Improved Analysis of the Tsallis-INF Algorithm in Stochastically Constrained Adversarial Bandits and Stochastic Bandits with Adversarial Corruptions

Saeed Masoudian   SAEED.MASOUDIAN@DI.KU.DK
Yevgeny Seldin   SELDIN@DI.KU.DK
University of Copenhagen, Denmark

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We derive improved regret bounds for the Tsallis-INF algorithm of Zimmert and Seldin (2021). We show that in adversarial regimes with a (Δ, C, T) self-bounding constraint the algorithm achieves

$$\mathcal{O}\left(\left(\sum_{i \neq i^*} \frac{1}{\Delta_i}\right) \log_+\left(\frac{(K-1)T}{\left(\sum_{i \neq i^*} \frac{1}{\Delta_i}\right)^2}\right) + \sqrt{C\left(\sum_{i \neq i^*} \frac{1}{\Delta_i}\right)\log_+\left(\frac{(K-1)T}{C\sum_{i \neq i^*} \frac{1}{\Delta_i}}\right)}\right)$$

regret bound, where T is the time horizon, K is the number of arms, Δ_i are the suboptimality gaps, i^* is the best arm, C is the corruption magnitude, and $\log_+(x) = \max(1, \log x)$. The regime includes stochastic bandits, stochastically constrained adversarial bandits, and stochastic bandits with adversarial corruptions as special cases. Additionally, we provide a general analysis, which allows to achieve the same kind of improvement for generalizations of Tsallis-INF to other settings beyond multiarmed bandits.

1. Introduction

Most of the literature on multiarmed bandits is focused either on the stochastic setting (Thompson, 1933; Robbins, 1952; Lai and Robbins, 1985; Auer et al., 2002a) or on the adversarial one (Auer et al., 2002b). However, in recent years there has been an increasing interest in algorithms that perform well in both regimes with no prior knowledge of the regime (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Wei and Luo, 2018), as well as algorithms that perform well in intermediate regimes between stochastic and adversarial (Seldin and Slivkins, 2014; Lykouris et al., 2018; Wei and Luo, 2018; Gupta et al., 2019). The quest for best-of-both-worlds algorithm culminated with the work of Zimmert and Seldin (2019), who proposed the Tsallis-INF algorithm and showed that its regret bound in both stochastic and adversarial environments matches the corresponding lower bounds within constants with no need of prior knowledge of the regime. Zimmert and Seldin (2021) further improved the analysis and introduced an *adversarial regime with a self-bounding constraint*, which is an intermediate regime between stochastic and adversarial environments, including *stochastically constrained adversaries* (Wei and Luo, 2018) and *stochastic bandits with adversarial corruptions* (Lykouris et al., 2018) as special cases. They have shown that the Tsallis-INF algorithm achieves the best known regret rate in this regime and its special cases.

The Tsallis-INF algorithm is based on regularization by Tsallis entropy with power $\frac{1}{2}$, which was also used in the earlier works by Audibert and Bubeck (2009, 2010) and Abernethy et al. (2015) for minimax optimal regret rates in the adversarial regime. The key novelty of the work of Zimmert

and Seldin (2019, 2021) is an analysis of the algorithm in the stochastic setting based on a self-bounding property of the regret. The idea has been subsequently extended to derive best-of-both-worlds algorithms for combinatorial semi-bandits (Zimmert et al., 2019), decoupled exploration and exploitation (Rouyer and Seldin, 2020), bandits with switching costs (Rouyer et al., 2021), and ergodic MDPs (Jin and Luo, 2020).

We present a refined analysis based on the self-bounding property, which improves the regret bound in the adversarial regime with a self-bounding constraint and its special cases: stochastic bandits, stochastically constrained adversarial bandits, and stochastic bandits with adversarial corruption. The adversarial regime with a self-bounding constraint is defined in the following way. Let $\ell_1, \ell_2, \ldots$ be a sequence of loss vectors with $\ell_t \in [0,1]^K$, let I_t be the action picked by the algorithm at round t, and let $\overline{Reg}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right]$ be the pseudo-regret. For a triplet (Δ, C, T) with $\Delta \in [0,1]^K$ and $C \geq 0$, Zimmert and Seldin (2021) define an *adversarial regime with a (Δ, C, T) self-bounding constraint* as an adversarial regime, where the adversary picks losses, such that the pseudo-regret of any algorithm at time T satisfies

$$\overline{Reg}_T \geq \sum_{t=1}^{T} \sum_i \Delta_i \mathbb{P}(I_t = i) - C.$$

(The above condition is only assumed to be satisfied at time T, but there is no requirement that it is satisfied at time $t < T$.) A special case of this regime is the stochastically constrained adversarial regime, where $\overline{Reg}_T = \sum_{t=1}^{T} \sum_i \Delta_i \mathbb{P}(I_t = i)$ with Δ being the vector of suboptimality gaps. In particular, the stochastic regime is a special case of the stochastically constrained adversarial regime. (In the stochastic regime the expected loss of each arm is fixed over time. Stochastically constrained adversarial regime relaxes this requirement by only assuming that the expected gaps between the losses of pairs of arms are fixed, but the expected losses are allowed to fluctuate over time.) Another special case of an adversarial regime with a self-bounding constraint are stochastic bandits with adversarial corruptions. For two sequences of losses $\overline{\mathcal{L}}_T = (\bar{\ell}_1, \ldots, \bar{\ell}_T)$ and $\mathcal{L}_T = (\ell_1, \ldots, \ell_T)$ the amount of corruption is measured by $\sum_{t=1}^{T} \|\bar{\ell}_t - \ell_t\|_\infty$. In stochastic bandits with adversarial corruptions the adversary takes a stochastic sequence of losses and injects corruption with corruption magnitude bounded by C. Zimmert and Seldin (2021) show that a stochastic, as well as a stochastically constrained adversarial regime with a vector of suboptimality gaps Δ and injected corruption of magnitude bounded by C, satisfy $(\Delta, 2C, T)$ self-bounding constraint. As C grows from zero to T, the stochastic regime with adversarial corruptions interpolates between stochastic and adversarial bandits.

Lykouris et al. (2018) were the first to introduce and study stochastic bandits with adversarial corruptions and their algorithm achieved $\mathcal{O}\left(\sum_{i:\Delta_i>0} \frac{KC+\log(T)}{\Delta_i}\log(T)\right)$ regret bound. Gupta et al. (2019) improved it to $\mathcal{O}\left(KC + \sum_{i:\Delta_i>0} \frac{1}{\Delta_i}\log^2(KT)\right)$. Zimmert and Seldin (2021) have shown that their best-of-both-worlds Tsallis-INF algorithm achieves $\mathcal{O}\left(\left(\sum_{i\neq i^*} \frac{\log T}{\Delta_i}\right) + \sqrt{C\sum_{i\neq i^*} \frac{\log T}{\Delta_i}}\right)$ regret bound in the more general adversarial regime with (Δ, C, T) self-bounding constraint under the assumption that Δ has a unique zero entry (the assumption corresponds to uniqueness of the best arm *before* corruption). Neither of the algorithms requires prior knowledge of C.

Our contributions are summarized in the enumerated list below. The improvements relative to the work by Zimmert and Seldin (2021) are further highlighted in Table 1.

Setting	Zimmert and Seldin (2021)	Our paper
Small C	$\mathcal{O}\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T\right)$	$\mathcal{O}\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log_+\left(T\frac{K-1}{(\sum_{i\neq i^*}1/\Delta_i)^2}\right)\right)$
Large C	$\mathcal{O}\left(\sqrt{C\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T}\right)$	$\mathcal{O}\left(\sqrt{C\sum_{i\neq i^*}\frac{1}{\Delta_i}\log_+\left(T\frac{K-1}{C(\sum_{i\neq i^*}1/\Delta_i)}\right)}\right)$

Table 1: Comparison of the leading terms in the regret bounds of Zimmert and Seldin (2021) and our paper, differences are highlighted in color. We define $\log_+(x) = \max(1, \log x)$. The "Small C" row compares the regret bounds in adversarial regimes with (Δ, C, T) self-bounding constraints with $C \leq \sum_{i\neq i*}\frac{1}{\Delta_i}\left(\left(\log\frac{T(K-1)}{(\sum_{i\neq i*}\frac{1}{\Delta_i})^2}\right)+1\right)$. Here, C is a subdominant term and does not show up in the big-$\mathcal{O}$ notation. The "Large C" row compares the regret bounds in adversarial regimes with (Δ, C, T) self-bounding constraints with $C \geq \sum_{i\neq i*}\frac{1}{\Delta_i}\left(\left(\log\frac{T(K-1)}{(\sum_{i\neq i*}\frac{1}{\Delta_i})^2}\right)+1\right)$. The regret bounds in the adversarial regime are identical, and hence omitted.

