An Optimal Algorithm for Stochastic and Adversarial Bandits

Julian Zimmert **Yevgeny Seldin**
University of Copenhagen
Copenhagen, Denmark

Abstract

We derive an algorithm that achieves the optimal (up to constants) pseudo-regret in both adversarial and stochastic multi-armed bandits without prior knowledge of the regime and time horizon. The algorithm is based on online mirror descent with Tsallis entropy regularizer. We provide a complete characterization of such algorithms and show that Tsallis entropy with power $\alpha = 1/2$ achieves the goal. In addition, the proposed algorithm enjoys improved regret guarantees in two intermediate regimes: the moderately contaminated stochastic regime defined by Seldin and Slivkins [22] and the stochastically constrained adversary studied by Wei and Luo [26]. The algorithm also achieves adversarial and stochastic optimality in the utility-based dueling bandit setting. We provide empirical evaluation of the algorithm demonstrating that it outperforms UCB1 and EXP3 in stochastic environments. In certain adversarial regimes the algorithm significantly outperforms UCB1 and THOMPSON SAMPLING, which exhibit almost linear regret.

1 Introduction

Stochastic (i.i.d.) and adversarial multi-armed bandits are two fundamental sequential decision making problems in online learning [24; 19; 17; 9; 10]. When prior information about the nature of environment is available, it is possible to achieve $\mathcal{O}(\sum_{i:\Delta_i>0} \frac{\log(T)}{\Delta_i})$ pseudo-regret in the stochastic case and $\mathcal{O}(\sqrt{KT})$ pseudo-regret in the adversarial case [6; 7], and both results match the lower bounds up to constants, see [13] for a survey.[1] The challenge in recent years has been to achieve the optimal regret rates without prior knowledge about the nature of the problem.

One approach pursued by Bubeck and Slivkins [14] and later refined by Auer and Chiang [8] is to start an algorithm under the assumption that the environment is i.i.d. and constantly monitor whether the assumption is satisfied. If a deviation from the i.i.d. assumption is detected, the algorithm performs an irreversible switch into adversarial operation mode. This approach recovers the optimal bound in the stochastic case, but suffers from an additional logarithmic factor in the regret in the adversarial case. Furthermore, the time horizon needs to be known in advance. The best known doubling schemes lead to extra multiplicative logarithmic factors in either the stochastic or the adversarial regime [11].

Another approach pioneered by Seldin and Slivkins [22] alters algorithms designed for adversarial bandits to achieve improved regret in the stochastic setting, without losing the adversarial guarantees. In this line of work, Seldin and Lugosi [21] have achieved an anytime regret of $\mathcal{O}(\sum_{i:\Delta_i>0} \frac{\log(T)^2}{\Delta_i})$ in the stochastic case while preserving optimality in the adversarial case. A related approach by Wei and Luo [26] provides an anytime regret bound scaling with $\log T$ in the stochastic case, but besides a suboptimal problem-dependent factor of $\mathcal{O}(\frac{K}{\min_{\Delta_i>0}\Delta_i})$, it is also suboptimal by a logarithmic factor in the adversarial regime. Seldin and Slivkins [22] and Wei and Luo [26] also obtained improved regret guarantees in a number of intermediate regimes between stochastic and adversarial bandits.

The question of existence of a universal trade-off preventing optimality in both worlds simultaneously has remained open for a while. Auer and Chiang [8]

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

[1] To be precise, the $\mathcal{O}(\sum_{i:\Delta_i>0} \frac{\log(T)}{\Delta_i})$ stochastic regret rate is optimal when the means of the rewards are close to $\frac{1}{2}$, see [17; 15; 16] for refined lower and upper bounds otherwise. However, the refined analysis assumes that the means are fixed, whereas we only assume that the gaps are fixed, but the means are allowed to fluctuate arbitrarily, see Section 2 for details.

	Regime	$\frac{\textit{Upper Bound}}{\textit{Lower Bound}}$	*Learning Rate*
$\lim_{\alpha\to 0}$	Sto.†*	$\mathcal{O}(\mathbf{1})$	$\Theta\left(\Delta_i\right)^*$
	Adv.	$\mathcal{O}\left(\sqrt{\log(T)}\right)$	$\Theta\left(\sqrt{\frac{\log(t)}{t}}\right)$
$\alpha=\frac{1}{2}$	Sto.† & Adv.	$\mathcal{O}(\mathbf{1})$	$\frac{1}{\sqrt{t}}$
$\lim_{\alpha\to 1}$	Sto.†*	$\mathcal{O}(\log(T))$	$\Theta\left(\frac{\log(t)}{\Delta_i t}\right)^*$
	Adv.	$\mathcal{O}\left(\sqrt{\log(K)}\right)$	$\Theta\left(\frac{1}{\sqrt{t}}\right)$

†novel results, *oracle knowledge of Δ_i is required

Table 1: Complete characterization of Online Mirror Descent algorithms regularized by Tsallis Entropy; α=1 corresponds to the EXP3 algorithm.

have shown that any algorithm obtaining the optimal stochastic pseudo-regret bound cannot simultaneously achieve the optimal high-probability adversarial regret bound or optimal expected regret bound for adaptive adversaries.[2] In addition, Abbasi-Yadkori et al. [1] have shown that in the pure exploration setting it is also impossible to obtain the optimal rates in both stochastic and adversarial regimes.

We show that for the pseudo-regret it is possible to achieve optimality in both regimes with a surprisingly simple algorithm. Additionally, we provide improved regret guarantees for two intermediate regimes and extend the results to utility-based dueling bandits. The algorithm is based on online mirror descent with regularization by Tsallis entropy with power α. We name it α-Tsallis-Inf or simply Tsallis-Inf, where INF stands for Implicitly Normalized Forecaster [6]. The proposed algorithm is anytime: it requires neither the knowledge of the time horizon nor doubling schemes.

