds the regret of Decoupled-Tsallis-INF in the adversarial regime.

Theorem 3 *In the adversarial regime, for any $\alpha \in (0,1)$ the pseudo-regret of Decoupled-Tsallis-INF with learning rate $\eta_t = \frac{2K^{1/2-\alpha}}{\sqrt{t}}$ and with q_t given by Lemma 2 satisfies:*

$$\mathcal{R}_T \leq \left(2 + \frac{1}{2\alpha(1-\alpha)}\right)\sqrt{KT} + 1.$$

We provide a proof of the theorem in Section 5.1. Avner et al. (2012) have derived a regret lower bound of $\Omega(\sqrt{KT})$ for the adversarial regime, which means that our algorithm is minimax optimal within constants. For comparison, the algorithm proposed by Avner et al. is suboptimal by a multiplicative factor of $\sqrt{\log K}$. In the next theorem we bound the regret of Decoupled-Tsallis-INF in the stochastically constrained adversarial setting.

Theorem 4 *In the stochastically constrained adversarial regime with a unique best action i^*, the pseudo-regret of Decoupled-Tsallis-INF with $\alpha \in (0, 2/3]$, $\eta_t = \frac{2K^{1/2-\alpha}}{\sqrt{t}}$, and with q_t given by Lemma 2 satisfies*

$$\mathcal{R}_T \leq O\left(\left(\sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \Delta_i^{\frac{\alpha}{\alpha-1}} \frac{\sqrt{K}}{t^{-\frac{1}{2(\alpha-1)}}}\right) + \frac{\sqrt{K}}{\Delta_{\min}}\right),$$

where $T_0 = \max_{i \neq i^*} \left\lceil D\left(\frac{8}{\Delta_i}\right)^2 \right\rceil$ for $D \geq 1$. For $\alpha > 1/2$ the bound is independent of the time horizon T. For $\alpha \in (1/2, 2/3]$ the pseudo-regret satisfies,

$$\mathcal{R}_T \leq \sum_{i \neq i^*} \left(C(\alpha)\frac{\sqrt{K}D^{\frac{2\alpha-1}{2\alpha-2}}}{\Delta_i}\right) + \frac{68\sqrt{KD}}{\Delta_{\min}} + 13\sqrt{K},$$

where

$$C(\alpha) = \frac{2-2\alpha}{2\alpha-1}\left(\frac{1}{2\alpha(1-\alpha)} + \frac{\left(\left(1 - \frac{(1-\alpha)}{4}\right)^{-\frac{2-\alpha}{1-\alpha}} + 1\right)}{2^{-1+\alpha/2}} + 2\right)^{\frac{1}{1-\alpha}} \left(\alpha^{\frac{\alpha}{1-\alpha}} - \alpha^{\frac{1}{1-\alpha}}\right) 8^{\frac{2\alpha-1}{\alpha-1}}.$$

A proof of the theorem is given in Section 5.2. The assumption on uniqueness of the best arm is a technical detail required in the analysis. The same assumption was used by Zimmert and Seldin (2019) in the analysis of Tsallis-INF. We conjecture that the assumption can be eliminated. The function $C(\alpha)$ is well-defined on the interval $(1/2, 2/3]$, and numerical evaluation shows that it is monotonically decreasing on the interval and minimized by $\alpha = 2/3$, for which we have $C(2/3) \leq 20$. Furthermore, for $D > 1$ the term $D^{\frac{2\alpha-1}{2\alpha-2}}$ is minimized by $\alpha = 2/3$, since $\frac{2\alpha-1}{2\alpha-2}$ decreases for $\alpha \in (1/2, 2/3]$. It is possible to derive a similar pseudo-regret bound for $\alpha \in [2/3, 1)$, however, for $\alpha > 2/3$ the dependency on K scales with $K^{\frac{2\alpha-1}{\alpha}}$ and the dependency on D with

$D^{1-1/\alpha}$. Both of these terms are increasing for $\alpha \in [2/3, 1)$, and thus minimized by $\alpha = 2/3$. Working in the $\alpha \in [2/3, 1)$ interval does not improve the dependency on neither t nor Δ. Thus, the bound is optimized by $\alpha = 2/3$.

The following corollary combines adversarial and stochastically constrained adversarial analysis (and stochastic regime as a special case of the latter).

Corollary 5 *For $\alpha = 2/3$, $\eta_t = \frac{2K^{1/2-\alpha}}{\sqrt{t}} = \frac{2K^{-1/6}}{\sqrt{t}}$, q_t given by Lemma 2, and $D \geq 1$, the regret of Decoupled-Tsallis-INF satisfies*

$$\mathcal{R}_T \leq 5\sqrt{KT} + 1$$

in the adversarial regime and

$$\mathcal{R}_T \leq \sum_{i \neq i^*} \frac{20}{\Delta_i} \frac{\sqrt{K}}{\sqrt{D}} + \frac{68\sqrt{KD}}{\Delta_{\min}} + 13\sqrt{K} \tag{4}$$

in the stochastically constrained adversarial regime with a unique best arm i^*. The two regret bounds hold simultaneously and with no need in prior knowledge of the regime. The bound in Equation (4) is minimized by $D = \max\left\{\frac{20}{68}\sum_{i \neq i^*} \frac{\Delta_{min}}{\Delta_i}, 1\right\}$, which gives

$$\mathcal{R}_T \leq 8\sqrt{\frac{K}{\Delta_{min}}} \max\left\{\sqrt{85\sum_{i \neq i^*} \frac{1}{\Delta_i}}, \frac{17}{\sqrt{\Delta_{min}}}\right\} + 13\sqrt{K}.$$

The corollary is a direct application of Theorem 3 and the second part of Theorem 4. By the corollary, the regret of Decoupled-Tsallis-INF in the stochastically constrained adversarial regime is $\mathcal{R}_T = O\left(\sqrt{\frac{K}{\Delta_{min}} \sum_{i \neq i^*} \frac{1}{\Delta_i}}\right)$. Since $\sum_{i \neq i^*} \frac{1}{\Delta_i} \leq \frac{K}{\Delta_{min}}$, we always have $\mathcal{R}_T = O\left(\frac{K}{\Delta_{min}}\right)$. However, if the delays are highly unbalanced, so that $\sum_{i \neq i^*} \frac{1}{\Delta_i} = O\left(\frac{1}{\Delta_{min}}\right)$, the regret bound can improve to $O\left(\frac{\sqrt{K}}{\Delta_{min}}\right)$.

Mourtada and Gaïffas (2019) have shown that in the full information stochastic case the regret lower bound is $\Omega\left(\frac{\log K}{\Delta}\right)$, assuming all the suboptimality gaps are equal, i.e., $\Delta_i = \Delta$ for all $i \neq i^*$. Since the decoupled setting is harder than the full information setting, the same lower bound applies to our case and there is a gap of $\frac{K}{\log K}$ between the upper and the lower bound in the case of identical suboptimality gaps. We conjecture that both the upper and the lower bound can be improved and that the actual regret scaling in the stochastically constrained adversarial regime should be of order $\sum_{i \neq i^*} \frac{1}{\Delta_i}$. Improvement of the upper and lower bounds is left for future work.

We note that unlike in the multi-armed bandit case, where $\alpha = 1/2$ is the optimal value both for the adversarial and the stochastically constrained adversarial regime, in the decoupled case there is a trade-off between the optimal values of α in the two regimes. However, the price of switching from the optimal $\alpha = 1/2$ to $\alpha = 2/3$ in the former is a minor multiplicative factor of $\frac{5}{4}$, whereas in the latter choosing $\alpha = 2/3$ eliminates the dependency of the regret on the time horizon and simultaneously provides a better dependency on K.