1. We present a refined analysis of the regret of Tsallis-INF in adversarial regimes with a (Δ, C, T) self-bounding constraint, achieving

$$\mathcal{O}\left(\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log_+\left(\frac{(K-1)T}{\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)^2}\right)+\sqrt{C\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log_+\left(\frac{(K-1)T}{C\sum_{i\neq i^*}\frac{1}{\Delta_i}}\right)}\right)$$

regret bound, where $\log_+(x) = \max(1, \log x)$.

2. In the stochastically constrained adversarial regime it improves the dominating term of the regret bound from $\mathcal{O}\left(\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log T\right)$ to $\mathcal{O}\left(\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log_+\left(\frac{(K-1)T}{\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)^2}\right)\right)$ relative to the work of Zimmert and Seldin (2021), see Table 1. A similar kind of improvement has been studied for UCB-type algorithms for stochastic bandits by Auer and Ortner (2010) and Lattimore (2018).

3. In the stochastic regime with adversarial corruptions the result yields an improvement by a multiplicative factor of $\mathcal{O}\left(\sqrt{\log T/\log(T/C)}\right)$ relative to the work of Zimmert and Seldin (2021), see Table 1 for a more refined statement. In particular, for $C = \Theta\left(\frac{TK}{(\log T)\sum_{i\neq i^*}\frac{1}{\Delta_i}}\right)$ it achieves an improvement by a multiplicative factor of $\sqrt{\frac{\log T}{\log\log T}}$.

4. While the analysis of Zimmert and Seldin (2021) used two different optimization problems to analyze the regret of Tsallis-INF in adversarial environments and in adversarial environments with a self-bounding constraint, we obtain both bounds from the same optimization problem. This provides continuity in the analysis in the sense that the $\mathcal{O}\left(\sqrt{KT}\right)$ adversarial regret bound is obtained as a natural limit case of the adversarial bound with a self-bounding constraint as C grows beyond $\mathcal{O}\left(\frac{KT}{\sum_{i\neq i^*}\frac{1}{\Delta_i}}\right)$. It also provides a better understanding of the self-bounding analysis technique.

5. We also provide a more general result, showing that any algorithm with adversarial pseudo-regret bound satisfying $\overline{Reg}_T \leq B \sum_{t=1}^{T} \sum_{i \neq i^*} \sqrt{\frac{\mathbb{E}[w_{t,i}]}{t}}$, where $w_{t,i}$ are the probabilities of playing action i at round t and B is a constant, achieves

$$\mathcal{O}\left(B^2 \left(\sum_{i \neq i^*} \frac{1}{\Delta_i} \right) \log_+ \left(\frac{(K-1)T}{\left(\sum_{i \neq i^*} \frac{1}{\Delta_i} \right)^2} \right) + B \sqrt{C \left(\sum_{i \neq i^*} \frac{1}{\Delta_i} \right) \log_+ \left(\frac{KT}{C \sum_{i \neq i^*} \frac{1}{\Delta_i}} \right)} \right)$$

regret in the adversarial regime with (Δ, C, T) self-bounding constraint. The result can be directly applied to achieve improved regret bounds for extensions of the Tsallis-INF algorithm, for example, the extension to episodic MDPs (Jin and Luo, 2020).

2. Problem Setting

We study multi-armed bandit problem in which at time $t = 1, 2, \ldots$ the learner chooses an arm I_t among a set of K arms $\{1, \ldots, K\}$. At the same time the environment selects a loss vector $\ell_t \in [0, 1]^K$ and the learner only observes and suffers the loss ℓ_{t,I_t}. The performance of the learner is evaluated using pseudo-regret, which is defined as

$$\overline{Reg}_T = \mathbb{E}\left[\sum_{t=t}^{T} \ell_{t,I_t} \right] - \min_{i \in [K]} \mathbb{E}\left[\sum_{t=t}^{T} \ell_{t,i} \right] = \mathbb{E}\left[\sum_{t=t}^{T} \left(\ell_{t,I_t} - \ell_{t,i_T^*} \right) \right],$$

where $i_T^* \in \operatorname{argmin}_{i \in [K]} \mathbb{E}\left[\sum_{t=t}^{T} \ell_{t,i} \right]$ is a best arm in hindsight in expectation over the loss generation model and, in case of an adaptive adversary, the randomness of the learner.

Like Zimmert and Seldin (2021) we consider (*adaptive*) *adversarial regimes* and *adversarial regimes with a* (Δ, C, T) *self-bounding constraint*. In the former the losses at round t are generated arbitrarily, potentially depending on the preceding actions of the learner, $I_1 \ldots, I_{t-1}$. In the latter the adversary selects losses, such that for some $\Delta \in [0, 1]^K$ and $C \geq 0$ the pseudo-regret of any algorithm at time T satisfies

$$\overline{Reg}_T \geq \left(\sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{P}(I_t = i) \Delta_i \right) - C. \tag{1}$$

The condition is only assumed to be satisfied at time T, but not necessarily at $t < T$. As we have already mentioned in the introduction, *stochastic* regime, *stochastically constrained adversarial* regime, and *stochastic bandits with adversarial corruptions* are all special cases of the adversarial regime with (Δ, C, T) self-bounding constraint.

Additional Notation: We use Δ^n to denote the probability simplex over $n + 1$ points. The characteristic function of a closed convex set $\mathcal{A}$ is denoted by $\mathcal{I}_{\mathcal{A}}(x)$ and satisfies $\mathcal{I}_{\mathcal{A}}(x) = 0$ for $x \in \mathcal{A}$ and $\mathcal{I}_{\mathcal{A}}(x) = \infty$ otherwise. We denote the indicator function of an event $\mathcal{E}$ by $\mathbb{1}(\mathcal{E})$ and use $\mathbb{1}_t(i)$ as a shorthand for $\mathbb{1}(I_t = i)$. The probability distribution over arms that is played by the learner at round t is denoted by $w_t \in \Delta^{K-1}$. The convex conjugate of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined by $f^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle x, y \rangle - f(x)\}$.

3. Background: the Tsallis-INF algorithm

In this section we provide a brief background on the Tsallis-INF algorithm of Zimmert and Seldin (2021). The algorithm is based on Follow The Regularized Leader (FTRL) framework with Tsallis entropy regularization (Tsallis, 1988). The best-of-both-worlds version of Tsallis-INF uses Tsallis entropy regularizer with power $\frac{1}{2}$, defined by

$$\Psi(w) = 4 \sum_{i=1}^{K} \left(\sqrt{w_i} - \frac{1}{2} w_i \right).$$

The regularization term at round t is given by

$$\Psi_t(w) = \eta_t^{-1} \Psi(w),$$

where η_t is the learning rate. The update rule for the distribution over actions is defined by

$$w_{t+1} = \nabla (\Psi_t + \mathcal{I}_{\Delta^{K-1}})^*(-\sum_{\tau=1}^{t} \hat{\ell}_\tau) = \arg \max_{w \in \Delta^{K-1}} \left(\left\langle -\sum_{\tau=1}^{t} \hat{\ell}_\tau, w \right\rangle - \Psi_t(w) \right),$$

where $\hat{\ell}_\tau$ is an estimate of the loss vector ℓ_τ. It is possible to use the standard importance-weighed loss estimate $\hat{\ell}_{t,i} = \frac{\ell_{t,i} \mathbb{1}(\mathcal{I}_t = i)}{w_{t,i}}$, but Zimmert and Seldin (2021) have shown that reduced-variance loss estimates defined by

$$\hat{\ell}_{t,i} = \frac{\mathbb{1}_t(i)(\ell_{t,i} - \mathbb{B}_t(i))}{w_{t,i}} + \mathbb{B}_t(i), \tag{2}$$

where $\mathbb{B}_t(i) = \frac{1}{2} \mathbb{1}\left(w_{t,i} \geq \eta_t^2 \right)$, lead to better constants. The complete algorithm is provided in Algorithm 1 box. The regret bound derived by Zimmert and Seldin (2021) is provided in Theorem 1.