The paper is structured in the following way: In Section 2 we provide a formal definition of the problem setting, including adversarial and stochastically constrained adversarial environments. Stochastic environments are a special case of the latter. In Section 3 we briefly review the framework of online mirror descent. We follow the techniques of Bubeck [12] to adapt the family of algorithms based on regularization by α-Tsallis Entropy [25; 3] to anytime setting. Section 4 contains the main theorems. We show that $\alpha=\frac{1}{2}$ provides an algorithm that is optimal in both adversarial and stochastically constrained adversarial regimes; the latter implies optimality in the stochastic regime. Furthermore, we show that any $\alpha\in[0,1)$ could potentially achieve the optimal regret bound against stochastically constrained adversaries, but it requires oracle access to the (unknown) gaps for tuning the learning rate. A summary of the results is provided in Table 1. In Section 5 we show that $\frac{1}{2}$-Tsallis-Inf also achieves the optimal regret rate in the moderately contaminated stochastic regime of Seldin and Slivkins [22]. In Section 6 we apply $\frac{1}{2}$-Tsallis-Inf to dueling bandits. Section 7 contains the proofs of our main theorems. In Section 8 we provide an empirical comparison of Tsallis-Inf with baseline stochastic and adversarial bandit algorithms from the literature. We show that in stochastic environments $\frac{1}{2}$-Tsallis-Inf outperforms UCB1 and EXP3, but lags behind Thompson Sampling, whereas in certain adversarial environments it significantly outperforms UCB1 and Thompson Sampling, which suffer almost linear regret, and also outperforms EXP3. To the best of our knowledge, this is also the first evidence that Thompson Sampling is vulnerable outside stochastic environments. We conclude with a summary in Section 9.

2 Problem setting

At time $t=1,2,\ldots$, the agent chooses an arm $I_t\in\{1,\ldots,K\}$ out of a set of K arms. The environment picks a loss vector $\ell_t\in[0,1]^K$ and the agent observes and suffers *only* the loss of the arm played, ℓ_{t,I_t}. In the (*adaptive*) *adversarial setting*, the adversary selects the losses arbitrarily, potentially based on the history of the agent's actions $(I_1,\ldots,I_{t-1})$ and the adversary's own internal randomization. In the *stochastically constrained adversarial setting* [26] the adversary is required to pick the best arm i^* and to sample losses from distributions that maintain a fixed gap to the best arm, $\mathbb{E}\left[\ell_{t,i}-\ell_{t,i^*}\right]=:\Delta_i\geq 0$. We emphasize that the means, as well as other parameters of the distributions of all arms are allowed to change with time and may depend on the agent's past actions $I_1,\ldots,I_{t-1}$. *Stochastic environments* are a special case of stochastically constrained adversarial setting, where the means are fixed throughout the game.

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

$$\begin{aligned}\overline{Reg}_T :=&\mathbb{E}\left[\sum_{t=1}^T \ell_{t,I_t}\right]-\min_i \mathbb{E}\left[\sum_{t=1}^T \ell_{t,i}\right]\\ =&\mathbb{E}\left[\sum_{t=1}^T\left(\ell_{t,I_t}-\ell_{t,i_T^*}\right)\right],\end{aligned}$$

where $i_T^*\in\arg\min_i \mathbb{E}\left[\sum_{t=1}^T \ell_{t,i}\right]$ is defined as a best arm in expectation in hindsight and the expectation is taken over internal randomization of the algorithm and the environment. For deterministic oblivious adversaries the definition of pseudo-regret coincides with the expected regret defined as $\mathbb{E}[Reg_T]:=$

[2] This does not contradict our result, because we bound the pseudo-regret, which is weaker than the expected regret.

$\mathbb{E}\left[\min_i \sum_{t=1}^T (\ell_{t,I_t} - \ell_{t,i})\right]$. In the stochastically constrained adversarial setting a best arm is fixed, $i_T^* = i^*$ for all T (if there is more than one best arm we can pick one arbitrarily), and the pseudo regret can be rewritten as

$$\overline{Reg}_T = \sum_{t=1}^T \sum_i \Delta_i \mathbb{P}\left[I_t = i\right].$$

3 Online Mirror Descent

We recall a number of basic definitions and facts from convex analysis. The convex conjugate[3] of a function $f : \mathbb{R}^K \to \mathbb{R}$ is defined by

$$f^*(y) = \max_{x \in \mathbb{R}^K} \left\{\langle x, y\rangle - f(x)\right\}.$$

We use

$$\mathcal{I}_A(x) := \begin{cases} 0, & \text{if } x \in A \\ \infty, & \text{otherwise} \end{cases}$$

to denote the characteristic function of a closed and convex set $A \subset \mathbb{R}^K$. Hence, $(f + \mathcal{I}_A)^*(y) = \max_{x \in A} \left\{\langle x, y\rangle - f(x)\right\}$. By standard results from convex analysis [20], for differentiable and convex f with invertible gradient $(\nabla f)^{-1}$ it holds that

$$\nabla (f + \mathcal{I}_A)^*(y) = \arg\max_{x \in A} \left\{\langle x, y\rangle - f(x)\right\} \in A.$$

3.1 General Framework

The traditional online mirror descent framework [23] uses a fixed regularizer Ψ, with certain regularity constraints. The update rule is $w_{t+1} = \nabla \Psi^*(-\sum_{s=1}^t a_s \ell_s)$, where $\sum_{s=1}^t a_s \ell_s$ is a weighted sum of past losses. This setting has been generalized to time-varying regularizers Ψ_t [18], where the updates are given by $w_{t+1} = \nabla \Psi_t^*(-\sum_{s=1}^t \ell_s)$. Note that this formulation uses no weighting a_s of the losses. In the bandit setting we do not observe the complete loss vector ℓ_t. Instead, an unbiased estimator $\hat{\ell}_t : \mathbb{E}_{I_t \sim w_t}\left[\hat{\ell}_t\right] = \ell_t$ is used for updating the cumulative losses. At every step, we need to choose a probability distribution over arms w_t, so we add $\mathcal{I}_{\Delta^K}$ to the regularizers Ψ_t, thereby ensuring that $w_t \in \Delta^K$.