5. Proofs of the Theorems

Using the decomposition of the regret presented in Equation (3), we present bounds for the stability and penalty terms. We take advantage of the decoupling to refine the bound of Zimmert and Seldin (2019) on the stability term. The penalty term does not depend on the exploration distribution and, therefore, we reuse the bound derived by Zimmert and Seldin.

Lemma 6 *For any $\alpha \in (0,1)$ and any positive learning rate, the stability term of the regret bound of Decoupled-Tsallis-INF with exploration distribution q_t given by Lemma 2 satisfies:*

1. $$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \leq \sum_{t=1}^{T} \frac{\eta_t}{2} K^{\alpha}.$$

2. *If further $\eta_t \leq \frac{1}{4}$, then for any fixed j:*
$$\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \leq \frac{\eta_t}{2}\left(K^{\alpha/2} + c(\alpha) + 1\right) \sum_{i \neq j} \mathbb{E}\left[p_{t,i}\right]^{1-\alpha/2},$$
where $c(\alpha) = \left(1 - \frac{(1-\alpha)}{4}\right)^{-\frac{2-\alpha}{1-\alpha}}$.

We present a proof of the lemma in Appendix B.2. Note that for $\alpha \in (0,1)$, we have $c(\alpha) \in [1,2]$.

Lemma 7 *For any $\alpha \in (0,1)$ and non-increasing positive learning rate sequence η_t, the penalty term of the regret bound of Decoupled-Tsallis-INF satisfies:*

1. $$\mathbb{E}\left[\sum_{t=1}^{T} \Phi_t(-\tilde{L}_{t-1}) - \Phi_t(-\tilde{L}_t) - \ell_{t,i_T^*}\right] \leq \frac{(K^{1-\alpha} - 1)(1 - T^{-\alpha})}{(1-\alpha)\alpha\eta_T} + 1.$$

2. *Furthermore, if $\eta_t = \frac{2\beta}{\sqrt{t}}$ for some $\beta > 0$, then the penalty further satisfies:*
$$\mathbb{E}\left[\sum_{t=1}^{T} \Phi_t(-\tilde{L}_{t-1}) - \Phi_t(-\tilde{L}_t) - \ell_{t,i_T^*}\right] \leq \frac{1}{4\alpha(1-\alpha)\beta} \sum_{i \neq i^*} \sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^{\alpha}}{\sqrt{t}} + \frac{K^{1-\alpha}}{2\alpha(1-\alpha)\beta}.$$

A proof of the lemma is provided in Appendix B.3.

5.1. Proof of Theorem 3

The proof of the theorem is based on application of the first parts of Lemmas 6 and 7.

Proof of Theorem 3 We use the first part of Lemma 6 to bound the stability term. We remind that the learning rate is $\eta_t = \frac{2K^{1/2-\alpha}}{\sqrt{t}}$.

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \leq \sum_{t=1}^{T} \frac{\eta_t}{2} K^{\alpha} = \sum_{t=1}^{T} \frac{K^{1/2-\alpha}}{\sqrt{t}} K^{\alpha} \leq 2\sqrt{KT}.$$

Similarly, we use the first part of Lemma 7 to bound the penalty term:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\Phi_t(-\tilde{L}_{t-1})-\Phi_t(-\tilde{L}_t)-\ell_{t,i_T^*}\right] \leq & \frac{(K^{1-\alpha}-1)(1-T^{-\alpha})}{\alpha(1-\alpha)2K^{1/2-\alpha}\sqrt{\frac{1}{T}}}+1\\ \leq & \frac{1}{2\alpha(1-\alpha)}\frac{K^{1-\alpha}}{K^{1/2-\alpha}\sqrt{\frac{1}{T}}}+1\\ = & \frac{1}{2\alpha(1-\alpha)}\sqrt{KT}+1.\end{aligned}$$

Summing the stability and the penalty terms finishes the proof.

■

5.2. Proof of Theorem 4

The following two lemmas are needed in order to take advantage of the self-bounding technique and obtain a time-independent bound. They are proven in Appendix A.

Lemma 8 *For $\alpha \in (0,1)$, $c > 0$ and $d \in (0,1]$, we have*

$$\max_{x\in[0,\infty)} cx^{\alpha}-dx = c^{\frac{1}{1-\alpha}}d^{\frac{\alpha}{\alpha-1}}\left(\alpha^{\frac{\alpha}{1-\alpha}}-\alpha^{\frac{1}{1-\alpha}}\right).$$

Lemma 9 *Let $T_0 = \max_{i\neq i^*}\left\lceil D\left(\frac{8}{\Delta_i}\right)^2\right\rceil$ for a constant $D \geq 1$. For $i \neq i^*$ define $S_i(T) = \frac{1}{\Delta_i^{-\frac{\alpha}{\alpha-1}}}\sum_{t=T_0+1}^{T}\frac{1}{t^{-\frac{1}{2(\alpha-1)}}}$. The series $S_i(T)$ converges for $T \to \infty$ if and only if $\alpha > \frac{1}{2}$. Furthermore, for $\alpha > \frac{1}{2}$, we have:*

$$\lim_{T\to\infty} S_i(T) \leq \frac{2-2\alpha}{2\alpha-1}\frac{\left(8\sqrt{D}\right)^{\frac{2\alpha-1}{\alpha-1}}}{\Delta_i}.$$

With the two lemmas at hand we move on to the proof. The proof strategy is the following. We define a time step T_0 from which we can achieve a refined upper bound for the instantaneous stability term introduced in Lemma 6. For $t \leq T_0$, the proof is the same as in the adversarial setting, which gives a contribution of order $O(\sqrt{KT_0}) = O\left(\frac{\sqrt{KD}}{\Delta_{\min}}\right)$.

Now, we focus on the part of the bound for $t > T_0$. Let B be an upper bound on the regret, $\mathcal{R}_T \leq B$. In the stochastically constrained adversarial regime we can use the alternative way of writing the regret given in equation (1) (what Zimmert and Seldin call the self-bounding property of the regret) to obtain

$$\mathcal{R}_T = 2\mathcal{R}_T - \sum_{t=1}^{T}\sum_{i\neq i^*}\mathbb{E}[p_{t,i}]\Delta_i \leq 2B - \sum_{t=1}^{T}\sum_{i\neq i^*}\mathbb{E}[p_{t,i}]\Delta_i.$$

For $t > T_0$ we derive a refined bound for instantaneous contributions to the right hand side. By using the second parts of Lemmas 6 and 7, for $\alpha \leq 2/3$ the instantaneous contributions to B can be bounded by $\sum_{i \neq i^*} C\mathbb{E}\left[p_{t,i}\right]^\alpha / \sqrt{t}$ for some constant C. The overall instantaneous contribution to the right hand side is then bounded by $\sum_{i \neq i^*} \left(2C\mathbb{E}\left[p_{t,i}\right]^\alpha / \sqrt{t} - \mathbb{E}[p_{t,i}]\Delta_i\right)$. By taking $x_i = \mathbb{E}\left[p_{t,i}\right]$ and using Lemma 8 we then bound the instantaneous contributions by

$$\sum_{i \neq i^*} \left(\frac{2Cx_i^\alpha}{\sqrt{t}} - \Delta_i x_i\right) \leq \sum_{i \neq i^*} \max_{x_i \in [0,\infty)} \left(\frac{2Cx_i^\alpha}{\sqrt{t}} - \Delta_i x_i\right) \leq \sum_{i \neq i^*} C' \Delta_i^{\frac{\alpha}{\alpha-1}} t^{\frac{1}{2(\alpha-1)}}$$

for some other constant C'. Note that the bound is meaningful only for $\Delta_i > 0$. This is why we need the assumption on uniqueness of the best arm. Summing the instantaneous contributions over t from T_0 to T completes the proof. The sum of the series of instantaneous contributions converges if and only if $\frac{1}{2(1-\alpha)} > 1$, which means that the bound is time-independent if and only if $\alpha > \frac{1}{2}$.