Algorithm 1 Tsallis-INF

1: **Input:** $(\Psi_t)_{t=1,2,\ldots}$
2: **Initialize:** Set $\hat{L}_0 = \mathbf{0}_K$ (where $\mathbf{0}_K$ is a zero vector in $\mathbb{R}^K$)
3: **for** t = 1, … **do**
4:     choose $w_t = \nabla (\Psi_t + \mathcal{I}_{\Delta^{K-1}})^*(-\hat{L}_{t-1})$
5:     sample $I_t \sim w_t$
6:     observe ℓ_{t,I_t}
7:     construct a loss estimator $\hat{\ell}_t$ using (2)
8:     update $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$
9: **end for**

Theorem 1 (Zimmert and Seldin, 2021) *The pseudo-regret of Tsallis-INF with $\eta_t = \frac{4}{\sqrt{t}}$ and reduced variance loss estimators defined in equation (2), in any adversarial bandit problem satisfies*

$$\overline{Reg}_T \leq 2\sqrt{KT} + 10K \log(T) + 16.$$

Furthermore, if there exists a vector $\Delta \in [0,1]^K$ with a unique zero entry i^* (i.e., $\Delta_{i^*} = 0$ and $\Delta_i > 0$ for all $i \neq i^*$) and a constant C, such that the pseudo-regret at time T satisfies the

(Δ, C, T) *self-bounding constraint (equation* (1)*), then the pseudo-regret additionally satisfies:*

$$\overline{Reg}_T \leq \left(\sum_{i \neq i^*} \frac{\log(T) + 3}{\Delta_i} \right) + 28K \log(T) + \frac{1}{\Delta_{min}} + \frac{3}{2}\sqrt{K} + 32 + C, \qquad (3)$$

where $\Delta_{min} = \min_{i \neq i^*} \{\Delta_i\}$. *Moreover, if* $C \geq \left(\sum_{i \neq i^*} \frac{\log(T)+3}{\Delta_i} \right) + \frac{1}{\Delta_{min}}$, *then the pseudo-regret also satisfies:*

$$\overline{Reg}_T \leq 2\sqrt{\left(\sum_{i \neq i^*} \frac{\log(T) + 3}{\Delta_i} + \frac{1}{\Delta_{min}} \right) C + 28K \log(T) + \frac{3}{2}\sqrt{K} + 32}. \qquad (4)$$

Remark 2 *While Theorem 1 requires uniqueness of the best arm for improved regret rates in the adversarial regime with a* (Δ, C, T) *self-bounding constraint, Zimmert and Seldin (2021) have shown experimentally that in the stochastic regime the presence of multiple best arms has no negative effect on the pseudo-regret of the algorithm. They conjecture that the requirement is an artifact of the analysis.*

4. Main Results

In this section we provide our two main results. First, in Theorem 3 we provide a refined analysis of Tsallis-INF, which improves the pseudo-regret bounds in the adversarial regime with a (Δ, C, T) self-bounding constraint. Then, in Theorem 4 we provide a more general result, which allows to improve pseudo-regret bounds in adversarial regimes with (Δ, C, T) self-bounding constraints for extensions of Tsallis-INF to other problems. An advantage of both results is that the bounds for adversarial regimes and adversarial regimes with a self-bounding constraint are achieved from a single optimization problem, rather than from two different optimization problems, as in prior work. As a result, the regret bounds for the adversarial regime are achieved as a limit case of the regret bounds for adversarial regimes with a self-bounding constraint for large C.

4.1. Improved analysis of the Tsallis-INF algorithm

We start with an improved regret bound for Tsallis-INF.

Theorem 3 *The pseudo-regret of Tsallis-INF with* $\eta_t = \frac{4}{\sqrt{t}}$ *and reduced variance loss estimators defined in equation* (2)*, in any adversarial bandit problem satisfies*

$$\overline{Reg}_T \leq 2\sqrt{(K-1)T} + \frac{1}{2}\sqrt{T} + 14K \log(T) + \frac{3}{4}\sqrt{K} + 15. \qquad (5)$$

Furthermore, if there exists a vector $\Delta \in [0,1]^K$ *with a unique zero entry* i^* *(i.e.,* $\Delta_{i^*} = 0$ *and* $\Delta_i > 0$ *for all* $i \neq i^*$*) and a constant* $C \geq 0$*, such that the pseudo-regret at time* T *satisfies the* (Δ, C, T) *self-bounding constraint (equation* (1)*), then the pseudo-regret additionally satisfies:*

$$\overline{Reg}_T \leq \sum_{i \neq i*} \frac{1}{\Delta_i} \left(\left(\log \frac{T(K-1)}{\left(\sum_{i \neq i^*} \frac{1}{\Delta_i} \right)^2} \right) + 6 \right) + 28K \log(T) + \frac{3}{2}\sqrt{K} + 30 + C. \qquad (6)$$

Moreover, for $\sum_{i \neq i^*} \frac{1}{\Delta_i} \left(\left(\log \frac{T(K-1)}{(\sum_{i \neq i^*} \frac{1}{\Delta_i})^2} \right) + 1 \right) \leq C \leq \frac{T(K-1)}{\sum_{i \neq i^*} \frac{1}{\Delta_i}}$ the regret also satisfies:

$$\overline{Reg}_T \leq \sqrt{C \sum_{i \neq i^*} \frac{1}{\Delta_i}} \left(\sqrt{\log \frac{T(K-1)}{C \sum_{i \neq i^*} \frac{1}{\Delta_i}}} + 5 \right) + Q, \tag{7}$$

where $Q = \sum_{i \neq i^*} \frac{1}{\Delta_i} \left(\log \frac{T(K-1)}{C \sum_{i \neq i^*} \frac{1}{\Delta_i}} + \sqrt{2 \log \frac{T(K-1)}{C \sum_{i \neq i^*} \frac{1}{\Delta_i}}} + 2 \right) + \frac{3\sqrt{K}}{2} + 28K \log(T) + 30$ is a subdominant term.

A proof of the theorem is provided in Appendix B. Theorem 3 improves on Theorem 1 in two ways. The bound in equation (6) improves the leading term of the regret bound under self-bounding constraint relative to equation (3) from $\sum_{i \neq i^*} \frac{1}{\Delta_i} \log T$ to $\sum_{i \neq i^*} \frac{1}{\Delta_i} \left(\log \frac{T(K-1)}{\left(\sum_{i \neq i^*} \frac{1}{\Delta_i} \right)^2} \right)$. Related refinements of regret bounds for UCB strategies for ordinary stochastic bandits have been studied by Auer and Ortner (2010) and Lattimore (2018). More importantly, for large amount of corruption $C \in \left[\sum_{i \neq i^*} \frac{1}{\Delta_i} \left(\log \left(\frac{T(K-1)}{(\sum_{i \neq i^*} \frac{1}{\Delta_i})^2} \right) + 1 \right), \frac{T(K-1)}{\sum_{i \neq i^*} \frac{1}{\Delta_i}} \right]$ the regret bound in equation (7) is of order $\mathcal{O} \left(\sqrt{C \left(\sum_{i \neq i^*} \frac{1}{\Delta_i} \right) \log_+ \left(\frac{KT}{C \sum_{i \neq i^*} \frac{1}{\Delta_i}} \right)} \right)$, whereas the regret bound in equation (4) is of order $\mathcal{O} \left(\sqrt{C \sum_{i \neq i^*} \frac{\log T}{\Delta_i}} \right)$. For $C = \Theta \left(\frac{TK}{(\log T) \sum_{i \neq i^*} \frac{1}{\Delta_i}} \right)$ Theorem 3 improves the pseudo-regret bound by a multiplicative factor of $\sqrt{\frac{\log T}{\log \log T}}$. Another observation is that Theorem 3 successfully exploits the self-bounding property even when the amount of corruption is almost linear in T.

4.2. A general analysis based on the self-bounding property

Now we provide a general result, which can be used to analyze extensions of Tsallis-INF to other problem settings.