The algorithm is provided in Algorithm 1. Note that this framework is equivalent to what Abernethy et al. [2] call Gradient-Based Prediction (Gbp), where they replace $\nabla(\Psi_t + \mathcal{I}_{\Delta^K})^*$ with suitable functions $\nabla \Phi_t : \mathbb{R}^K \to \Delta^K$. We adopt the notation of $\Phi_t := (\Psi_t + \mathcal{I}_{\Delta^K})^*$.

[3] Also known as Fenchel conjugate.

Algorithm 1: Online Mirror Descent (OMD) for bandits

Input: $(\Psi_t)_{t=1,2,\ldots}$
Initialize: $\hat{L}_0 = \mathbf{0}_K$ (the zero vector of dimension K)
for $t = 1, \ldots$ **do**
  choose $w_t = \nabla(\Psi_t + \mathcal{I}_{\Delta^K})^*(-\hat{L}_{t-1})$
  sample $I_t \sim w_t$
  observe ℓ_{t,I_t}
  construct $\hat{\ell}_t$
  update $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$

3.2 OMD with Tsallis Entropy Regularization

We now consider a family of algorithms that is parameterized by the (negative) α-Tsallis entropy: $H_\alpha(x) := \frac{1}{1-\alpha}(1 - \sum_i x_i^\alpha)$. Constant terms in Ψ_t do not change the gradient $\nabla \Phi_t$. We change the scaling of α-Tsallis entropy and drop all constant terms, resulting in the following time-dependent regularizers: $\Psi_{t,\alpha}(w) := -\sum_i \frac{w_i^\alpha}{\alpha \eta_{t,i}}$. This family of algorithms is a subset of Inf [6], which we call α-Tsallis-Inf. When the learning rate $\eta_{t,i}$ is constant over time and arms, the algorithm is equivalent to the Gbp algorithm proposed by Abernethy et al. [3].

As it has been observed earlier [3; 4], α-Tsallis-Inf includes Exp3 and algorithms based on the log-barrier potential as special cases. This can be easily seen by taking the limit of the properly scaled and shifted regularizers:

$$\lim_{\alpha \to 1} \frac{1}{1-\alpha}(1 - \sum_i x_i^\alpha) = \sum_i x_i \log(x_i),$$
$$\lim_{\alpha \to 0} \frac{1}{\alpha}(K - \sum_i x_i^\alpha) = -\sum_i \log(x_i).$$

We use importance sampling to construct the loss estimates $\hat{\ell}_{t,i} = \mathbb{1}_t(i)\frac{\ell_{t,i}}{w_{t,i}}$, where $\mathbb{1}_t(i)$ is used as a shorthand for the indicator function $\mathbb{1}\left(I_t = i\right)$.

4 Main results

In this section we present regret bounds for Tsallis-Inf in adversarial and stochastically constrained adversarial regimes. The latter also provides bounds for the stochastic regime, since it is a special case.

4.1 Adversarial Regime

Tsallis-Inf has been previously analyzed by Abernethy et al. [3] and Agarwal et al. [4]. Abernethy et al. provide a finite-time analysis for $\alpha \in (0, 1]$, while Agar-

wal et al. analyze the case of $\alpha = 0$. The main contribution of the following theorem is that it provides a unified and anytime treatment of all $\alpha \in [0,1]$. The bound recovers the constants from Abernethy et al. without the need of tuning the learning rate by the time horizon T.

Theorem 1. *For any $\alpha \in [0,1]$, and any adversarial bandit problem the pseudo-regret at any time T of* Tsallis-Inf *with the learning rate $\eta_{t,i} = \eta_t = \sqrt{\frac{(1-\alpha)(K^{1-2\alpha}-K^{-\alpha})(1-t^{-\alpha})}{\alpha t}}$ satisfies*

$$\overline{Reg}_T \leq 2\sqrt{\min\left\{\frac{1}{\alpha-\alpha^2}, \frac{\log(K)}{\alpha}, \frac{\log(T)}{1-\alpha}\right\} KT} + 1.$$

The proof is postponed to section 7.

4.2 Stochastically Constrained Adversarial Regime

We show that with a carefully tuned learning rate Tsallis-Inf can achieve $\log(T)$ regret rate in stochastically constrained adversarial environments. This ensures the same regret scaling in stochastic environments as a special case. At the moment, for any $\alpha \neq \frac{1}{2}$ the algorithm requires oracle access to the gaps for tuning the learning rates. We leave it to future research to explore the possibility of replacing the true gaps with gap estimates, as in Seldin and Lugosi [21]. For $\alpha = \frac{1}{2}$ we provide a tuning of the learning rate that requires no knowledge of the gaps and achieves the optimal problem-dependent constant of $\sum_{i \neq i^*} \frac{1}{\Delta_i}$ in the regret bound.

The learning rates in this section are $\eta_{t,i} = \Theta(\Delta_i^{1-2\alpha} t^{-\alpha})$. Note that for $\alpha = \frac{1}{2}$ the learning rates for all arms are identical, $\eta_{t,i} = \eta_t$, and do not depend on the gaps. Therefore, taking $\alpha = \frac{1}{2}$ circumvents the need of tuning the learning rates based on the unknown gaps, which has hindered progress in prior work [26].