Proof of Theorem 4 We bound the stability and the penalty terms. Concerning the stability term, we want to use the second part of Lemma 6 when t is large enough. We choose the threshold $T_0 = \max_{i \neq i^*} \left\lceil D\left(\frac{8}{\Delta_i}\right)^2 \right\rceil \geq 64$ for some $D \geq 1$. This choice allows us to use the second part of Lemma 6 because for all $t > T_0$ we have $\eta_t = \frac{2K^{1/2-\alpha}}{\sqrt{t}} \leq \frac{2}{\sqrt{t}} \leq \frac{1}{4}$. We use the second part of Lemma 6 with $j = i^*$ for $t > T_0$ and the first part of Lemma 6 for $t \leq T_0$. We have:

$$\begin{aligned}
\text{stability} = &\,\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \\
= &\sum_{t=1}^{T_0} \mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] + \sum_{t=T_0+1}^{T} \mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \\
\leq &\sum_{t=1}^{T_0} \frac{K^{1/2-\alpha}}{\sqrt{t}} K^\alpha + \sum_{t=T_0+1}^{T} \frac{K^{1/2-\alpha}}{\sqrt{t}} \left(K^{\alpha/2} + c(\alpha) + 1\right) \left(\sum_{i \neq i^*} \mathbb{E}\left[p_{t,i}\right]^{1-\alpha/2}\right) \\
\leq &\, 2\sqrt{KT_0} + \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \frac{\left(\frac{c(\alpha)+1}{2^{\alpha/2}} + 1\right) K^{1/2-\alpha/2}}{\sqrt{t}} \mathbb{E}\left[p_{t,i}\right]^{1-\alpha/2},
\end{aligned}$$

where we used the inequalities $\sum_{t=1}^{T_0} \frac{1}{\sqrt{t}} \leq 2\sqrt{T_0}$ and $K^{\alpha/2} \geq 2^{\alpha/2}$.

To bound the penalty term, we use the second part of Lemma 7 with $\beta = K^{1/2-\alpha}$.

$$\begin{aligned}
\text{penalty} \leq &\frac{1}{4\alpha(1-\alpha)K^{1/2-\alpha}} \sum_{i \neq i^*} \sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}} + \frac{\sqrt{K}}{2\alpha(1-\alpha)} \\
\leq &\left(\frac{K^{\alpha-1/2}}{4\alpha(1-\alpha)} \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}}\right) + \left(\frac{K^{\alpha-1/2}}{4\alpha(1-\alpha)} \sum_{i \neq i^*} \sum_{t=1}^{T_0} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}}\right) + \frac{\sqrt{K}}{2\alpha(1-\alpha)} \\
\leq &\left(\frac{K^{\alpha-1/2}}{4\alpha(1-\alpha)} \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}}\right) + \left(\frac{K^{\alpha-1/2}}{4\alpha(1-\alpha)} \sum_{t=1}^{T_0} \frac{K^{1-\alpha}}{\sqrt{t}}\right) + \frac{\sqrt{K}}{2\alpha(1-\alpha)} \\
\leq &\left(\frac{K^{\alpha-1/2}}{4\alpha(1-\alpha)} \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}}\right) + \frac{\sqrt{KT_0}}{2\alpha(1-\alpha)} + \frac{\sqrt{K}}{2\alpha(1-\alpha)},
\end{aligned}$$

where we used the inequalities $\sum_{i\neq i*} \mathbb{E}\left[p_{t,i}\right]^\alpha \leq (K-1)\left(\frac{1}{K-1}\right)^\alpha = (K-1)^{1-\alpha} \leq K^{1-\alpha}$ and $\sum_{t=1}^{T_0} \frac{1}{\sqrt{t}} \leq 2\sqrt{T_0}$.

We make two observations regarding the powers: for $\alpha \leq 2/3$ we have $1/2 - \alpha/2 \geq \alpha - 1/2$, so for all $K \geq 2$ it holds that $K^{\alpha-1/2} \leq K^{1/2-\alpha/2}$. Furthermore, for $\alpha \leq 2/3$ we have $\alpha \leq 1 - \alpha/2$, and for all $t \in [T]$ and $i \in [K]$ we have $\mathbb{E}[p_{t,i}] \leq 1$ and $\mathbb{E}[p_{t,i}]^{1-\alpha/2} \leq \mathbb{E}[p_{t,i}]^\alpha$. Thus, we have:

$$\text{stability} \leq 2\sqrt{KT_0} + \sum_{i\neq i^*} \sum_{t=T_0+1}^{T} \frac{\left(\frac{c(\alpha)+1}{2^{\alpha/2}} + 1\right) K^{1/2-\alpha/2}}{\sqrt{t}} \mathbb{E}\left[p_{t,i}\right]^\alpha$$

and

$$\text{penalty} \leq \left(\frac{K^{1/2-\alpha/2}}{4\alpha(1-\alpha)} \sum_{i\neq i^*} \sum_{t=T_0+1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}} \right) + \frac{\sqrt{KT_0}}{2\alpha(1-\alpha)} + \frac{\sqrt{K}}{2\alpha(1-\alpha)}.$$

We combine the two bounds and use the alternative way of writing the regret in the stochastically constrained adversarial regime (the self-bounding technique) to obtain

$$\begin{aligned}
\mathcal{R}_T &= 2\mathcal{R}_T - \sum_{i\neq i*} \sum_{t=1}^{T} \mathbb{E}[p_{t,i}]\Delta_i \\
&\leq \sum_{i\neq i^*} \sum_{t=T_0+1}^{T} \left(\left(\frac{1}{2\alpha(1-\alpha)} + \frac{c(\alpha)+1}{2^{(\alpha/2)-1}} + 2 \right) \frac{K^{1/2-\alpha/2}}{\sqrt{t}} \mathbb{E}[p_{t,i}]^\alpha - \mathbb{E}[p_{t,i}]\Delta_i \right) \\
&\quad + \left(\frac{1}{\alpha(1-\alpha)} + 4 \right) \sqrt{KT_0} + \frac{\sqrt{K}}{\alpha(1-\alpha)} \\
&\leq \sum_{i\neq i^*} \sum_{t=T_0+1}^{T} \max_{x\in[0,\infty)^K} \left(\left(\frac{1}{2\alpha(1-\alpha)} + \frac{c(\alpha)+1}{2^{(\alpha/2)-1}} + 2 \right) \frac{K^{1/2-\alpha/2}}{\sqrt{t}} x^\alpha - x\Delta_i \right) \\
&\quad + \left(\frac{1}{\alpha(1-\alpha)} + 4 \right) \sqrt{KT_0} + \frac{\sqrt{K}}{\alpha(1-\alpha)}.
\end{aligned}$$