Theorem 4 *For any algorithm for an arbitrary problem domain with K possible actions that satisfies*

$$\overline{Reg}_T \leq B \sum_{t=1}^{T} \sum_{i \neq i^*} \sqrt{\frac{\mathbb{E}[w_{t,i}]}{t}} + D, \tag{8}$$

where $B, D \geq 0$ are some constants, the pseudo-regret of the algorithm in any adversarial environment satisfies

$$\overline{Reg}_T \leq 2B \sqrt{(K-1)T} + D. \tag{9}$$

Furthermore, if there exists a vector $\Delta \in [0,1]^K$ with a unique zero entry i^* (i.e., $\Delta_{i^*} = 0$ and $\Delta_i > 0$ for all $i \neq i^*$) and a constant $C \geq 0$, such that the pseudo-regret at time T satisfies the (Δ, C, T) self-bounding constraint (equation (1)), then the pseudo-regret additionally satisfies:

$$\overline{Reg}_T \leq B^2 \sum_{i \neq i^*} \frac{1}{\Delta_i} \left(\left(\log \frac{T(K-1)}{\left(\sum_{i \neq i^*} \frac{1}{\Delta_i} \right)^2} \right) + 3 - 2 \log B \right) + C + 2D. \tag{10}$$

Moreover, for $B^2 \sum_{i \neq i*} \frac{1}{\Delta_i} \left(\left(\log \frac{T(K-1)}{B^2 (\sum_{i \neq i*} \frac{1}{\Delta_i})^2} \right) + 1 \right) \leq C \leq \frac{T(K-1)}{\sum_{i \neq i*} \frac{1}{\Delta_i}}$ the pseudo-regret also satisfies:

$$\overline{Reg}_T \leq B \sqrt{C \sum_{i \neq i^*} \frac{1}{\Delta_i} \left(\sqrt{\log \frac{T(K-1)}{C \sum_{i \neq i^*} \frac{1}{\Delta_i}}} + 2 \right)} + M, \tag{11}$$

where $M = B^2 \sum_{i \neq i*} \frac{1}{\Delta_i} \left(\log \frac{T(K-1)}{C \sum_{i \neq i*} \frac{1}{\Delta_i}} + \sqrt{2 \log \frac{T(K-1)}{C \sum_{i \neq i*} \frac{1}{\Delta_i}}} + 2 \right) + 2D$ is a subdominant term.

A proof is provided in Section 5. The Tsallis-INF algorithm satisfies the condition in equation (8) with $B = \frac{5}{4}$ (see equation (12) in Section 5, which follows from intermediate results by Zimmert and Seldin (2021)). Although the specialized analysis of Tsallis-INF in Theorem 3 is a bit tighter than the general result in Theorem 4, the latter can be applied to extensions of Tsallis-INF. One such example is the best-of-both-worlds algorithm of Jin and Luo (2020) for episodic MDPs. Jin and Luo (2020, Theorem 4) show that their algorithm satisfies the condition in (8) and use this result to achieve $\mathcal{O}\left((\log T) + \sqrt{C \log(T)} \right)$ pseudo-regret bound in the stochastic case with adversarial corruptions (Jin and Luo, 2020, Corollary 3). Application of our Theorem 4 improves the pseudo-regret bound to $\mathcal{O}\left((\log T) + \sqrt{C \log(T/C)} \right)$. In particular, for $C = \Theta(\frac{T}{\log T})$ the bound gets tighter by a multiplicative factor of $\frac{\log T}{\log \log T}$.

5. Proofs

In this section we provide a proof of Theorem 4. The proof of Theorem 3 is analogous, but more technical due to fine-tuning of the constants and is deferred to Appendix B. Before showing the proof we revisit the key steps in the analysis of Tsallis-INF by Zimmert and Seldin (2021), which show that the pseudo-regret of Tsallis-INF satisfies the condition in equation (8) of Theorem 4.

Standard FTRL analysis (Lattimore and Szepesvári, 2020) uses a potential function $\Phi_t(-L) = \max_{w \in \Delta^{K-1}} \{ \langle w, -L \rangle - \Psi_t(w) \}$ for breaking the pseudo-regret into *penalty* and *stability* terms, $\overline{Reg}_T = stability + penalty$, where

$$stability = \mathbb{E} \left[\sum_{t=1}^{T} \ell_{t,I_t} + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1}) \right],$$

$$penalty = \mathbb{E} \left[\sum_{t=1}^{T} -\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \ell_{t,i_T^*} \right].$$

The two terms are then typically analyzed separately. Zimmert and Seldin (2021) proved the following bounds for the two terms for Tsallis-INF with reduced-variance loss estimators:

$$stability \leq \sum_{t=1}^{T} \left(\sum_{i \neq i^*} \frac{\mathbb{E}[w_{t,i}]^{\frac{1}{2}}}{2\sqrt{t}} + \frac{\mathbb{E}[w_{t,i}]}{2\sqrt{t}} \right) + 14K \log(T) + 15,$$

$$penalty \leq \sum_{t=1}^{T} \left(\sum_{i \neq i^*} \frac{\mathbb{E}[w_{t,i}]^{\frac{1}{2}}}{2\sqrt{t}} - \frac{\mathbb{E}[w_{t,i}]}{4\sqrt{t}} \right) + \frac{3}{4}\sqrt{K}.$$

By summation of the two bounds the pseudo-regret satisfies

$$\overline{Reg}_T \leq \sum_{t=1}^{T} \left(\sum_{i \neq i^*} \frac{\mathbb{E}[w_{t,i}]^{\frac{1}{2}}}{\sqrt{t}} + \frac{\mathbb{E}[w_{t,i}]}{4\sqrt{t}} \right) + 14K \log(T) + \frac{3}{4}\sqrt{K} + 15. \tag{12}$$

Since $\mathbb{E}[w_{t,i}] \leq \mathbb{E}[w_{t,i}]^{\frac{1}{2}}$, the pseudo-regret of Tsallis-INF with reduced-variance loss estimators satisfies the condition in equation (8) with $B = \frac{5}{4}$ and $D = \frac{3}{4}\sqrt{K} + 14K \log(T) + 15$. (In the proof of Theorem 3 we keep the refined bound on the pseudo-regret from equation (12) to obtain better constants.) Now, after we have shown how the condition in equation (8) can be satisfied, we present a proof of Theorem 4. We start with a high-level overview of the key ideas and then present the technical details.

5.1. Overview of the Key Ideas Behind the Proof of Theorem 4

As observed by Zimmert and Seldin (2021), for any $\lambda \in [0, 1]$ we have

$$\overline{Reg}_T = (\lambda + 1)\overline{Reg}_T - \lambda\overline{Reg}_T. \tag{13}$$

The condition on $\overline{Reg}_T$ in equation (8) can be used to upper bound the first term and the self-bounding constraint (1) to lower bound the second, giving

$$\overline{Reg}_T \leq (\lambda + 1)\left(B \sum_{i \neq i^*} \sum_{t=1}^{T} \frac{\mathbb{E}[w_{t,i}]^{\frac{1}{2}}}{\sqrt{t}} + D \right) - \lambda\left(\sum_{t=1}^{T}\left(\sum_{i \neq i^*} \mathbb{E}[w_{t,i}]\Delta_i \right) - C \right)$$

$$\leq \sum_{t=1}^{T}\sum_{i \neq i^*}\left(B(\lambda+1)\frac{\mathbb{E}[w_{t,i}]^{\frac{1}{2}}}{\sqrt{t}} - \lambda\mathbb{E}[w_{t,i}]\Delta_i \right) + \lambda C + (\lambda+1)D. \tag{14}$$

In the adversarial analysis, we take $\lambda = 0$ and maximize the right hand side of (14) (which for $\lambda = 0$ is identical to the right hand side of (8)) under the constraint that $w_{t,i}$ is a probability distribution to obtain $\mathcal{O}(\sqrt{KT})$ regret bound. This is almost identical to the approach of Zimmert and Seldin (2021), except that in this case instead of the bound in equation (8) they use a bound involving summation over all arms, including i^*.

In the self-bounding analysis, Zimmert and Seldin (2021) relax the inequality in (14) to

$$\overline{Reg}_T \leq \sum_{t=1}^{T}\sum_{i \neq i^*}\left(2B\sqrt{\mathbb{E}[w_{t,i}]/t} - \lambda\Delta_i\mathbb{E}[w_{t,i}] \right) + \lambda C + 2D$$

and apply *individual* maximization of each $2B\sqrt{\mathbb{E}[w_{t,i}]/t} - \lambda\Delta_i\mathbb{E}[w_{t,i}]$ term, dropping the constraint that w_t is a probability distribution. We use (14) directly for bounding the regret and introduce two key novelties:

(a) we keep the constraint that w_t are probability distributions; and

(b) we jointly optimize with respect to all $w_{t,i}$ and λ, whereas Zimmert and Seldin (2021) first optimize w.r.t. $w_{t,i}$ and then w.r.t. λ.