The rationale behind the selection of the learning rate is the following. The target regret of $\Theta(\sum_{i \neq i^*} \frac{\log t}{\Delta_i})$ dictates that suboptimal arms should be explored at a rate of $\Theta(\frac{1}{\Delta_i^2 t})$ per round. Exploring more than that leads to excessive regret from the exploration alone. Exploring less is also prohibitive, because it leads to an overly high probability of misidentifying the best arm. Tsallis-Inf pulls suboptimal arms at a rate of about $(\eta_{t,i}(\hat{L}_{t,i} - \hat{L}_{t,\hat{i}_t^*}))^{-\frac{1}{1-\alpha}}$. Since $\mathbb{E}\left[\hat{L}_{t,i} - \hat{L}_{t,\hat{i}_t^*}\right] = \Delta_i t$, it seems straightforward to use a learning rate that ensures $(\eta_{t,i}\Delta_i t)^{-\frac{1}{1-\alpha}} = \Theta\left(\frac{1}{\Delta_i^2 t}\right)$. However, previous techniques to analyze the regret faced major obstacles. Since $x^{-\frac{1}{1-\alpha}}$ is a convex function, $\mathbb{E}\left[(\eta_{t,i}(\hat{L}_{t,i} - \hat{L}_{t,\hat{i}_t^*}))^{-\frac{1}{1-\alpha}}\right] \geq (\eta_{t,i}\Delta_i t)^{-\frac{1}{1-\alpha}}$. Therefore, control of the exploration rate requires control of the variance of $(\hat{L}_{t,i} - \hat{L}_{t,\hat{i}_t^*})$. Due to importance sampling, the variance of $\hat{L}_{t,i}$ is of order $\sum_{s=1}^{t} \frac{1}{w_{s,i}}$, where $w_{s,i}$ is the probability of pulling arm i at round s. If the suboptimal arms are pulled according to the optimal rate of $\frac{1}{\Delta_i^2 t}$, then the variance is of order $\Theta(\Delta_i^2 t^2)$. This is prohibitively large, because the square root of the variance is of the same order as the expected difference of the cumulative losses and standard tools, such as Bernstein's inequality, do not guarantee concentration of $\hat{L}_{t,i} - \hat{L}_{t,\hat{i}_t^*}$ around $\Delta_i t$. Therefore, in prior work authors have increased the targeted rate of exploration, thereby decreasing the variance at the cost of multiplicative $\log T$ factor(s) in the regret [22; 21].

We circumvent this challenge with a novel way of bounding the regret. Following a refined analysis analogous to the adversarial case, we demonstrate a self-bounding property of the regret, leading to our main theorem that bounds the regret of $\frac{1}{2}$-Tsallis-Inf in adversarial and stochastically constrained adversarial regimes simultaneously.

Theorem 2. *For learning rates $\eta_{t,i} = \eta_t = \sqrt{\frac{1}{t}}$, the pseudo-regret of $\frac{1}{2}$-Tsallis-Inf in any adversarial bandit problem satisfies:*

$$\overline{Reg}_T \leq 4\sqrt{KT} + 1.$$

If the optimal arm i_T^* is unique throughout the game and there exists a gap vector Δ, such that the pseudo regret at time T satisfies

$$\mu\mathbb{E}\left[\sum_{i \neq i^*} \sum_{t=1}^{T} w_{t,i}\Delta_i\right] \leq \overline{Reg}_T, \tag{1}$$

then the pseudo-regret further satisfies

$$\overline{Reg}_T \leq \sum_{i \neq i^*} \frac{4\log(T) + 68}{\mu\Delta_i} + 4\sqrt{K}.$$

The proof is postponed to section 7.

Remark 1. *In stochastically constrained adversarial environments and stochastic bandits as their special case $\overline{Reg}_T = \mathbb{E}\left[\sum_{i \neq i^*} \sum_{t=1}^{T} w_{t,i}\Delta_i\right]$ by definition and $\mu = 1$. The relaxed condition in equation (1) is required for extension of the results to contaminated stochastic bandits.*

Remark 2. *Uniqueness of the best arm is a technical condition we had to use in our proofs, but our experiments show that it can most likely be eliminated. We leave elimination of this condition for future work.*

As a direct corollary, we obtain an asymptotic regret bound for the standard multi-armed bandits (MAB), which shows that $\frac{1}{2}$-Tsallis-Inf is optimal (up to constants) in both stochastic and adversarial environments simultaneously.

Corollary 1. *The asymptotic regret of* $\frac{1}{2}$-Tsallis-Inf *with learning rates* $\eta_t = \sqrt{\frac{1}{t}}$ *in a stochastic* K*-armed MAB with a unique best arm* i^* *satisfies*

$$\liminf_{t\to\infty} \frac{\overline{Reg}_t}{\log(t)} \leq \sum_{i\neq i^*} \frac{4}{\Delta_i}.$$

The worst case lower bound for Stochastic MAB with Bernoulli losses is achieved when the mean losses are close to $\frac{1}{2}$. Let $\mathbb{E}[\ell_{t,i}] = \frac{1}{2} + \Delta_i$ and let Δ denote the vector of gaps, then for any consistent algorithm

$$\lim_{||\Delta||\to 0} \left(\left(\sum_{i\neq i^*} \frac{1}{\Delta_i} \right)^{-1} \liminf_{t\to\infty} \frac{\mathbb{E}\left[\overline{Reg}_t\right]}{\log(t)} \right) \geq \frac{1}{2}.$$

This can be derived from the well known divergence dependent lower bound [17] (see Appendix B). Therefore, the asymptotic regret upper bound of $\frac{1}{2}$-Tsallis-Inf is suboptimal by a factor of 8, which is arguably a small price for a significant gain in robustness. We leave it to future work to close the gap or prove that it is impossible to do so without losing adversarial guarantees.

Finally, we present a full characterization of Tsallis-Inf with $\alpha \in [0,1]$ in the stochastic setting. We let $\bar{t} = \max\{e, t\}$. For learning rates $\eta_{t,i} = \Delta_i^{1-2\alpha} \frac{32^\alpha}{8} \left(\frac{1-\bar{t}^{-1+\alpha}}{t}\right)^\alpha$ for $i \neq i^*$ and $\eta_{t,i^*} = \min_{i\neq i^*} \eta_{t,i}$ we prove the following theorem.