In the last step we take $x = \mathbb{E}\left[p_{t,i}\right]$ and drop the constraint that p_t is a probability distribution. Using Lemma 8, for any $i \neq i^*$ and $t > T_0$, we have

$$\begin{aligned}
&\max_{x\in[0,\infty)} \left(\left(\frac{1}{2\alpha(1-\alpha)} + \frac{c(\alpha)+1}{2^{(\alpha/2)-1}} + 2 \right) \frac{K^{1/2-\alpha/2}}{\sqrt{t}} x^\alpha - x\Delta_i \right) \\
&\qquad\qquad \leq \Delta_i^{\frac{\alpha}{\alpha-1}} \frac{\sqrt{K}}{t^{-\frac{1}{2(\alpha-1)}}} \left(\frac{1}{2\alpha(1-\alpha)} + \frac{c(\alpha)+1}{2^{(\alpha/2)-1}} + 2 \right)^{\frac{1}{1-\alpha}} \left(\alpha^{\frac{\alpha}{1-\alpha}} - \alpha^{\frac{1}{1-\alpha}} \right).
\end{aligned}$$

Using $\widetilde{C}(\alpha) = \left(\frac{1}{2\alpha(1-\alpha)} + \frac{c(\alpha)+1}{2^{(\alpha/2)-1}} + 2 \right)^{\frac{1}{1-\alpha}} \left(\alpha^{\frac{\alpha}{1-\alpha}} - \alpha^{\frac{1}{1-\alpha}} \right)$, we can incorporate this result in the regret bound and deduce that:

$$\mathcal{R}_T \leq \sum_{i\neq i^*} \sum_{t=T_0+1}^{T} \left(\widetilde{C}(\alpha) \Delta_i^{\frac{\alpha}{\alpha-1}} \frac{\sqrt{K}}{t^{-\frac{1}{2(\alpha-1)}}} \right) + \left(\frac{1}{\alpha(1-\alpha)} + 4 \right) \sqrt{KT_0} + \frac{\sqrt{K}}{\alpha(1-\alpha)},$$

which gives the first statement of the lemma. Finally, we use Lemma 9 to deduce that the bound is time-independent if and only if $\alpha > 1/2$. Furthermore, by definition of T_0 we have $\sqrt{T_0} \leq \frac{8\sqrt{D}}{\Delta_{\min}}+1$.

We deduce that for $\alpha \in (1/2, 2/3]$, the pseudo-regret is upper bounded as:

$$\begin{aligned}\mathcal{R}_T &\leq \sum_{i \neq i^*} \left(\widetilde{C}(\alpha) \frac{2-2\alpha}{2\alpha-1} 8^{\frac{2\alpha-1}{\alpha-1}} \frac{\sqrt{K} D^{\frac{2\alpha-1}{2\alpha-2}}}{\Delta_i} \right) + \left(\frac{1}{\alpha(1-\alpha)} + 4 \right) \sqrt{K T_0} + \frac{\sqrt{K}}{\alpha(1-\alpha)} \\ &\leq \sum_{i \neq i^*} \left(C(\alpha) \frac{\sqrt{K} D^{\frac{2\alpha-1}{2\alpha-2}}}{\Delta_i} \right) + 68 \frac{\sqrt{KD}}{\Delta_{\min}} + 13\sqrt{K},\end{aligned}$$

where $C(\alpha) = \widetilde{C}(\alpha) \frac{2-2\alpha}{2\alpha-1} 8^{\frac{2\alpha-1}{\alpha-1}}$. This gives the second statement of the theorem. ■

6. Discussion

We have derived an algorithm for the problem of decoupled exploration and exploitation in multi-armed bandits. We have shown that it achieves the minimax optimal $O(\sqrt{KT})$ regret bound in the adversarial regime and simultaneously a time-independent $O\left(\sqrt{\frac{K}{\Delta_{min}} \sum_{i \neq i^*} \frac{1}{\Delta_i}}\right)$ regret bound in the stochastically constrained adversarial regime, where Δ_{min} is the minimal positive gap. The results improve on the work of Avner et al. (2012) in both regimes without requiring prior knowledge of the regime.

We conjecture that the regret bound in the stochastically constrained adversarial regime can be improved further and that the dependence on $\frac{K}{\Delta_{min}}$ can be replaced with the refined complexity measure $\sum_{i:\Delta_i>0} \frac{1}{\Delta_i}$. This research direction and derivation of matching lower bounds is left for future work.

As we have mentioned, the decoupled setting is an important bridge between full information and bandit problems, as well as a bridge between pure exploration and exploration-exploitation trade-off. An interesting direction for future research would be to use our techniques to improve results along these two directions. One possibility is to apply our regularization and exploration technique to tighten best of both worlds guarantees for prediction with limited advice and bandits with paid observations (Seldin et al., 2014; Thune and Seldin, 2018). Another direction is to explore the relations with pure exploration problems. Abbasi-Yadkori et al. (2018) have shown that in the pure exploration setting it is impossible to achieve simultaneous optimality in both adversarial and stochastic settings. An interesting question is whether the decoupled formulation can be used to reformulate the objective and to achieve some alternative results there.

Acknowledgments

We would like to thank Julian Zimmert for insightful discussions and helpful comments. We acknowledge support by the Independent Research Fund Denmark, grant number 9040-00361B.

References

Yasin Abbasi-Yadkori, Peter Bartlett, Victor Gabillon, Alan Malek, and Michal Valko. Best of both worlds: Stochastic & adversarial best-arm identification. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2018.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2009.

Peter Auer and Chao-Kai Chiang. 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.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1), 2002b.

Orly Avner, Shie Mannor, and Ohad Shamir. Decoupling exploration and exploitation in multiarmed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multiarmed bandit problems. *Foundations and Trends in Machine Learning*, 5, 2012.

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

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuously-armed bandits. *Theoretical Computer Science*, 412, 2011.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7, 2006.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6, 1985.

Tor Lattimore and Csaba Svepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5, 2004.

Jaouad Mourtada and Stéphane Gaïffas. On the optimality of the hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20, 2019.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

Ralph Tyrell Rockafellar. *Convex analysis*. Princeton University Press, 1970.

Yevgeny Seldin and Gábor Lugosi. 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.

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

Yevgeny Seldin, Peter L. Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2), 2012.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 12, 2019.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25, 1933.

Tobias Sommer Thune and Yevgeny Seldin. Adaptation to easy data in prediction with limited advice. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 1988.

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

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019. https://arxiv.org/pdf/1807.07623.

Appendix A. Proofs of Auxiliary Lemmas

Proof of Lemma 8 Let $f(x) = cx^\alpha - dx$. Then $f'(x) = \alpha c x^{\alpha-1} - d$ and $f''(x) = \alpha(\alpha-1)cx^{\alpha-2} \leq 0$. Thus, the solution of $f'(x) = 0$ gives the maximum of f.

$$f'(\tilde{x}) = 0 \Leftrightarrow \alpha c \tilde{x}^{\alpha-1} = d \Leftrightarrow \tilde{x} = \left(\frac{d}{\alpha c}\right)^{\frac{1}{\alpha-1}}.$$

Finally, we calculate $f(\tilde{x})$.

$$\max_{x\in[0,\infty)} cx^\alpha - dx = f(\tilde{x}) = c\left(\frac{d}{\alpha c}\right)^{\frac{\alpha}{\alpha-1}} - d\left(\frac{d}{\alpha c}\right)^{\frac{1}{\alpha-1}} = c^{\frac{1}{1-\alpha}} d^{\frac{\alpha}{\alpha-1}} \left(\alpha^{\frac{\alpha}{1-\alpha}} - \alpha^{\frac{1}{1-\alpha}}\right).$$

■

Proof of Lemma 9 Consider the series $S_i(T) = \sum_{t=T_0+1}^{T} \frac{1}{t^{-\frac{1}{2(\alpha-1)}}}$. We first show that when $\alpha > 1/2$, the series converges and upper bound its limit. Then, we show that the series diverges when $\alpha \leq 1/2$.