9

Joint optimization over all $w_{t,i}$ and λ under the constraint that w_t are probability distributions is the major technical challenge that we resolve. Our analysis yields three advantages:

(A) The dependence on time is improved from $\log T$ to $\log(T(K-1)/(\sum_{i \neq i^*} \frac{1}{\Delta_i})^2)$ due to (a);

(B) We gain the $\sqrt{\log T / \log(T/C)}$ factor due to (b);

(C) Our adversarial and stochastic bounds come out of the same optimization problem, highlighting the relation and continuity between the two.

5.2. Proof of Theorem 4

Now we provide a detailed proof of Theorem 4.

Proof of the regret bound for an unconstrained adversarial regime (equation (9))

In the unconstrained adversarial regime we take $\lambda = 0$ and plug the inequalities

$$\sum_{i \neq i^*} \mathbb{E}[w_{t,i}]^{\frac{1}{2}} \leq \sqrt{K-1}, \tag{15}$$

which holds since $\sum_{i \neq i^*} \mathbb{E}[w_{t,i}] \leq 1$, and $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ into equation (14) and obtain the bound in equation (9).

Proof of the general regret bound for an adversarial regime with a self-bounding constraint (equation (10))

In the adversarial regime with a self-bounding constraint, we keep the constraint that w_t is a probability distribution, and thus $\sum_{i \neq i^*} \mathbb{E}[w_{t,i}]^{\frac{1}{2}} \leq \sqrt{K-1}$, and apply maximization directly to the sum over i under this constraint.

To simplify the notation, we use $a_{t,i} := \mathbb{E}[w_{t,i}]^{\frac{1}{2}}$, $S := \sum_{i \neq i^*} \frac{1}{\Delta_i}$, and w.l.o.g. assume that $i^* = K$. We denote $R_t := \sum_{i \neq i^*} \left(B(\lambda+1)\frac{a_{t,i}}{\sqrt{t}} - \lambda\Delta_i a_{t,i}^2 \right)$ and $R := \sum_{t=1}^{T} R_t + \lambda C$. With this notation, by equation (14) we have

$$\overline{Reg}_T \leq R + (1+\lambda)D. \tag{16}$$

We bound R_t under the constraint that $\mathbb{E}[w_{t,i}]^{\frac{1}{2}}$ satisfy equation (15). We have

$$R_t \leq \max_{a_1,\ldots,a_{K-1}} \sum_{i=1}^{K-1} B(\lambda+1)\frac{a_i}{\sqrt{t}} - \lambda\Delta_i a_i^2$$

$$\text{s.t.} \quad \sum_{i=1}^{K-1} a_i \leq \sqrt{K-1}.$$

By Lemma 6 provided in Appendix A, the answer to this optimization problem is as follows:

1. If $\frac{B(\lambda+1)S}{2\lambda\sqrt{t}} \leq \sqrt{K-1}$, then $R_t \leq \frac{SB^2(\lambda+1)^2}{4\lambda t}$.

2. If $\frac{B(\lambda+1)S}{2\lambda\sqrt{t}} \geq \sqrt{K-1}$, then $R_t \leq \frac{\sqrt{K-1}B(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S}$.

This gives a threshold $T_0 = \frac{B^2(\lambda+1)^2S^2}{4\lambda^2(K-1)}$, so that for $t \leq T_0$ the second case applies to R_t, and otherwise the first case applies. We break the time steps into those before T_0 and after T_0 and obtain:

$$
\begin{aligned}
R &= \sum_{t=1}^{T_0} R_t + \sum_{T_0+1}^{T} R_t + \lambda C \\
&\leq \sum_{t=1}^{T_0} \left(\frac{\sqrt{K-1}B(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S} \right) + \sum_{t=T_0+1}^{T} \frac{SB^2(\lambda+1)^2}{4\lambda t} + \lambda C \\
&\leq 2\sqrt{T_0(K-1)}B(\lambda+1) - \frac{\lambda(K-1)T_0}{S} + \frac{SB^2(\lambda+1)^2}{4\lambda}\log\frac{T}{T_0} + \lambda C \\
&= \frac{B^2(\lambda+1)^2S}{\lambda} - \frac{B^2(\lambda+1)^2S}{4\lambda} + \frac{B^2(\lambda+1)^2S}{4\lambda}\left(\log\frac{T(K-1)}{S^2} - 2\log\frac{B(\lambda+1)}{2\lambda} \right) + \lambda C \\
&= \frac{B^2(\lambda+1)^2S}{4\lambda}\left[3 + \log\frac{T(K-1)}{S^2} \right] - \frac{B^2(\lambda+1)^2S}{2\lambda}\log\frac{B(\lambda+1)}{2\lambda} + \lambda C.
\end{aligned}
\tag{17}
$$

By taking $\lambda = 1$ we obtain

$$
R \leq B^2 S \left(\log\frac{T(K-1)}{S^2} - 2\log(B) + 3 \right) + C,
$$

which together with (16) gives the bound (10) in the theorem.

Proof of the refined regret bound for an adversarial regime with a self-bounding constraint (equation (11))

We continue from equation (17). We improve on the bound of Zimmert and Seldin (2021) in equation (4) by applying a smarter optimization over λ. We let $\alpha = \frac{2\lambda}{B(\lambda+1)}$ and rewrite the inequality in (17) as

$$
R \leq \underbrace{\frac{B}{2-B\alpha}\underbrace{\left[\frac{S}{\alpha}\left(3 + \log\left(\frac{T(K-1)}{S^2} \right) \right) + \frac{2S}{\alpha}\log(\alpha) + \alpha C \right]}_{f(\alpha)}}_{h(B,\alpha)}.
\tag{18}
$$

We denote the right hand side of the expression by $h(B, \alpha)$. We restrict the range of α, so that $T \geq T_0 = \frac{S^2}{\alpha^2(K-1)}$, which gives $\alpha \geq \frac{S}{\sqrt{T(K-1)}}$. Since $\lambda \in [0, 1]$, we also have $\alpha \leq \frac{1}{B}$. In order to bound $h(B, \alpha)$ we need to solve an optimization problem in α over the above interval. However, $h(B, \alpha)$ is not convex in α, but we show that the expression in the brackets, which we denote by $f(\alpha)$, is convex. We take the point $\alpha^* = \text{argmin}_{\alpha \in [\frac{S}{\sqrt{T(K-1)}}, \frac{1}{B}]} f(\alpha)$, which achieves the minimum of $f(\alpha)$, and use $h(B, \alpha^*) = \frac{B}{2-B\alpha^*}f(\alpha^*)$ as an upper bound for R. Since $R \leq h(B, \alpha)$ for any α, in particular we have $R \leq h(B, \alpha^*)$.

In order to show that $f(\alpha)$ is convex and find its minimum we take the first and second derivatives.

$$f'(\alpha) = \frac{-1}{\alpha^2}\left[2S\log(\alpha) - C\alpha^2 + S\log\frac{(K-1)T}{S^2} + S\right] = 0,$$

$$f''(\alpha) = \frac{2S}{\alpha^3}\left(2\log\alpha + \log\frac{T(K-1)}{S^2}\right).$$

For $\alpha \geq \frac{S}{\sqrt{T(K-1)}}$ the second derivative is positive and, therefore, $f(\alpha)$ is convex and the minimum is achieved when $f'(\alpha) = 0$. This happens when

$$-\log\frac{\alpha^2(K-1)T}{S^2} + \frac{C}{S}\alpha^2 - 1 = 0.$$

We define $\beta = \frac{\alpha^2(K-1)T}{S^2}$, then

$$g(\beta) = \frac{CS}{(K-1)T}\beta - \log(\beta) - 1 = 0.$$

Since $\alpha \in [\frac{S}{\sqrt{T(K-1)}}, \frac{1}{B}]$, we have $\beta \in [1, \frac{(K-1)T}{B^2S^2}]$. We recall that equation (11) holds under the assumption that $B^2S\left(\log\frac{(K-1)T}{B^2S^2} + 1\right) \leq C \leq \frac{(K-1)T}{S}$. We note that for $C \leq \frac{(K-1)T}{S}$ we have $g(1) = \frac{CS}{(K-1)T} - 1 \leq 0$. We also note that for $C \geq B^2S\left(\log\frac{(K-1)T}{B^2S^2} + 1\right)$ we have $g\left(\frac{(K-1)T}{B^2S^2}\right) \geq 0$. Since $g(\beta)$ is continuous, the root of $g(\beta) = 0$ for C in the above range is thus achieved by $\beta \in [1, \frac{(K-1)T}{B^2S^2}]$ and since $g(\beta)$ is convex the solution is unique.