Theorem 3. *For any* $\alpha \in [0,1]$ *and any stochastically constrained adversarial regime with a unique best arm, there exists an arm-dependent learning rate schedule, such that the pseudo-regret of* Tsallis-Inf *at any time* T *satisfies*

$$\overline{Reg}_T \leq \sum_{i\neq i^*} \left(\frac{56 \min\{\frac{1}{1-\alpha}, \log(T)\} \log(T)}{\Delta_i} + \frac{144 \log(\frac{8}{\Delta_i})^2}{\Delta_i} \right) + 2.$$

A proof is provided in Appendix E.

Remark 3. *We reemphasize that for* $\alpha \neq \frac{1}{2}$ *the result in Theorem 3 requires knowledge of the gaps* Δ_i *for tuning the learning rate. For* $\alpha = \frac{1}{2}$ *this knowledge is not required. Therefore, at the moment Theorem 3 is primarily interesting from the theoretical perspective of characterization of behavior of* Tsallis-Inf *in stochastically constrained adversarial environments, whereas* $\alpha = \frac{1}{2}$ *is the only practically interesting value.*

5 Moderately Contaminated Stochastic Regime

In many real world examples the systems behave mostly stochastically, but not all the time. In such situations it is desirable to stay close to the $\log(T)$ bound instead of resorting to the weaker $\sqrt{T}$ worst-case regret guarantee. In order to model such situations Seldin and Slivkins [22] have defined the moderately contaminated stochastic regime. In this regime, the adversary picks some round-arm pairs (t, i) ("locations") before the game starts and assigns the loss values there in an arbitrary way. The remaining losses are generated according to the stochastic regime. A contaminated stochastic regime is called *moderately contaminated after* τ *rounds*, if for all $t \geq \tau$ the total number of contaminated locations for each suboptimal arm up to time t is at most $\frac{t\Delta_i}{4}$ and the number of contaminated locations for the best arm is at most $\min_{\Delta_i} \frac{t\Delta_i}{4}$. By this definition, it follows directly that

$$\frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \sum_{i\neq i^*} w_{t,i}\Delta_i\right] \leq \overline{Reg}_T$$

for $T \geq \tau$. Therefore, by Theorem 2, $\frac{1}{2}$-Tsallis-Inf with $\eta_t = \sqrt{\frac{1}{t}}$ has $\mathcal{O}(\log(T))$ regret guarantee simultaneously with $\mathcal{O}(\sqrt{KT})$ worst-case regret.

6 Dueling Bandits

In the sparring approach to stochastic utility-based dueling bandits [5] each side in the sparring can be modeled by stochastically constrained adversarial environment. This makes it a perfect application domain for $\frac{1}{2}$-Tsallis-INF. The problem is defined by K arms with utilities $u_i \in [0,1]$. At each round, the agent has to select two arms, I_t and J_t, to "duel". The feedback is the winner W_t of the "duel", which is chosen according to $\mathbb{P}[W_t = I_t] = \frac{1+u_{I_t}-u_{J_t}}{2}$. The regret is defined by the distance to the optimal utility:

$$\overline{Reg}_T = \sum_{t=1}^{T} 2u_{i^*} - \mathbb{E}\left[\sum_{t=1}^{T}(u_{I_t} + u_{J_t})\right].$$

In an adversarial version of the problem, the utilities u_i are not constant, but time dependent, $u_{t,i}$, and selected by an adversary. The regret in this case is the difference to the optimal utility in hindsight:

$$\overline{Reg}_T = \max_i \sum_{t=1}^{T} 2u_{t,i} - \mathbb{E}\left[\sum_{t=1}^{T}(u_{t,I_t} + u_{t,J_t})\right].$$

Ailon et al. [5] proposed the Sparring algorithm, in which two black-box MAB algorithms spar with each

other. The first algorithm selects I_t and receives the loss $\ell_{t,I_t} = \mathbb{I}(W_t \neq I_t)$. The second algorithm selects J_t and receives the loss $\ell_{t,J_t} = \mathbb{I}(W_t \neq J_t)$. They have shown that the regret is the sum of individual regret values for both MABs, thereby recovering $\mathcal{O}(\sqrt{KT})$ regret in the adversarial case, if MABs with $\mathcal{O}(\sqrt{KT})$ adversarial regret bound are used. The stochastic case for each black-box MAB is a system with stochastically constrained adversary. Since no algorithm has been known to achieve $\log(T)$ for this problem, Ailon et al. [5] provide no analysis of SPARRING in the stochastic case. Applying Theorem 1 and Theorem 2, we immediately get

Corollary 2. SPARRING *with two independent versions of* $\frac{1}{2}$-TSALLIS-INF *with learning rates* $\eta_t = \sqrt{\frac{1}{t}}$ *suffers a regret of*

$$\overline{Reg}_T \leq \mathcal{O}\left(\sum_{i:\Delta_i>0} \frac{\log(T)}{\Delta_i}\right) + \mathcal{O}(K)$$

in the stochastic case and

$$\overline{Reg}_T \leq \mathcal{O}\left(\sqrt{KT}\right)$$

in the adversarial case.

7 Proofs

In this section, we provide proofs of Theorems 1 and 2. A proof of Theorem 3 along with proofs of all lemmas in this section are provided in the Appendix.