When $\alpha > 1/2$, we have $\frac{-1}{2(\alpha-1)} > 1$ so the Riemann's series $\sum_{t=1}^{\infty} \frac{1}{t^{-\frac{1}{2(\alpha-1)}}}$ converges. This is an upper bound on $s(T)$, which converges as well. However, when we derive the upper bound on $\lim_{T\to\infty} s(T)$, we want to take advantage of the fact that we sum for $t \geq T_0 + 1$. We have:

$$\begin{aligned}\lim_{T\to\infty} \sum_{t=T_0+1}^{T} \frac{1}{t^{-\frac{1}{2(\alpha-1)}}} &\leq \lim_{T\to\infty} \int_{T_0}^{T} \frac{1}{t^{-\frac{1}{2(\alpha-1)}}} dt \\ &= \lim_{T\to\infty} \frac{T^{1+\frac{1}{2(\alpha-1)}} - T_0^{1+\frac{1}{2(\alpha-1)}}}{1+\frac{1}{2(\alpha-1)}} \\ &\leq \frac{-T_0^{1+\frac{1}{2(\alpha-1)}}}{1+\frac{1}{2(\alpha-1)}},\end{aligned}$$

where in the last step we use the fact that $1 + \frac{1}{2(\alpha-1)} = \frac{2\alpha-1}{2\alpha-2}$ is negative. This also implies that as $T_0 \geq D\left(\frac{8}{\Delta_{min}}\right)^2$, we can upper bound $T_0^{1+\frac{1}{2(\alpha-1)}}$ by $\left(D\left(\frac{8}{\Delta_{min}}\right)^2\right)^{\frac{2\alpha-1}{2\alpha-2}} = \left(\frac{8\sqrt{D}}{\Delta_{min}}\right)^{\frac{2\alpha-1}{\alpha-1}} \leq \left(\frac{8\sqrt{D}}{\Delta_i}\right)^{\frac{2\alpha-1}{\alpha-1}}$ for any $i \neq i^*$, because $\frac{2\alpha-1}{2(\alpha-1)} \leq 0$ for $\alpha > 1/2$. Incorporating this result in $S_i(T)$ finishes this part of the proof.

For the second part of the proof, for $\alpha \leq 1/2$, we have:

$$\begin{aligned}\sum_{t=T_0+1}^{T} \frac{1}{t^{-\frac{1}{2(\alpha-1)}}} &\geq \sum_{t=T_0+1}^{T} \frac{1}{t} \\ &\geq \int_{T_0+1}^{T+1} \frac{1}{t} dt \\ &= \log(T+1) - \log(T_0+1),\end{aligned}$$

which diverges, because T_0 is a constant. Thus, for any $\alpha \in (0, 1/2]$, we cannot obtain a time-independent upper bound on $S_i(T)$. ■

Appendix B. Analysis in the Follow the Regularized Leader Framework

We first introduce some tools needed to work in the Follow the Regularized Leader framework and then derive bounds on the stability and the penalty terms.

B.1. Follow the Regularized Leader Framework and Tsallis Entropy

Follow the Regularized Leader (FTRL) has been widely used in online learning in the past few years. We use Tsallis entropy as our regularizer, defined as

$$\Psi_t(w) = -\frac{1}{\eta_t}\sum_i \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)}, \tag{5}$$

and its convex conjugate, defined as

$$\Psi_t^*(y) = \max_{x\in\mathbb{R}^K}\left\{\langle x, y\rangle - \Psi_t(x)\right\} = \max_{x\in\mathbb{R}^K}\left\{\langle x, y\rangle + \frac{1}{\eta_t}\sum_i \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)}\right\}.$$

We let Δ^{K-1} denote the probability simplex over K vertices and define $\mathcal{I}_{\Delta^{K-1}}(x) = \begin{cases} 0, & \text{if } x \in \Delta^{K-1} \\ \infty, & \text{otherwise}\end{cases}$. Using results from convex analysis (Rockafellar, 1970), Ψ_t is a convex differentiable function with an invertible gradient $(\nabla\Psi)^{-1}$, and we have

$$\nabla(\Psi_t + \mathcal{I}_{\Delta^{K-1}})^*(y) = \underset{x\in\Delta^{K-1}}{\operatorname{argmax}}\left\{\langle x, y\rangle - \Psi_t(x)\right\}.$$

Note that $\nabla(\Psi + \mathcal{I}_{\Delta^{K-1}})^*(y) \in \Delta^{K-1}$. We define the potential function Φ_t as

$$\Phi_t(-L) = (\Psi_t + \mathcal{I}_{\Delta^{K-1}})^*(-L) = \max_{w\in\Delta^{K-1}}\left\{\langle w, -L\rangle + \frac{1}{\eta_t}\sum_{i=1}^K \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)}\right\}.$$

Φ_t is a restriction of Ψ_t^* to the probability simplex. The weights p_t of the Decoupled-Tsallis-INF algorithm fulfil:

$$\begin{aligned} p_t = \nabla\Phi_t(-\tilde{L}_{t-1}) &= \arg\max_{p\in\Delta^{K-1}}\left\{\left\langle p, -\tilde{L}_{t-1}\right\rangle + \frac{1}{\eta_t}\sum_{i=1}^K \frac{p_i^\alpha - \alpha p_i}{\alpha(1-\alpha)}\right\} \\ &= \arg\min_{p\in\Delta^{K-1}}\left\{\left\langle p, \tilde{L}_{t-1}\right\rangle - \frac{1}{\eta_t}\sum_{i=1}^K \frac{p_i^\alpha - \alpha p_i}{\alpha(1-\alpha)}\right\}. \end{aligned}$$

Ψ_t^* is the unconstrained definition of Φ_t. Thus there exists a Lagrange multiplier ν such that

$$p_t = \nabla\Psi_t^*(-\hat{L}_{t-1} + \nu 1_K),$$

where we use 1_K to denote a K-dimensional vector of ones. Using the fact that Ψ_t is a Legendre function (Rockafellar, 1970), which implies that it is gradient inversible and $\nabla\Psi_t^{-1} = \nabla\Psi_t^*$ we deduce that:

$$\nabla\Psi_t(p_t) = -\hat{L}_{t-1} + \nu 1_K. \tag{6}$$

B.2. Analysis of the Stability term

The analysis of the stability term is based on tools developed by Zimmert and Seldin (2019). We note that $\tilde{\ell}_{t,A_t}$ is an unbiased estimate of ℓ_{t,A_t} and

$$\mathbb{E}\left[\ell_{t,A_t}\right] = \mathbb{E}\left[\left\langle p_t, \tilde{\ell}_t\right\rangle\right]. \tag{7}$$

The following result is analogous to the result of Zimmert and Seldin (2019, Lemma 11). We use 1_K to denote a K-dimensional vector of ones.

