Adaptive Best-of-Both-Worlds Algorithm for Heavy-Tailed Multi-Armed Bandits

Jiatai Huang [*1] **Yan Dai** [*1] **Longbo Huang** [1]

Abstract

In this paper, we generalize the concept of heavy-tailed multi-armed bandits to adversarial environments, and develop robust best-of-both-worlds algorithms for heavy-tailed multi-armed bandits (MAB), where losses have α-th ($1 < \alpha \leq 2$) moments bounded by σ^α, while the variances may not exist. Specifically, we design an algorithm `HTINF`, when the heavy-tail parameters α and σ are known to the agent, `HTINF` simultaneously achieves the optimal regret for both stochastic and adversarial environments, without knowing the actual environment type a-priori. When α, σ are unknown, `HTINF` achieves a $\log T$-style instance-dependent regret in stochastic cases and $o(T)$ no-regret guarantee in adversarial cases. We further develop an algorithm `AdaTINF`, achieving $\mathcal{O}(\sigma K^{1-1/\alpha} T^{1/\alpha})$ minimax optimal regret even in adversarial settings, without prior knowledge on α and σ. This result matches the known regret lower-bound (Bubeck et al., 2013), which assumed a stochastic environment and α and σ are both known. To our knowledge, the proposed `HTINF` algorithm is the first to enjoy a best-of-both-worlds regret guarantee, and `AdaTINF` is the first algorithm that can adapt to both α and σ to achieve optimal gap-indepedent regret bound in classical heavy-tailed stochastic MAB setting and our novel adversarial formulation.

1. Introduction

In this paper, we focus on the multi-armed bandit problem with heavy-tailed losses. Specifically, in our setting, there is an agent facing K feasible actions (called bandit arms) to sequentially make decisions on. For each time step $t \in [T]$,[1] each arm $i \in [K]$ is associated with a loss distribution $\nu_{t,i}$ which is unknown to the agent. The only constraint on $\nu_{t,i}$ is that the α-th moment ($\alpha \in (1, 2]$) is bounded by some constant σ^α, i.e., $\mathbb{E}_{\ell \sim \nu_{t,i}}[|\ell|^\alpha] \leq \sigma^\alpha$ for all $t \in [T]$ and $i \in [K]$. However, *neither* α *nor* σ is known to the agent.

At each step t, the agent picks an arm i_t and observes a loss ℓ_{t,i_t} drawn from the distribution ν_{t,i_t}, which is independent to all previous steps. The goal of the agent is to minimize the *pseudo-regret*, which is defined by the expected difference between the loss he suffered and the loss of always pulling the best arm in hindsight (formally defined in Definition 3.2), where the expectation is taken with respect to the randomness both in the algorithm and the environment.

Prior MAB literature mostly studies settings where the loss distributions are supported on a bounded interval I (e.g., $I = [0, 1]$) known to the agent before-hand, which is a special case of our setting where all $\nu_{t,i}$'s are Dirac measures centered within I (Seldin & Slivkins, 2014; Zimmert & Seldin, 2019). By constrast, there is another common existing MAB setting called scale-free MAB (De Rooij et al., 2014; Orabona & Pál, 2018), where the range of losses are not known. In this case, the loss range itself can even depend on other scale parameter of the problem instance (e.g., T and K) rather than being a constant. Our heavy-tailed setting can be seen as an intermediate setting between bounded-loss MAB and scale-free MAB, where loss feedback can be indefinitely large, but not in a completely arbitrary manner. This setting naturally extends classical MAB settings, including bounded-loss MAB and sub-Gaussian-loss MAB.

Following the convention of prior MAB literature, we further distinguish the environment into two typical types. Environment of the first type consists with time homogeneous distributions, i.e., for each $i \in [K]$, $\nu_{t,i} = \nu_{1,i}$ holds for all $t \in [T]$. We call them *stochastic* environments. Bubeck et al. (2013) proved that, for heavy-tailed stochastic bandits, even when α and σ are both known to the agent, an $\Omega(\sigma K^{1-1/\alpha} T^{1/\alpha})$ instance-independent regret and $\Omega(\sigma^{\frac{\alpha}{\alpha-1}} \sum_{i \neq i^*} \log T \Delta_i^{-\frac{1}{\alpha-1}})$ instance-dependent regret is

[*]Equal contribution [1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. Correspondence to: Longbo Huang <longbohuang@tsinghua.edu.cn>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Throughout the paper, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$.

unavoidable, where i^* denotes the optimal arm in hindsight, and $\Delta_i \triangleq \mathbb{E}_{\ell\sim\nu_{1,i}}[\ell] - \mathbb{E}_{\ell\sim\nu_{1,i^*}}[\ell]$ is the sub-optimally gap between i and i^*. They also designed an algorithm that matches these lower-bounds up to logarithmic factors when both α and σ are known.

In the second type of environments, loss distributions can be time inhomogeneous, and we call them *adversarial* environments. To our knowledge, no previous work studied similar adversarial heavy-tailed MAB problems. It can be seen that the instance-independent lower-bound $\Omega(\sigma K^{1-1/\alpha}T^{1/\alpha})$ for stochastic heavy-tailed MAB proved by Bubeck et al. (2013) is also a lower-bound for this adversarial extension.