7.1 Proof of Theorem 1

Lemma 1. *For any* α *and any learning rate schedule the loss of* TSALLIS-INF *at any time* t *satisfies*

$$\ell_{t,I_t} \leq \frac{\eta_{t,I_t}}{2(1-\alpha){w_{t,I_t}}^{\alpha}} + \Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t).$$

Lemma 2. *For any* α *and any arm-independent non-increasing learning rate schedule the sum of potential differences of* TSALLIS-INF *satisfies*

$$\sum_{t=1}^{T} \Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t) \leq \frac{(K^{1-\alpha}-1)(1-T^{-\alpha})}{\alpha\eta_T} + (1-T^{-1})\hat{L}_{T,i_T^*} + \frac{T^{-1}}{K-1}\sum_{i\neq i_T^*}\hat{L}_{T,i}.$$

Proof of Theorem 1. Using Lemma 1 and Lemma 2, we can bound the sum of the agent's losses as

$$\sum_{t=1}^{T}\ell_{t,I_t} \leq \sum_{t=1}^{T}\frac{\eta_t}{2(1-\alpha){w_{t,I_t}}^{\alpha}} + \frac{(K^{1-\alpha}-1)(1-T^{-\alpha})}{\alpha\eta_T} + (1-T^{-1})\hat{L}_{T,i_T^*} + \frac{T^{-1}}{K-1}\sum_{i\neq i_T^*}\hat{L}_{T,i}.$$

Subtracting the optimal loss $L_{i_T^*}$, taking expectation over both sides and using $\mathbb{E}[L_{i_T^*}] = \mathbb{E}[\hat{L}_{T,i_T^*}]$ leads to

$$\begin{aligned}\overline{Reg}_T &= \mathbb{E}\left[\left(\sum_{t=1}^{T}\ell_{t,I_t}\right) - L_{i_T^*}\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta_t}{2(1-\alpha){w_{t,I_t}}^{\alpha}}\right] + \frac{(K^{1-\alpha}-1)(1-T^{-\alpha})}{\alpha\eta_T} \\ &\quad + \mathbb{E}\left[-T^{-1}\hat{L}_{T,\hat{i}_T^*} + \frac{T^{-1}}{K-1}\sum_{i\neq i_T^*}\hat{L}_{T,i}\right].\end{aligned}$$

The first expectation can be expressed as $\mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta_t}{2(1-\alpha){w_{t,I_t}}^{\alpha}}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_i \frac{\mathbb{1}_t(i)\eta_t}{2(1-\alpha){w_{t,I_t}}^{\alpha}}\right]$. We note that the conditional expectation of $\mathbb{1}_t(i)$ (conditioned on all the randomness prior to selection of I_t) is $w_{t,i}$. Therefore, $\mathbb{E}\left[\sum_{t=1}^{T}\sum_i \frac{\mathbb{1}_t(i)\eta_t}{2(1-\alpha){w_{t,I_t}}^{\alpha}}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_i \frac{\eta_t {w_{t,i}}^{1-\alpha}}{2(1-\alpha)}\right] \leq \left(\sum_{t=1}^{T}\eta_t\right)\frac{K^{\alpha}}{2(1-\alpha)}$, where we use that $w_i = K^{-1}$ maximizes $\sum_i {w_{t,i}}^{1-\alpha}$. We use that $0 \leq \mathbb{E}[\hat{L}_{T,i}] \leq T$ to obtain

$$\begin{aligned}\overline{Reg}_T &= \mathbb{E}\left[\left(\sum_{t=1}^{T}\ell_{t,I_t}\right) - L_{i_T^*}\right] \\ &\leq \left(\sum_{t=1}^{T}\eta_t\right)\frac{K^{\alpha}}{2(1-\alpha)} + 1 + \frac{(K^{1-\alpha}-1)(1-T^{-\alpha})}{\alpha\eta_T}.\end{aligned} \tag{2}$$

Finally, we plug in the learning rate $\eta_t = \sqrt{\frac{(1-\alpha)(K^{1-2\alpha}-K^{-\alpha})(1-t^{-\alpha})}{\alpha t}}$, bound $\sum_{t=1}^{T}\sqrt{\frac{1-t^{-\alpha}}{t}} \leq \sum_{t=1}^{T}\sqrt{\frac{1-T^{-\alpha}}{t}} \leq 2\sqrt{T(1-T^{-\alpha})}$ and get

$$\overline{Reg}_T \leq 2\sqrt{\left(\frac{1-K^{\alpha-1}}{1-\alpha}\right)\left(\frac{1-T^{-\alpha}}{\alpha}\right)KT} + 1.$$

The first factor is bounded by $\sqrt{\frac{1}{1-\alpha}}$ and monotonically increasing in α with the limit $\lim_{\alpha\to 1}\sqrt{\frac{1-K^{\alpha-1}}{1-\alpha}} = \sqrt{\log(K)}$ (details in Lemma 5 in the Appendix). By the same argument, the second factor is bounded by $\sqrt{\frac{1}{\alpha}}$ and monotonically decreasing in α with the limit $\lim_{\alpha\to 0}\sqrt{\frac{1-T^{-\alpha}}{\alpha}} = \sqrt{\log(T)}$. □

7.2 Proof of Theorem 2

Lemma 3. *The loss at time* t *of* $\frac{1}{2}$-TSALLIS-INF *with learning rates* $\eta_t = \sqrt{\frac{1}{t}}$ *satisfies*

$$\ell_{t,I_t} \leq \sum_{i\neq I_t}\frac{4\eta_t w_{t,i}}{w_{t,I_t}} + \Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t).$$

Lemma 4. *The expected sum of potentials of* $\frac{1}{2}$-Tsallis-Inf *with learning rates* $\eta_t = \sqrt{\frac{1}{t}}$ *satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T} \Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*}\sqrt{\frac{w_{t,i}}{t}}\right] + \mathbb{E}[L_{i^*}] + 2\sqrt{K}.$$

Proof of Theorem 3. The adversarial part follows directly from line (2) in the previous proof. For the second part of the theorem, we start by bounding the expectation of ℓ_{t,I_t}. By using Lemma 1 when $I_t \neq i^*$ and Lemma 3 otherwise we have for all t:

$$\ell_{t,I_t} \leq \sum_{i\neq i^*}\left(\frac{\mathbb{1}_t(i)\eta_{t,i}}{\sqrt{w_{t,i}}} + \frac{4\mathbb{1}_t(i^*)\eta_{t,i}w_{t,i}}{w_{t,I_t}}\right) + \Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t).$$