Lemma 10 *Let $p_t = \nabla\Phi_t(-\tilde{L}_{t-1})$ for $\tilde{L}_t = \tilde{L}_{t-1} + \tilde{\ell}_t$, where $\tilde{\ell}_t$ is an unbiased estimate of ℓ_t. For any $x \in [0, \infty)$, the instantaneous stability of the pseudo-regret of Algorithm 1 satisfies*

$$\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \leq \mathbb{E}\left[\sum_{i=1}^{K} \max_{\tilde{p}_i \in [p_{t,i}, \nabla\Psi^*(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x1_K)_i]} \frac{\eta_t}{2}(\tilde{\ell}_{t,i} - x)^2 (\tilde{p}_i)^{2-\alpha}\right].$$

Proof of Lemma 10 Note that $\Phi_t(L+x\mathbf{1}_K) = \Phi(L)+x$ since we take the argmax over a probability distribution.

$$\begin{align}
&\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \nonumber\\
&\qquad = \mathbb{E}\left[\left\langle p_t, \tilde{\ell}_t \right\rangle + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \nonumber\\
&\qquad = \mathbb{E}\left[\left\langle p_t, \tilde{\ell}_t \right\rangle + \Phi_t(\nabla\Psi_t(p_t) - \tilde{\ell}_t) - \Phi_t(\nabla\Psi_t(p_t))\right] \nonumber\\
&\qquad = \mathbb{E}\left[\left\langle p_t, \tilde{\ell}_t - x\mathbf{1}_K \right\rangle + \Phi_t(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K) - \Phi_t(\nabla\Psi_t(p_t))\right] \nonumber\\
&\qquad \leq \mathbb{E}\left[\left\langle p_t, \tilde{\ell}_t - x\mathbf{1}_K \right\rangle + \Psi_t^*(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K) - \Psi_t^*(\nabla\Psi_t(p_t))\right] \tag{8}\\
&\qquad \leq \mathbb{E}\left[D_{\Psi_t^*}(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K, \nabla\Psi_t(p_t))\right], \nonumber
\end{align}$$

where the first step follows by Equation (7). Then, we used the fact that $-\tilde{L}_{t-1} + \nu 1_K = \nabla\Psi_t(p_t)$ for some constant ν using Equation (6), and the definition of Φ_t to add $x\mathbf{1}_K$. Finally, we recall that Ψ_t^* is the unrestricted version of Φ_t. On step 8, we recognize the Bregman divergence of Ψ_t, and using Taylor's expansion, there is some $z \in \text{conv}(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K, \nabla\Psi_t(p_t))$ such that

$$D_{\Psi_t^*}(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K, \nabla\Psi_t(p_t)) = \frac{1}{2}\|\tilde{\ell}_t - x\mathbf{1}_K\|^2_{\nabla^2\Psi_t(z)}.$$

We deduce that:

$$\begin{aligned}
&\mathbb{E}\left[D_{\Psi_t^*}(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K, \nabla\Psi_t(p_t))\right]\\
&\qquad \leq \mathbb{E}\left[\max_{z \in \text{conv}(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K, \nabla\Psi_t(p_t))} \frac{1}{2}\|\tilde{\ell}_t - x\mathbf{1}_K\|^2_{\nabla^2\Psi_t(z)}\right]\\
&\qquad \leq \mathbb{E}\left[\sum_{i=1}^{K} \max_{\tilde{p}_i \in [p_{t,i}, \nabla\Psi^*(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x1_K)_i]} \frac{\eta_t}{2}(\tilde{\ell}_{t,i} - x)^2 (\tilde{p}_i)^{2-\alpha}\right],
\end{aligned}$$

where we used the fact that $\nabla\Psi_t(p_t)$ is in the probability simplex so $\Phi_t(\nabla\Psi_t(p_t)) = \Psi_t^*(\nabla\Psi_t(p_t))$, and finally the fact that $\nabla^2\Psi_t(p) = \text{diag}\left(\frac{p_i^{\alpha-2}}{\eta_t}\right)_{i=1,\dots,K}$. ■

Using the stability analysis in Lemma 10 we move on to bound the stability term. First, we focus on bounding the stability term when the distribution to query the exploration arm is arbitrary.
Proof of Lemma 1 We start by bounding the instantaneous stability at a fixed round t. We start from Lemma 10 with $x = 0$. We recall that $\nabla\Psi^*(\nabla\Psi_t(p_t) - \tilde{\ell}_t) \leq \nabla\Psi^*(\nabla\Psi_t(p_t)) = p_t$, because

the losses are non-negative and $\nabla\Psi_t^*$ is monotonically increasing.

$$\begin{aligned}\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] &\leq \mathbb{E}\left[\sum_{i=1}^K \frac{\eta_t}{2}(\tilde{\ell}_{t,i})^2(p_{t,i})^{2-\alpha}\right] \\ &= \mathbb{E}\left[\sum_{i=1}^K \frac{\eta_t}{2}\frac{(\ell_{t,i})^2}{q_{t,i}}(p_{t,i})^{2-\alpha}\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^K \frac{\eta_t}{2}\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right],\end{aligned}$$

where we used the definition of $\tilde{\ell}_t$ and that $\mathbb{E}\left[\mathbb{1}\left\{B_t = i\right\}\right] = q_{t,i}$. The last steps relies on the fact that the losses are bounded in the $[0,1]$ interval. Finally, summing for $t = 1$ to T finishes the proof. ■

In order to derive a proof of Lemma 6, we first need to bound the weights estimators $\tilde{p}$ in Lemma 10. The proof follows from Zimmert and Seldin (2019, Proof of Lemma 16).

Lemma 11 *Let $p \in \Delta^{K-1}$ and $\tilde{p} = \nabla\Psi_t^*(\nabla\Psi_t(p) - \ell)$. If $\eta_t \leq 1/4$, then for all $\ell_i \geq -1$ it holds that $\tilde{p}_i^{2-\alpha} \leq c(\alpha)p_i^{2-\alpha}$, where $c(\alpha) = (1 - \frac{(1-\alpha)}{4})^{-\frac{2-\alpha}{1-\alpha}}$.*

Proof of Lemma 11 $\nabla\Psi_t$ is the inverse of $\nabla\Psi_t^*$, which gives

$$\nabla\Psi_t(\tilde{p}) = \nabla\Psi_t(p) - \ell.$$

Using our lower bound on ℓ, we deduce that in each dimension:

$$\begin{aligned}&\nabla\Psi_t(p)_i - \nabla\Psi_t(\tilde{p})_i = \ell_i \geq -1, \\ &\frac{p_i^{\alpha-1} - 1}{(1-\alpha)\eta_t} - \frac{\tilde{p}_i^{\alpha-1} - 1}{(1-\alpha)\eta_t} \leq 1, \\ &\tilde{p}_i^{1-\alpha} \leq \frac{p_i^{1-\alpha}}{1 - \eta_t(1-\alpha)p_i^{1-\alpha}} \leq \frac{p_i^{1-\alpha}}{1 - \eta_t(1-\alpha)}, \\ &\tilde{p}_i^{2-\alpha} \leq \frac{p_i^{2-\alpha}}{(1 - \eta_t(1-\alpha))^{\frac{2-\alpha}{1-\alpha}}}.\end{aligned}$$

Now we need to upper bound $(1 - \eta_t(1-\alpha))^{-\frac{2-\alpha}{1-\alpha}}$. This function is monotonically decreasing in η_t for all $\alpha \in [0,1]$. Therefore,

$$(1 - \eta_t(1-\alpha))^{-\frac{2-\alpha}{1-\alpha}} \leq \left(1 - \frac{(1-\alpha)}{4}\right)^{-\frac{2-\alpha}{1-\alpha}} = c(\alpha).$$

■

Using the above results and Lemma 2, we move on to the proof of Lemma 6. The first part of the lemma is a direct application of Lemma 1.