We find the root of $g(\beta) = 0$ by using the -1-branch of the *Lambert W function*, called $W_{-1}(x)$, which is defined as the solution of equation $we^w = x$. If $g(\beta) = 0$, then β satisfies

$$\frac{-CS\beta}{(K-1)T}e^{\frac{-CS\beta}{(K-1)T}} = \frac{-CS}{e(K-1)T},$$

and thus

$$\beta = \frac{-T(K-1)}{CS}W_{-1}\left(\frac{-CS}{e(K-1)T}\right).$$

We conclude that the minimum of $f(\alpha)$ is attained at

$$\alpha^* = \sqrt{\frac{-S}{C}W_{-1}\left(\frac{-CS}{e(K-1)T}\right)} \tag{19}$$

and, consequently, $\log\left(\frac{T(K-1)(\alpha^*)^2}{S^2}\right) = \frac{C}{S}(\alpha^*)^2 - 1$. By substituting this identity into $h(B, \alpha^*)$, we obtain:

$$h(B, \alpha^*) = \frac{B}{2 - B\alpha^*}\left(2\frac{S}{\alpha^*} + 2C\alpha^*\right) \leq B(1 + B\alpha^*)\left(\frac{S}{\alpha^*} + C\alpha^*\right)$$

$$= B\left(\frac{S}{\alpha^*} + C\alpha^* + BS + BC(\alpha^*)^2\right) = B\left(\sqrt{\frac{CS}{w}} + \sqrt{CSw} + BS + BSw\right), \tag{20}$$

where $w := -W_{-1}\left[\frac{-CS}{e(K-1)T}\right]$ and the inequality follows by the fact that $\forall x \in [0,1] : \frac{2}{2-x} \leq 1+x$. This provides a closed form upper bound for the pseudo-regret, but we still need an estimate of w to obtain an explicit bound. We use the result of Chatzigeorgiou (2013), who provides the following bounds for $W_{-1}(x)$.

Lemma 5 (Chatzigeorgiou 2013) *For any $x \leq 1$*

$$1 + \sqrt{2\log(1/x)} + \frac{2}{3}\log(1/x) \leq -W_{-1}(-x/e) \leq 1 + \sqrt{2\log(1/x)} + \log(1/x).$$

To complete the proof it suffices to use Lemma 5 with $x = \frac{CS}{(K-1)T}$, which gives

$$1 \leq w \leq 1 + \sqrt{2\log\frac{T(K-1)}{CS}} + \log\frac{T(K-1)}{CS} \leq \left(1 + \sqrt{\log\frac{T(K-1)}{CS}}\right)^2.$$

By substituting this into (20) we obtain:

$$
\begin{aligned}
h(B,\alpha^*) \leq{}& B\sqrt{CS} + B\sqrt{CS}\left(1 + \sqrt{\log\frac{T(K-1)}{CS}}\right) \\
& + 2B^2 S + B^2 S\log\frac{T(K-1)}{CS} + B^2 S\sqrt{2\log\frac{T(K-1)}{CS}} \\
={}& B\sqrt{CS}\left(\sqrt{\log\frac{T(K-1)}{CS}} + 2\right) + B^2 S\left(\log\frac{T(K-1)}{CS} + \sqrt{2\log\frac{T(K-1)}{CS}} + 2\right).
\end{aligned}
$$

$$\tag{21}$$

Finally, by (18) we have $R \leq h(B,\alpha^*)$, which together with (16) and the fact that $\lambda \leq 1$ completes the proof. ∎

6. Discussion

We have presented a refined analysis of the Tsallis-INF algorithm in adversarial regimes with a self-bounding constraint. The result improves on prior work in two ways. First, it improves the dependence of the regret bound on time horizon from $\log T$ to $\log\frac{(K-1)T}{(\sum_{i\neq i^*}\frac{1}{\Delta_i})^2}$. Second, it improves the dependence of the regret bound on corruption amount C. In particular, for $C = \Theta\left(\frac{TK}{(\log T)\sum_{i\neq i^*}\sum\frac{1}{\Delta_i}}\right)$ it improves the pseudo-regret bound by a multiplicative factor of $\sqrt{\frac{\log T}{\log\log T}}$. Moreover, we have provided a generalized result that can be used to improve regret bounds for extensions of Tsallis-INF to other problem settings, where the regret satisfies a self-bounding constraint. Due to versatility and rapidly growing popularity of regret analysis based on the self-bounding property, the result provides a powerful tool for tightening regret bounds in a broad range of corrupted settings.

Acknowledgments

References

Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2015.

Jean-Yves Audibert and Sébastien Bubeck. Regret Bounds and Minimax Policies under Partial Monitoring. *Journal of Machine Learning Research*, 11, 2010.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the Conference on 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 Conference on Learning Theory (COLT)*, 2016.

Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61, 2010.

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 multi-armed bandit problem. *SIAM Journal on Computing*, 32, 2002b.

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

Ioannis Chatzigeorgiou. Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17, 2013.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the Conference on Learning Theory (COLT)*, 2019.

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

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

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, 2018.

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

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the Annual ACM SIGACT Symposium on Theory of Computing*, 2018.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58, 1952.

Chloé Rouyer and Yevgeny Seldin. Tsallis-INF for decoupled exploration and exploitation in multi-armed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2020.

Chloé Rouyer, Yevgeny Seldin, and Nicolò Cesa-Bianchi. An algorithm for stochastic and adversarial bandits with switching costs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the Conference on 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.

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

Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52, 1988.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the Conference on 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.

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

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

Appendix A. Technical Lemmas

Lemma 6 *Let b and $c_1, \ldots, c_n$ be non-negative real numbers and let*

$$Z = \max_{x \in \mathbb{R}^n} \sum_{i=1}^{n} (bx_i - c_i x_i^2)$$

$$s.t. \sum_{i=1}^{n} x_i \leq M.$$

Then

$$Z = \begin{cases} bM - \frac{M^2}{\sum_{i=1}^{n} \frac{1}{c_i}}, & \text{if } \sum_{i=1}^{n} \frac{b}{2c_i} > M, \\ \frac{b^2}{4} \sum_{i=1}^{n} \frac{1}{c_i}, & \text{otherwise.} \end{cases}$$

Moreover, we always have $bM - \frac{M^2}{\sum_{i=1}^{n} \frac{1}{c_i}} \leq \frac{b^2}{4} \sum_{i=1}^{n} \frac{1}{c_i}$ and, therefore, we always have $Z \leq \frac{b^2}{4} \sum_{i=1}^{n} \frac{1}{c_i}$.

Proof Since $c_i \geq 0$, the objective function is a sum of downward-pointing parabolas and, therefore, concave. Thus, the maximum is attained when the first derivative of the Lagrangian with Lagrange variable $v \geq 0$ for the inequality constraint satisfies

$$b - 2c_i x_i - v = 0,$$

where $v(\sum_{i=1}^{n} x_i - M) = 0$. Thus, $x_i = \frac{b-v}{2c_i}$. The KKT conditions provide two cases:

   i) If $\sum_{i=1}^{n} \frac{b}{2c_i} > M$, then $v > 0$ and $\sum_{i=1}^{n} x_i = M$. As a consequence, $v = b - \frac{M}{\sum_{i=1}^{n} \frac{1}{2c_i}}$. So $x_i = \frac{M}{c_i \sum_{i=1}^{n} \frac{1}{c_i}}$ and $Z = bM - \frac{M^2}{\sum_{i=1}^{n} \frac{1}{c_i}}$.

   ii) If $\sum_{i=1}^{n} \frac{b}{2c_i} \leq M$, then $v = 0$ and, as a consequence, $x_i = \frac{b}{2c_i}$ and $Z = \frac{b^2}{4} \sum_{i=1}^{n} \frac{1}{c_i}$.