We note again that the conditional expectation of $\mathbb{1}_t(i)$ is $w_{t,i}$. Therefore, taking the expectation and plugging in the learning rate $\eta_{t,i} = \sqrt{\frac{1}{t}}$, we obtain:

$$\mathbb{E}[\ell_{t,I_t}] \leq \mathbb{E}\left[\sum_{i\neq i^*}\left(\frac{\mathbb{1}_t(i)\eta_{t,i}}{\sqrt{w_{t,i}}} + \frac{4\mathbb{1}_t(i^*)\eta_{t,i}w_{t,i}}{w_{t,I_t}}\right)\right] + \mathbb{E}\left[\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t)\right]$$

$$= \mathbb{E}\left[\sum_{i\neq i^*}\frac{\sqrt{w_{t,i}} + 4w_{t,i}}{\sqrt{t}}\right] + \mathbb{E}\left[\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t)\right].$$

In the initial time period we are using Lemma 1 for all actions, including $I_t = i^*$, and in the same way as above obtain:

$$\mathbb{E}[\ell_{t,I_t}] \leq \mathbb{E}\left[\sum_{i}\frac{\sqrt{w_{t,i}}}{\sqrt{t}}\right] + \mathbb{E}\left[\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t)\right]$$

$$\leq \mathbb{E}\left[\sum_{i\neq i^*}\frac{\sqrt{w_{t,i}}}{\sqrt{t}}\right] + \frac{1}{\sqrt{t}} + \mathbb{E}\left[\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t)\right].$$

Summing over t, using the first bound for $t \geq T_0 := \lceil \max_{\Delta_i>0}\frac{256}{\mu^2\Delta_i^2}\rceil$, the second bound otherwise and then applying Lemma 4:

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,I_t}\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*}\frac{\sqrt{w_{t,i}} + \mathbb{I}_{(t>T_0)}4w_{t,i}}{\sqrt{t}}\right] + \sum_{t=1}^{T_0}\frac{1}{\sqrt{t}} + \mathbb{E}\left[\sum_{t=1}^{T}\Phi_t(-\hat{L}_{t-1}) - \Phi_t(-\hat{L}_t)\right]$$

$$\leq \sum_{i\neq i^*}\mathbb{E}\left[\sum_{t=1}^{T_0}\frac{2\sqrt{w_{t,i}}}{\sqrt{t}} + \sum_{t=T_0+1}^{T}\frac{2\sqrt{w_{t,i}} + 4w_{t,i}}{\sqrt{t}}\right] + 2\sqrt{T_0} + 2\sqrt{K} + \mathbb{E}[L_{i^*}].$$

We are using the inequality $a\sqrt{x} + cx \leq bx + \frac{a^2}{4(b-c)}$, which holds for all $a, b, x > 0$ and $c < b$ (see Lemma 8 in the Appendix). Subtracting the optimal expected loss $\mathbb{E}[L_{i^*}]$ on both sides

$$\overline{Reg}_T \leq 2\sqrt{K} + \sum_{i\neq i^*}\mathbb{E}\left[\sum_{t=1}^{T_0}\left(\frac{\mu\Delta_i}{2}w_{t,i} + \frac{2}{\mu\Delta_i t}\right) + \sum_{t=T_0+1}^{T}\left(\frac{\mu\Delta_i}{2}w_{t,i} + \frac{2}{(\mu\Delta_i - 8\sqrt{t^{-1}})t}\right)\right] + 2\sqrt{T_0}$$

$$= \sum_{i\neq i^*}\sum_{t=1}^{T}\mathbb{E}\left[\frac{\mu\Delta_i}{2}w_{t,i} + \frac{2}{\mu\Delta_i t}\right] + 2\sqrt{K} + 2\sqrt{T_0} + \sum_{i\neq i^*}\sum_{t=T_0+1}^{T}\frac{16\sqrt{t}^{-1}}{\mu\Delta_i(\mu\Delta_i - 8\sqrt{t^{-1}})t}.$$

Since for any $t \geq \min_{\Delta_i>0}\frac{256}{\mu^2\Delta_i^2} : \mu\Delta_i - 8\sqrt{t^{-1}} \geq \frac{\mu\Delta_i}{2}$,

$$\sum_{t=T_0+1}^{T}\frac{16\sqrt{t}^{-1}}{\mu\Delta_i(\mu\Delta_i - 8\sqrt{t^{-1}})t} \leq \sum_{t=T_0+1}^{T}\frac{32}{\mu^2\Delta_i^2\sqrt{t^3}} \leq \frac{32}{\mu^2\Delta_i^2\sqrt{T_0}} \leq \frac{2}{\mu\Delta_i}.$$

Bounding $\sqrt{T_0}$ by $\sum_{i\neq\Delta_i}\frac{16}{\mu\Delta_i}$ and using $\overline{Reg}_T \geq \mu\sum_{t=1}^{T}\sum_{i\neq i^*}\mathbb{E}\left[\mu\Delta_i w_{t,i}\right]$:

$$\overline{Reg}_T \leq \frac{\overline{Reg}_T}{2} + \sum_{i\neq i^*}\frac{2\log(T) + 34}{\mu\Delta_i} + 2\sqrt{K}.$$

Subtracting $\frac{1}{2}\overline{Reg}_T$ on both sides and multiplying by 2 concludes the proof. □

8 Experiments

We evaluate the classical Ucb1($\alpha = 1.5$) and Thompson Sampling[4] for stochastic bandit algorithms and Exp3 as an algorithm optimized for adversarial environments. We also evaluate Broad and Exp3++ as state-of-the-art all-purpose algorithms. The pseudo-regret is estimated by 1000 repetitions experiments and we plot the standard deviation of our regret estimation in a shaded area. We always show the first 10000 time steps in a linear plot and then the time horizon from 10^4 to 10^7 in a separate log-log plot.

The first two experiments illustrate the superiority of Tsallis-Inf. In both experiments, we use the same number of arms $K = 8$ and the same gap for all suboptimal arms $\Delta = 0.125$.

[4]Other algorithms, such as KL-UCB and MOSS, perform comparably to Thompson Sampling in our experiments and, therefore, omitted.