In this paper, we develop algorithms for heavy-tailed bandits in *both* stochastic and adversarial cases. In contrast to existing (stochastic) heavy-tailed MAB algorithms (Bubeck et al., 2013; Lee et al., 2020) that heavily use well-designed mean estimators for heavy-tailed distributions, our algorithms are mainly designed based on the Follow-the-Regularized-Leader (FTRL) framework, which has been applied in a number of adversarial MAB works (Zimmert & Seldin, 2019; Seldin & Lugosi, 2017). Our proposed algorithms enjoy optimal or near-optimal regret guarantees and require much less prior knowledge compared to prior works. When σ, α are known before-hand, our algorithm *matches* existing gap dependent and independent regret lower-bounds, while previous algorithms suffer extra $\log$-factors (check Table 1 for a comparison). Finally, we propose an algorithm with $\mathcal{O}(\sigma K^{1-1/\alpha}T^{1/\alpha})$ regret even when σ, α are *both unknown*, which shows the existing $\Omega(\sigma K^{1-1/\alpha}T^{1/\alpha})$ lower-bound is tight even when all prior knowledge on σ, α is absent.

1.1. Our Contributions

We first introduce a novel adversarial MAB setting where losses are heavy-tailed, which generalizes the existing heavy-tailed stochastic MAB setting and scalar-loss adversarial MAB setting. Three novel algorithms are proposed. `HTINF` enjoys an optimal best-of-both-worlds regret guarantee when α, σ are known. Without the knowledge of α, σ, `OptTINF` guarantees $o(T)$ adversarial regret (a.k.a. "*no-regret* guarantee") and $\mathcal{O}(\log T)$ gap-dependent bound for stochastically constrained environments. `AdaTINF` guarantees *minimax optimal* $\mathcal{O}(\sigma K^{1-1/\alpha}T^{1/\alpha})$ adversarial regret.

1.1.1. Known α, σ case

When α, σ are both known to the agent, we provide a novel algorithm called **H**eavy-**T**ail Tsallis-INF (`HTINF`, Algorithm 1), based on the Follow-the-Regularized-Leader (FTRL) framework. In `HTINF`, We introduce a novel skipping technique equipped with an *action-dependent* skipping threshold (r_t in Algorithm 1) to handle the heavy-tailed losses, which can be of independent interest.

`HTINF` enjoys the so-called *best-of-both-worlds* property (Bubeck & Slivkins, 2012) to achieve $\mathcal{O}\left(\sigma K^{1-1/\alpha}T^{1/\alpha}\right)$ regret in adversarial settings and $\mathcal{O}\left(\sigma^{\frac{\alpha}{\alpha-1}}\sum_{i\neq i^*}\Delta_i^{-\frac{1}{\alpha-1}}\log T\right)$ regret in stochastically constrained adversarial settings (which contains stochastic cases; see Section 3 for definition) *simultaneously*, without knowing the actual environment type a-priori. The claimed regret bounds both match the corresponding lower-bounds by Bubeck et al. (2013), showing that these bounds are indeed tight even for our adversarial setting.

1.1.2. Unknown α, σ case

When the agent does not access to α and σ, running `HTINF` **opt**imistically with $\alpha = 2$ and $\sigma = 1$ (named `OptTINF`; Algorithm 2) also gives non-trivial regret guarantees. Specifically, we showed that it enjoys a near-optimal regret of $\mathcal{O}\left(\sum_{i\neq i^*}(\frac{\sigma^{2\alpha}}{\Delta_i^{3-\alpha}})^{\frac{1}{\alpha-1}}\log T\right)$ in stochastically constrained adversarial environments and $\mathcal{O}\left(\sigma^\alpha K^{\frac{\alpha-1}{2}}T^{\frac{3-\alpha}{2}}\right)$ regret in adversarial cases, which is still $o(T)$.

We further present another novel algorithm called **Ada**ptive Tsallis-INF (`AdaTINF`, Algorithm 3) for heavy-tailed bandits. Without knowing the heavy-tail parameters α and σ before-hand, `AdaTINF` is capable of guaranteeing an $\mathcal{O}\left(\sigma K^{1-1/\alpha}T^{1/\alpha}\right)$ regret in the adversarial setting, matching the regret lower-bound from Bubeck et al. (2013).

To the best of our knowledge, all prior algorithms for MAB with heavy-tailed losses need to know α before-hand. The proposed two algorithms, `OptTINF` and `AdaTINF`, are the first algorithms to have the adaptivity for *both* unknown heavy-tail parameters α and σ, while achieving near-optimal regrets in stochastic or adversarial settings.

1.2. Related Work

Heavy-tailed losses: The heavy-tailed (stochastic) bandit model was first introduced by Bubeck et al. (2013), where instance-dependent and independent lower-bounds were given. They designed an algorithm nearly matching these lower-bounds (with an extra $\log T$ factor in the gap-independent regret), when σ, α are both known to the agent. Vakili et al. (2013) derived a tighter upper-bound with α, σ and $\min \Delta_i$ all presented to the agent. Kagrecha et al. (2019) gave an algorithm adaptive to σ in a pure exploration setting. Lee et al. (2020) got rid of the requirement of σ, yielding near-optimal regret bounds with a prior knowledge on α only. Moreover, all above algorithms built on the `UCB` framework, which does not directly apply to adversarial environments. One can refer to Table 1 for a comparison.