Finally, by the AM-GM inequality we have

$$\frac{M^2}{\sum_{i=1}^{n} \frac{1}{c_i}} + \frac{b^2}{4} \sum_{i=1}^{n} \frac{1}{c_i} \geq bM,$$

which gives the final statement of the lemma. ∎

We also use the following result by Zimmert and Seldin (2021, Lemma 15).

Lemma 7 (Zimmert and Seldin, 2021) *For any $b > 0$ and $c > 0$ and $T_0, T \in \mathbb{N}$, such that $T_0 < T$ and $b\sqrt{T_0} > c$, it holds that*

$$\sum_{t=T_0+1}^{T} \frac{1}{bt^{\frac{3}{2}} - ct} \leq \frac{2}{b\sqrt{T_0} - c}.$$

By doubling the lower threshold on $b\sqrt{T_0}$ we obtain the following corollary.

Corollary 8 *For any $b > 0$ and $c > 0$ and $T_0, T \in \mathbb{N}$, such that $T_0 < T$ and $b\sqrt{T_0} \geq 2c$, it holds that*

$$\sum_{t=T_0+1}^{T} \frac{1}{bt^{\frac{3}{2}} - ct} \leq \frac{2}{c}.$$

Appendix B. Proof of Theorem 3

Proof Similar to the proof of Theorem 4, for any $\lambda \in [0, 1]$ we use the self-bounding constraint and the regret bound of Zimmert and Seldin (2021) given in equation (12) to provide the following bound for the pseudo-regret:

$$\overline{Reg}_T = (\lambda + 1)\overline{Reg}_T - \lambda\overline{Reg}_T$$

$$\leq (\lambda + 1) \left(\sum_{i \neq i^*} \left[\sum_{t=1}^{T} \frac{\mathbb{E}[w_{t,i}]}{4\sqrt{t}} + \sum_{t=1}^{T} \frac{\sqrt{\mathbb{E}[w_{t,i}]}}{\sqrt{t}} \right] + \frac{3}{4}\sqrt{K} + 14K\log(T) + 15 \right)$$

$$- \lambda \left(\left[\sum_{t=1}^{T} \sum_{i \neq i^*} \mathbb{E}[w_{t,i}]\Delta_i \right] - C \right).$$

As before, to simplify the notation, let $a_{t,i} = \mathbb{E}[w_{t,i}]^{\frac{1}{2}}$ and $S = \sum_{i \neq i^*} \frac{1}{\Delta_i}$ and w.l.o.g. assume that $i^* = K$ and define

$$R_t = \sum_{i \neq i^*} \left(\frac{\lambda + 1}{\sqrt{t}} a_{t,i} - \left(\lambda\Delta_i - \frac{\lambda + 1}{4\sqrt{t}} \right) a_{t,i}^2 \right), \tag{22}$$

$$R = \sum_{t=1}^{T} R_t + \lambda C.$$

Then

$$\overline{Reg}_T \leq R + (1 + \lambda) \left(\frac{3}{4}\sqrt{K} + 14K\log(T) + 15 \right). \tag{23}$$

Hence, in order to obtain a bound for the pseudo-regret, it suffices to derive a bound for R. We start with the bound for a general adversarial environment and then prove the refinements.

Proof of the regret bound for an unconstrained adversarial regime (equation (5))

We take $\lambda = 0$. By plugging it into the definition of R_t in equation (22) we obtain

$$R_t \leq \frac{\sqrt{K-1}}{\sqrt{t}} + \frac{1}{4\sqrt{t}}$$

and

$$R = \sum_{t=1}^{T} R_t \leq 2\sqrt{(K-1)T} + \frac{1}{2}\sqrt{T}.$$

Plugging this into (23) completes the proof of (5).

Proof of the regret bounds for an adversarial regime with a self-bounding constraint (equations (6) and (7))

Now we prove the refined bounds for adversarial environments satisfying the self-bounding constraint with unique best arm. Similarly to the proof of Theorem 4, we bound R_t for each $t \geq 1$ by

solving a constrained maximization problem over $\{a_{t,i}\}_{i=1}^n$, where the constraint is $\sum_{i=1}^{K-1} a_{t,i} \leq \sqrt{K-1}$. But the challenge here is that the coefficients $\lambda\Delta_i - \frac{\lambda+1}{4\sqrt{t}}$ in front of $a_{t,i}^2$ in the definition of R_t are not necessarily positive, and if they are not, then Lemma 6 cannot be applied. More precisely, if

$$\forall i \neq i^* : \lambda\Delta_i \geq \frac{\lambda+1}{4\sqrt{t}} \Rightarrow t \geq \left(\frac{\lambda+1}{4\lambda\Delta_{min}}\right)^2, \tag{24}$$

where $\Delta_{min} = \min_{i \neq i^*}\{\Delta_i\}$, then all the coefficients are positive. We denote $\alpha = \frac{2\lambda}{\lambda+1}$ and define a threshold $T_1 = \left(\frac{\lambda+1}{2\lambda\Delta_{min}}\right)^2 = \left(\frac{1}{\alpha\Delta_{min}}\right)^2$. We note that T_1 is four times larger than what is required for satisfaction of the condition in equation (24). The reason is that at a later point in the proof we apply Corollary 8 for $t \geq T_1$ and we need to satisfy the condition of the corollary. For $t \geq T_1$ we can use Lemma 6 to bound R_t. By the lemma we obtain:

$$R_t \leq \frac{(\lambda+1)^2}{4t} \sum_{i=1}^{K-1} \frac{1}{\lambda\Delta_i - \frac{\lambda+1}{\sqrt{t}}} = \sum_{i=1}^{K-1} \frac{\lambda+1}{\frac{4\lambda}{\lambda+1}\Delta_i t - \sqrt{t}} = \sum_{i=1}^{K-1} \frac{\lambda+1}{2\alpha\Delta_i t - \sqrt{t}}.$$

We rewrite each term in the summation in the following way

$$\frac{\lambda+1}{2\alpha\Delta_i t - \sqrt{t}} = \frac{\lambda+1}{2\alpha\Delta_i t} + \frac{\lambda+1}{4\alpha^2\Delta_i^2 t^{\frac{3}{2}} - 2\alpha\Delta_i t}$$

and obtain

$$\text{for } t \geq T_1: \qquad R_t \leq \frac{S(\lambda+1)}{2\alpha t} + \sum_{i=1}^{K-1} \frac{\lambda+1}{4\alpha^2\Delta_i^2 t^{\frac{3}{2}} - 2\alpha\Delta_i t}. \tag{25}$$

In order to bound R_t for $t < T_1$, we break it into two parts as follows:

$$R_t = \sum_{i \neq i^*} \left(\frac{\lambda+1}{\sqrt{t}} a_{t,i} - \lambda\Delta_i a_{t,i}^2\right) + \sum_{i \neq i^*} \left(\frac{\lambda+1}{4\sqrt{t}} a_{t,i}^2\right)$$
$$\leq \sum_{i \neq i^*} \left(\frac{\lambda+1}{\sqrt{t}} a_{t,i} - \lambda\Delta_i a_{t,i}^2\right) + \frac{1}{2\sqrt{t}},$$

where the inequality holds because $\lambda \leq 1$ and $\sum_{i \neq i^*} a_{t,i}^2 \leq 1$. We use Lemma 6 to bound the summation in the latter expression. The solution depends on a threshold $T_2 = \frac{(\lambda+1)^2 S^2}{4\lambda^2(K-1)} = \frac{S^2}{(K-1)\alpha^2}$:

$$\text{for } t \leq T_2: \qquad R_t \leq \frac{\sqrt{K-1}(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S} + \frac{1}{2\sqrt{t}}, \tag{26}$$

$$\text{for } t \geq T_2: \qquad R_t \leq \frac{S(\lambda+1)^2}{4\lambda t} + \frac{1}{2\sqrt{t}} = \frac{S(\lambda+1)}{2\alpha t} + \frac{1}{2\sqrt{t}}. \tag{27}$$

Note that for $t \geq T_1$ we have a choice between using the bound in equation (25) or one of the bounds in (26) or (27), depending on whether $t \leq T_2$ or $t \geq T_2$. The relation between the thresholds, $T_1 \leq T_2$ or $T_2 \leq T_1$, depends on the relation between $\left(\frac{1}{\Delta_{min}}\right)^2$ and $\frac{S^2}{K-1}$. Also note that the choice of α (which determines λ) affects the thresholds T_1 and T_2, but not their relation. Similar

to the proof of Theorem 4, we restrict the range of α, so that $T \geq T_2 = \frac{S^2}{\alpha^2(K-1)}$, which gives $\alpha \geq \frac{S}{\sqrt{T(K-1)}}$.