Proof of Lemma 6

The first statement of the Lemma Using Lemma 1 and the distribution given in Lemma 2, we can bound the instantaneous stability at round t by:

$$\begin{aligned}\mathbb{E}\left[\ell_{t,A_t}+\Phi_t(-\tilde{L}_t)-\Phi_t(-\tilde{L}_{t-1})\right] &\leq \mathbb{E}\left[\sum_{i=1}^{K}\frac{\eta_t}{2}\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right]\\ &=\mathbb{E}\left[\sum_{i=1}^{K}\frac{\eta_t}{2}\frac{(p_{t,i})^{2-\alpha}}{(p_{t,i})^{1-\alpha/2}}\sum_{j=1}^{K}(p_{t,j})^{1-\alpha/2}\right]\\ &=\mathbb{E}\left[\frac{\eta_t}{2}\left(\sum_{i=1}^{K}(p_{t,i})^{1-\alpha/2}\right)^2\right].\end{aligned}$$

We can upper bound the expectation by replacing p_t with a distribution which maximizes the expression. Because $f(x)=x^2$ is an increasing function for $x\in\mathbb{R}^+$, this expression is maximized when $\sum_{i=1}^{K}(p_{t,i})^{1-\alpha/2}$ is maximized. Since $1-\alpha/2\leq 1$, the expression is maximized by the uniform distribution, which gives

$$\begin{aligned}\mathbb{E}\left[\ell_{t,A_t}+\Phi_t(-\tilde{L}_t)-\Phi_t(-\tilde{L}_{t-1})\right] &\leq \mathbb{E}\left[\frac{\eta_t}{2}\left(\sum_{i=1}^{K}K^{-1+\alpha/2}\right)^2\right]\\ &\leq \mathbb{E}\left[\frac{\eta_t}{2}\left(K^{\alpha/2}\right)^2\right]\\ &\leq \frac{\eta_t}{2}K^{\alpha}.\end{aligned}$$

Summing over t finishes the proof of the first part of the lemma.

The second statement of the Lemma We move on to the second part of the lemma. This time, we start from Lemma 10, where we choose $x=\mathbb{1}_t\left[B_t=j\right]\ell_{t,j}$. The analysis depends on whether $B_t\neq j$ or $B_t=j$. In the first case, we have $x=0$, and the expression is maximized by $\tilde{p}_i=p_{t,i}$, because the losses are non-negative, and $\nabla\Psi_t^*$ is monotonically increasing. In the second case, we have $B_t=j$, which means that for $i\neq j$ we have $\tilde{\ell}_{t,i}-x=0-x\geq -1$ and we can apply Lemma 11 and bound $\tilde{p}_i^{2-\alpha}$ by $c(\alpha)p_i^{2-\alpha}$, where $c(\alpha)=\left(1-\frac{(1-\alpha)}{4}\right)^{-\frac{2-\alpha}{1-\alpha}}$. Otherwise, we have $\tilde{\ell}_{t,j}-x\geq 0$, and using the fact that $\nabla\Psi_t^*$ is monotonically increasing we deduce that $\tilde{p}_j=p_{t,j}$. Combining all the cases, we have:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{K} \max_{\tilde{p}\in[p_{t,i},\nabla\Psi^*(\nabla\Psi_t(p_t)-\tilde{\ell}_t+x1_K)_i]} \frac{\eta_t}{2}(\tilde{\ell}_{t,i}-x)^2(\tilde{p}_i)^{2-\alpha}\right] \\
&\quad\le \sum_{i\neq j}\frac{\eta_t}{2}\mathbb{E}\left[\mathbb{1}\left[B_t=i\right]\sum_{k=1}^{K}(\tilde{\ell}_{t,k})^2(p_{t,k})^{2-\alpha}\right] \\
&\qquad\quad + \mathbb{E}\left[\mathbb{1}\left[B_t=j\right]\left(\left(\sum_{i\neq j}\frac{\eta_t}{2}(\tilde{\ell}_{t,i}-\ell_{t,j})^2 c(\alpha)(p_{t,i})^{2-\alpha}\right)+\frac{\eta_t}{2}(\tilde{\ell}_{t,j}-\ell_{t,j})^2(p_{t,j})^{2-\alpha}\right)\right] \\
&\quad\le \sum_{i\neq j}\frac{\eta_t}{2}\mathbb{E}\left[q_{t,i}\left(\frac{\ell_{t,i}}{q_{t,i}}\right)^2\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right] \\
&\qquad\quad + \mathbb{E}\left[\left(\sum_{i\neq j}\frac{\eta_t c(\alpha)}{2}(\ell_{t,j})^2(p_{t,i})^{2-\alpha}q_{t,j}\right)+\frac{\eta_t}{2}\left(\frac{\ell_{t,j}}{q_{t,j}}-\ell_{t,j}\right)^2(p_{t,j})^{2-\alpha}q_{t,j}\right],
\end{aligned}
$$

where in both inequalities the first term concerns the cases where $B_t \neq j$, and the second term the case where $B_t = j$. In the second term, the index j is taken out of the sum as it is treated differently. Continuing the derivation above, we have

$$
\begin{aligned}
&\le \sum_{i\neq j}\frac{\eta_t}{2}\mathbb{E}\left[\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right] \\
&\qquad + \mathbb{E}\left[\left(\sum_{i\neq j}\frac{\eta_t c(\alpha)}{2}(\ell_{t,j})^2(p_{t,i})^{2-\alpha}q_{t,j}\right)+\frac{\eta_t}{2}\frac{1}{(q_{t,j})^2}(\ell_{t,j})^2(1-q_{t,j})^2(p_{t,j})^{2-\alpha}q_{t,j}\right] \\
&\le \sum_{i\neq j}\frac{\eta_t}{2}\mathbb{E}\left[\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right]+\mathbb{E}\left[\left(\sum_{i\neq j}\frac{\eta_t c(\alpha)}{2}(p_{t,i})^{2-\alpha}q_{t,j}\right)+\frac{\eta_t}{2}\frac{(p_{t,j})^{2-\alpha}}{q_{t,j}}(1-q_{t,j})^2\right].
\end{aligned}
$$

Now we consider each term separately. We can replace q_t by its definition from Lemma 2, $\forall i \in [K], q_{t,i} = \frac{(p_{t,i})^{1-\alpha/2}}{\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}}$. Note that $1-\alpha/2 \le 1$, so $\sum_{j=1}^{K}(p_{t,j})^{1-\alpha/2} \le K\left(\frac{1}{K}\right)^{1-\alpha/2} = K^{\alpha/2}$. In the previous expression, the first term is bounded as:

$$
\sum_{i\neq j}\frac{\eta_t}{2}\mathbb{E}\left[\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right] \le \sum_{i\neq j}\frac{\eta_t}{2}\mathbb{E}\left[(p_{t,i})^{1-\alpha/2}\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}\right] \le \frac{\eta_t}{2}K^{\alpha/2}\sum_{i\neq j}\mathbb{E}\left[(p_{t,i})^{1-\alpha/2}\right].
$$

In order to bound the third term, we observe that $(1 - q_{t,j})^2 \leq 1 - q_{t,j} = \sum_{i \neq j} q_{t,i}$. This gives:

$$\begin{aligned}
\mathbb{E}\left[\frac{\eta_t}{2}\frac{(p_{t,j})^{2-\alpha}}{q_{t,j}}(1-q_{t,j})^2\right] &= \mathbb{E}\left[\frac{\eta_t}{2}(1-q_{t,j})^2(p_{t,j})^{1-\alpha/2}\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}\right] \\
&\leq \mathbb{E}\left[\frac{\eta_t}{2}\sum_{i\neq j} q_{t,i}\left(\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}\right)\right] \\
&= \mathbb{E}\left[\frac{\eta_t}{2}\sum_{i\neq j}\frac{(p_{t,i})^{1-\alpha/2}}{\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}}\left(\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}\right)\right] \\
&= \mathbb{E}\left[\frac{\eta_t}{2}\sum_{i\neq j}(p_{t,i})^{1-\alpha/2}\right].
\end{aligned}$$

Finally, the second term can be bounded as:

$$\mathbb{E}\left[\sum_{i\neq j}\frac{\eta_t c(\alpha)}{2}(p_{t,i})^{2-\alpha}q_{t,j}\right] \leq c(\alpha)\mathbb{E}\left[\frac{\eta_t}{2}\sum_{i\neq j}(p_{t,i})^{1-\alpha/2}\right].$$

The inequality relies on the fact that $\forall t, i \ : 0 \leq p_{t,i} \leq 1$, so $(p_{t,i})^{2-\alpha} \leq \sqrt{(p_{t,i})^{2-\alpha}} = (p_{t,i})^{1-\alpha/2}$ and that $q_{t,j} \leq 1$.