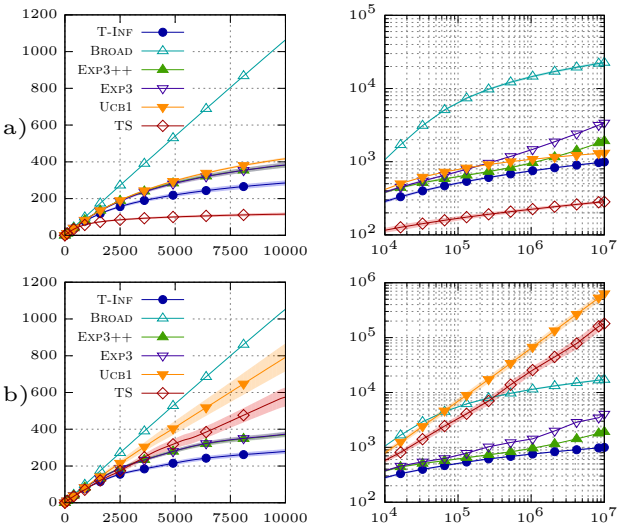

Figure 1: Comparison of several bandit algorithms with $K = 8$ and $\Delta = 1/8$ under a) stochastic and b) stochasticaly constrained adversary regime. The left side is in linear scale and the right is in log-log scale.

The first experiment shown in Figure 1.a) is a standard stochastic MAB, where the mean rewards are given by $(1+\Delta)/2$ and $(1-\Delta)/2$, respectively. Unsurprisingly, Thompson Sampling exhibits the lowest regret followed by Tsallis-Inf and then Ucb1. Exp3++ is a clear improvement over Exp3, however $T = 10^7$ is not large enough to reach the $\log(T)^2$ regime. Broad suffers from extremely large leading factors in the regret and is out of question for practical applications.

The second experiment shown in Figure 1.b) is identical in the number of arms and gaps, however the means are not fixed. The mean loss of (optimal arm, any sub-optimal arm) switches between $(1-\Delta, 1)$ and $(0, \Delta)$, while staying unchanged for phases that are increasing exponentially in length. Both Ucb1 and Thompson-Sampling suffer almost linear regret. To the best of our knowledge, this is the first empirical evidence clearly showing that Thompson Sampling is unsuitable for the adversarial regime. All other algorithms are almost unaffected by the shifting of means, with Tsallis-Inf being the only algorithm that obtains $\log(T)$ regret with practical leading factors.

In the Appendix we provide an additional experiment that addresses the main shortcoming of our proof, namely, that the stochastic guarantee requires uniqueness of the best arm. We use the same gap as before, but only 1 suboptimal arm. The experiment is repeated with increasing number of copies of the best arm, see Figure 2 in the Appendix. We observe that the regret decreases with the growth of the number of suboptimal arms. Therefore, we conjecture that the requirement of uniqueness is merely an artifact of the analysis.

9 Discussion

We have presented a complete characterization of online mirror descent algorithms regularized by Tsallis entropy. As the main contribution, we have shown that the special case of $\alpha = \frac{1}{2}$ achieves optimality in both adversarial and stochastic regimes, while being oblivious to the environment at hand. Thereby, we have closed logarithmic gaps to lower bounds, which were present in existing best-of-both-worlds algorithms. We introduced a novel proof technique based on the self-bounding property of the regret, circumventing the need of controlling the variance of loss estimates. We have provided empirical evidence that our algorithm is competitive with UCB1 in stochastic environments and significantly more robust than UCB1 and Thompson Sampling in non i.i.d. settings. Finally, we have shown that our results extend to two intermediate settings from prior literature, stochastically constrained adversaries and moderately contaminated stochastic regimes, as well as the utility-based dueling bandit problem.

A weak point of the current proof is the requirement of uniqueness of the best arm in stochastic and stochastically constrained adversarial settings. Our experiments suggest that this is most likely an artifact of the analysis and we aim to address this shortcoming in future work.

Another open question is whether it is possible to achieve optimality in problem-independent constants. We have reduced the multiplicative gap with the lower bound down to a factor of 8. There is a potential of reducing the gap to a factor of 4 by using loss estimators with smaller variance, such as $\hat{\ell}_{t,i} = \frac{(\ell_{t,i}-z)}{w_{t,i}}\mathbb{1}_t(i) + z$. However, this requires careful analysis, since parts of the proof rely on non-negative (or lower bounded) losses. On the other hand, it might be more suitable to compare the upper bound with the best achievable regret in the stochastically constrained adversarial setting. To the best of our knowledge, there is no refined lower bound known for this problem.

An additional direction for future research is the application of Tsallis-Inf to further problems. The fact that the algorithm relies solely on importance weighted losses makes it a suitable candidate for partial monitoring games.

References

[1] Y. Abbasi-Yadkori, P. Bartlett, V. Gabillon, A. Malek, and M. Valko. Best of both worlds: Stochastic & adversarial best-arm identification. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2018.

[2] J. Abernethy, C. Lee, A. Sinha, and A. Tewari. Online linear optimization via smoothing. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2014.

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

[4] A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2017.

[5] N. Ailon, Z. Karnin, and T. Joachims. Reducing dueling bandits to cardinal bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

[6] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2009.

[7] J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct), 2010.

[8] P. Auer and C.-K. 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.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3), 2002.

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

[11] L. Besson and E. Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

[12] S. Bubeck. *Bandits games and clustering foundations*. PhD thesis, Université des Sciences et Technologie de Lille-Lille I, 2010.

[13] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1), 2012.

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

[15] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41, 2013.

[16] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An optimal finite time analysis. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2012.

[17] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1), 1985.

[18] F. Orabona, K. Crammer, and N. Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3), 2015.

[19] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5), 1952.

[20] R. T. Rockafellar. *Convex analysis*. Princeton university press, 2015.

[21] Y. Seldin and G. 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.

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

[23] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2), 2012.

[24] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4), 1933.

[25] C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52 (1-2), 1988.

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