We now derive a bound on R. We consider three cases: $T_2 \leq T \leq T_1$, $T_2 \leq T_1 \leq T$, and $T_1 \leq T_2 \leq T$.

First case: $T_2 \leq T \leq T_1$. By (26) and (27) we have:

$$\sum_{t=1}^{T} R_t \leq \sum_{t=1}^{T_2} R_t + \sum_{t=T_2+1}^{T} R_t$$

$$\leq \sum_{t=1}^{T_2} \left(\frac{\sqrt{K-1}(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S} \right) + \sum_{t=T_2+1}^{T} \left(\frac{S(\lambda+1)}{2\alpha t} \right) + \sqrt{T}$$

$$\leq 2\sqrt{T_2(K-1)}(\lambda+1) - \frac{\lambda(K-1)T_2}{S} + \frac{S(\lambda+1)}{2\alpha} \log(\frac{T}{T_2}) + \sqrt{T_1}, \qquad (28)$$

where in the second line we used $\sum_{t=1}^{T} \frac{1}{2\sqrt{t}} \leq \sqrt{T}$ and in the third line $\sum_{t=T_2+1}^{T} \frac{1}{t} \leq \log(T/T_2)$ and $\lambda \leq 1$ and $T \leq T_1$.

Second case: $T_2 \leq T_1 \leq T$. By (26), (27), and (25) we have:

$$\sum_{t=1}^{T} R_t \leq \sum_{t=1}^{T_2} R_t + \sum_{t=T_2+1}^{T_1} R_t + \sum_{t=T_1+1}^{T} R_t$$

$$\leq \sum_{t=1}^{T_2} \left(\frac{\sqrt{K-1}(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S} \right) + \sum_{t=T_2+1}^{T} \left(\frac{S(\lambda+1)}{2\alpha t} \right) + \sqrt{T_1} + \sum_{i=1}^{K-1} \sum_{t=T_1+1}^{T} \frac{\lambda+1}{4\alpha^2 \Delta_i^2 t^{\frac{3}{2}} - 2\alpha\Delta_i t}$$

$$\leq 2\sqrt{T_2(K-1)}(\lambda+1) - \frac{\lambda(K-1)T_2}{S} + \frac{S(\lambda+1)}{2\alpha} \log(\frac{T}{T_2}) + \sqrt{T_1} + \sum_{i=1}^{K-1} \sum_{t=T_1+1}^{T} \frac{1}{2\alpha^2 \Delta_i^2 t^{\frac{3}{2}} - \alpha\Delta_i t},$$

$$(29)$$

where in the second line we used $\sum_{t=1}^{T_1} \frac{1}{2\sqrt{t}} \leq \sqrt{T_1}$ and in the third line $\sum_{t=T_2+1}^{T} \frac{1}{t} \leq \log(T/T_2)$ and $\lambda \leq 1$.

Third case: $T_1 \leq T_2 \leq T$. By (26) and (25) we have:

$$\sum_{t=1}^{T} R_t \leq \sum_{t=1}^{T_2} R_t + \sum_{t=T_2+1}^{T} R_t$$

$$\leq \sum_{t=1}^{T_2} \left(\frac{\sqrt{K-1}(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S} \right) + \sqrt{T_2} + \sum_{t=T_2+1}^{T} \left(\frac{S(\lambda+1)}{2\alpha t} \right) + \sum_{i=1}^{K-1} \sum_{t=T_2+1}^{T} \frac{\lambda+1}{4\alpha^2 \Delta_i^2 t^{\frac{3}{2}} - 2\alpha\Delta_i t}$$

$$\leq 2\sqrt{T_2(K-1)}(\lambda+1) - \frac{\lambda(K-1)T_2}{S} + \sqrt{T_2} + \frac{S(\lambda+1)}{2\alpha} \log(\frac{T}{T_2}) + \sum_{i=1}^{K-1} \sum_{t=T_1+1}^{T} \frac{1}{2\alpha^2 \Delta_i^2 t^{\frac{3}{2}} - \alpha\Delta_i t}.$$

$$(30)$$

Merging the cases: Corollary 8 provides an upper bound for the last terms of (29) and (30):

$$\sum_{t=T_1+1}^{T} \frac{1}{2\alpha^2 \Delta_i^2 t^{\frac{3}{2}} - \alpha \Delta_i t} \leq \frac{2}{\alpha \Delta_i}.$$

Now we combine (28), (29), and (30), and obtain following bound for R:

$$
\begin{aligned}
R &= \sum_{t=1}^{T} R_t + \lambda C \\
&\leq 2\sqrt{T_2(K-1)}(\lambda+1) - \frac{\lambda(K-1)T_2}{S} + \frac{S(\lambda+1)}{2\alpha} \log(\frac{T}{T_2}) + \lambda C \\
&\quad + \sqrt{\max\{T_1, T_2\}} + \sum_{i=1}^{K-1} \frac{2}{\alpha \Delta_i}.
\end{aligned}
\tag{31}
$$

We note that $\max\{T_1, T_2\} = \max\left\{\frac{S^2}{(K-1)\alpha^2}, \frac{1}{\Delta_{min}^2 \alpha^2}\right\} \leq \frac{S^2}{\alpha^2}$. Moreover, by substituting $T_2 = \frac{S^2}{\alpha^2(K-1)}$ into (31) we obtain:

$$
\begin{aligned}
R &\leq 2(\lambda+1)\frac{S}{\alpha} - \frac{\lambda S}{\alpha^2} + \frac{S(\lambda+1)}{2\alpha} \log\left(\frac{\alpha^2(K-1)T}{S^2}\right) + \lambda C + \frac{S}{\alpha} + \sum_{i=1}^{K-1} \frac{2}{\alpha \Delta_i} \\
&= \frac{\lambda+1}{2}\left[4\frac{S}{\alpha} - \frac{S}{\alpha} + \frac{S}{\alpha}\log\left(\frac{(K-1)T}{S^2}\right) + \frac{2S}{\alpha}\log(\alpha) + \alpha C\right] + \frac{3S}{\alpha} \\
&= \underbrace{\frac{1}{2-\alpha}\left[\frac{S}{\alpha}\left(3 + \log\left(\frac{T(K-1)}{S^2}\right)\right) + \frac{2S}{\alpha}\log(\alpha) + \alpha C\right] + \frac{3S}{\alpha}}_{h(1,\alpha)}.
\end{aligned}
\tag{32}
$$

We recognize that the first term in equation (32) is $h(1, \alpha)$, which was defined earlier in equation (18).

Proof of the general bound in equation (6): By taking $\lambda = 1$, which corresponds to $\alpha = 1$, we obtain

$$
\begin{aligned}
R &\leq S\left(\log\left(\frac{T(K-1)}{S^2}\right) + 3\right) + C + 3S \\
&= S\left(\log\left(\frac{T(K-1)}{S^2}\right) + 6\right) + C.
\end{aligned}
$$

Plugging this and the value of λ into (16) completes the proof of (6).

Proof of the refined bound in equation (7): We note that the range of C in the refined bound in equation (7) is the same as in the refined bound in (11) in Theorem 4 for $B = 1$. We take α^* as in equation (19), i.e., $\alpha^* = \sqrt{\frac{-S}{C}W_{-1}\left(\frac{-CS}{e(K-1)T}\right)}$. By Lemma 5 we have $-W_{-1}\left(\frac{-CS}{e(K-1)T}\right) \geq 1$,

and thus $\alpha^* \geq \sqrt{\frac{S}{C}}$. By plugging this bound and the bound on $h(1, \alpha^*)$ from equation (21) into equation (32), we obtain:

$$R \leq \sqrt{CS} \left(\sqrt{\log \frac{T(K-1)}{CS}} + 2 \right) + S \left(\log \frac{T(K-1)}{CS} + \sqrt{2 \log \frac{T(K-1)}{CS}} + 2 \right) + 3\sqrt{CS}$$

$$= \sqrt{CS} \left(\sqrt{\log \frac{T(K-1)}{CS}} + 5 \right) + S \left(\log \frac{T(K-1)}{CS} + \sqrt{2 \log \frac{T(K-1)}{CS}} + 2 \right).$$

Plugging this bound into (16) and using the fact that $\lambda \leq 1$ completes the proof of (7). ∎