Combining the three terms and using Jensen's inequality finishes the proof of the lemma.

$$\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \leq \frac{\eta_t}{2}\left(K^{\alpha/2} + c(\alpha) + 1\right)\sum_{i\neq j}\mathbb{E}\left[p_{t,i}\right]^{1-\alpha/2}.$$

■

B.3. Analysis of the Penalty term

The analysis of the penalty term follows from the work of Zimmert and Seldin (2019, Lemma 12), which we cite below.

Lemma 12 *(Zimmert and Seldin 2019, Lemma 12) For any $\alpha \in [0,1]$ and any unbiased loss estimator, the penalty term of α-TSALLIS-INF satisfies:*

1. *For the symmetric regularizer and a non-increasing sequence of learning rates*

$$\mathbb{E}\left[\sum_{i\neq i^*}\left(\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) - \ell_{t,i_T^*}\right)\right] \leq \frac{(K^{1-\alpha}-1)(1-T^{-\alpha})}{(1-\alpha)\alpha\eta_T} + 1.$$

2. *For an arbitrary regularizer, a non-increasing sequence of learning rates, and any $x \in [1, x]$*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\neq i^*}\left(\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) - \ell_{t,i_T^*}\right)\right] \\
&\leq \frac{1-T^{-\alpha x}}{\alpha}\sum_{i\neq i_T^*}\left(\frac{\mathbb{E}\left[p_{1,i}\right]^{\alpha} - \alpha\mathbb{E}\left[p_{1,i}\right]}{\eta_1\xi_i(1-\alpha)} + \sum_{t=2}^{T}\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right)\frac{\mathbb{E}\left[p_{t,i}\right]^{\alpha} - \alpha\mathbb{E}\left[p_{t,i}\right]}{\xi_i(1-\alpha)}\right) + T^{1-x}.
\end{aligned}$$

In our case the regularizer is symmetric and $\xi_i = 1$ for all i. Now we provide a proof of Lemma 7.

Proof of Lemma 7 Algorithm 1 uses the TSALLIS-INF framework introduced by Zimmert and Seldin (2019). Furthermore, by assumption the learning rate η_t is non-increasing, and we have $\forall i \in [K],\ \xi_i = 1$, which implies that the regularizer Ψ_t that we use is symmetric. This means that both statements of Zimmert and Seldin (2019, Lemma 12) apply. The first statement of Lemma 7 follows directly from the first statement of Zimmert and Seldin (2019, Lemma 12).

Concerning the second statement, we consider the second statement of Zimmert and Seldin (2019, Lemma 12) for $x = \infty$. By assumption, we have $\eta_t = \frac{2\beta}{\sqrt{t}}$ for some constant $\beta > 0$.

$$\begin{aligned}
\text{penalty} &\leq \frac{1}{\alpha(1-\alpha)} \sum_{i \neq i^*} \left(\frac{\mathbb{E}\left[p_{1,i}\right]^\alpha}{\eta_1} + \sum_{t=2}^{T} (\eta_t^{-1} - \eta_{t-1}^{-1}) \mathbb{E}\left[p_{t,i}\right]^\alpha \right) \\
&= \frac{1}{\alpha(1-\alpha)\beta} \sum_{i \neq i^*} \left(\frac{1}{2} \mathbb{E}\left[p_{1,i}\right]^\alpha + \sum_{t=2}^{T} \frac{1}{2} (\sqrt{t} - \sqrt{t-1}) \mathbb{E}\left[p_{t,i}\right]^\alpha \right) \\
&= \frac{1}{4\alpha(1-\alpha)\beta} \sum_{i \neq i^*} \left(2\mathbb{E}\left[p_{1,i}\right]^\alpha + \sum_{t=2}^{T} (2\sqrt{t} - 2\sqrt{t-1}) \mathbb{E}\left[p_{t,i}\right]^\alpha + \sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}} - \sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}} \right) \\
&= \frac{1}{4\alpha(1-\alpha)\beta} \sum_{i \neq i^*} \left(\sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}} + \mathbb{E}\left[p_{1,i}\right]^\alpha + \sum_{t=2}^{T} \left(2\sqrt{t} - 2\sqrt{t-1} - \frac{1}{\sqrt{t}} \right) \mathbb{E}\left[p_{t,i}\right]^\alpha \right).
\end{aligned}$$

For the last two terms we use the inequality $\sum_{i \neq i^*} \mathbb{E}\left[p_{t,i}\right]^\alpha \leq (K-1) \left(\frac{1}{K-1}\right)^\alpha = (K-1)^{1-\alpha} \leq K^{1-\alpha}$. Continuing the derivation we then have

$$\begin{aligned}
\text{penalty} &\leq \frac{1}{4\alpha(1-\alpha)\beta} \left(\left(\sum_{i \neq i^*} \sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}} \right) + \left(1 + \sum_{t=2}^{T} \left(2\sqrt{t} - 2\sqrt{t-1} - \frac{1}{\sqrt{t}} \right) \right) K^{1-\alpha} \right) \\
&\leq \frac{1}{4\alpha(1-\alpha)\beta} \left(\left(\sum_{i \neq i^*} \sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}} \right) + 2K^{1-\alpha} \right) \\
&= \frac{1}{4\alpha(1-\alpha)\beta} \sum_{i \neq i^*} \sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}} + \frac{K^{1-\alpha}}{2\alpha(1-\alpha)\beta},
\end{aligned}$$

where we telescoped the sum, and used the fact that $2\sqrt{T} \leq 2 + \sum_{t=1}^{T} \frac{1}{\sqrt{t}}$. ■

Appendix C. Choice of the Exploration Distribution

The exploration distribution q_t is designed to minimize the bound on the stability term in Lemma 1. The analysis is similar to the work of Avner et al. (2012).

Proof of Lemma 2 We optimize q_t for each round separately through a solution of the constrained optimization problem:

$$\begin{aligned} \min_{q_t} \quad & \sum_{i=1}^{K} \frac{(p_{t,i})^{2-\alpha}}{q_{t,i}} \\ \text{s.t.} \quad & \sum_{i=1}^{K} q_{t,i} = 1, \\ & \forall i \in [K], 0 \leq q_{t,i} \leq 1. \end{aligned}$$

The minimum is achieved by the distribution $q_{t,i} = \frac{(p_{t,i})^{1-\alpha/2}}{\sum_{j=1}^{K}(p_{t,j})^{1-\alpha/2}}$. ■