Other variations with heavy-tailed losses are also studied in the literature, e.g., linear bandits (Medina & Yang, 2016;

Table 1. An overview of the proposed algorithms and related works.

Algorithm	Loss Type	Prior Knowledge	Total Regret
Lower-bounds (Bubeck et al., 2013)	Stochastic[a]	α, σ	$\Omega\left(\sigma^{\frac{\alpha}{\alpha-1}} \sum_{i \neq i^*} \Delta_i^{-\frac{1}{\alpha-1}} \log T\right)$
			$\Omega\left(\sigma K^{1-1/\alpha} T^{1/\alpha}\right)$
`RobustUCB` (Bubeck et al., 2013)	Stochastic	α, σ	$\mathcal{O}\left(\sum_{i \neq i^*} (\frac{\sigma^\alpha}{\Delta_i})^{\frac{1}{\alpha-1}} \log T\right)$ **(optimal)**
			$\mathcal{O}\left(\sigma (K \log T)^{1-1/\alpha} T^{1/\alpha}\right)$ **(sub-optimal for $\log T$ factors)**
Lee et al. (2020)	Stochastic	α; require $\mu_i \in [0,1]$	$\mathcal{O}\left(K^{1-1/\alpha} T^{1/\alpha} \log K\right)$[b] **(sub-optimal for $\log K$ factors)**
`1/2-Tsallis-INF` (Zimmert & Seldin, 2019)	SCA-unique[c]	require $\alpha = 2$ and $[0,1]$-bounded losses	$\mathcal{O}\left(\sum_{i \neq i^*} \frac{1}{\Delta_i} \log T\right)$ **(optimal for $\alpha = 2, \sigma = 1$ case)**
	Adversarial		$\mathcal{O}\left(\sqrt{KT}\right)$ **(optimal for $\alpha = 2, \sigma = 1$ case)**
`HTINF` **(ours)**	SCA-unique	α, σ	$\mathcal{O}\left(\sum_{i \neq i^*} (\frac{\sigma^\alpha}{\Delta_i})^{\frac{1}{\alpha-1}} \log T\right)$ **(optimal)**
	Adversarial		$\mathcal{O}\left(\sigma K^{1-1/\alpha} T^{1/\alpha}\right)$ **(optimal)**
Optimistic `HTINF` **(ours)**	SCA-unique	None	$\mathcal{O}\left(\sum_{i \neq i^*} \left(\frac{\sigma^{2\alpha}}{\Delta_i^{3-\alpha}}\right)^{\frac{1}{\alpha-1}} \log T\right)$
	Adversarial		$\mathcal{O}(\sigma^\alpha K^{\frac{\alpha-1}{2}} T^{\frac{3-\alpha}{2}})$
`AdaTINF` **(ours)**	Adversarial	None[d]	$\mathcal{O}\left(\sigma K^{1-1/\alpha} T^{1/\alpha}\right)$ **(optimal)**

[a] As discussed in Section 3, the instance-independent lower bounds automatically apply to adversarial settings, and the main result of this paper shows that it is indeed tight even for adversarial settings.

[b] Lee et al. (2020) regarded σ as a constant when stating their regret bounds. By designing different estimators, they also gave various instance-dependent bounds, each with $(\log T)^{\frac{\alpha}{\alpha-1}}$ (sub-optimal) dependency on T. One can check Table 1 in their paper for more details.

[c] Abbreviation for stochastically constrained adversarial settings with a unique optimal arm.

[d] Though the time horizon T is assumed to be known in Algorithm 3, it is in fact non-essential for `AdaTINF`. The removal of T, via a usual doubling trick, will not cause extra factors. Check Appendix D for more discussions.

Xue et al., 2021), contextual bandits (Shao et al., 2018) and Lipschitz bandits (Lu et al., 2019). However, none of above algorithms removes the dependency on α.

Best-of-both-worlds: This concept of designing a single algorithm to yield near-optimal regret in both stochastic and adversarial environments was first proposed by Bubeck & Slivkins (2012). Bubeck & Slivkins (2012); Auer & Chiang (2016); Besson & Kaufmann (2018) designed algorithms that initially run a policy for stochastic settings, and may permanently switch to a policy for adversarial settings during execution. Seldin & Slivkins (2014); Seldin & Lugosi (2017); Wei & Luo (2018); Zimmert & Seldin (2019) designed algorithm using the Online Mirror Descent (OMD) or Follow the Regularized Leader (FTRL) framework. Our work falls into the second category.

Adaptive algorithms: There is a rich literature in deriving algorithms adaptive to the loss sequences, for either full information setting (Luo & Schapire, 2015; Orabona & Pal, 2016), stochastic bandits (Garivier & Cappé, 2011; Lattimore, 2015) or adversarial bandits (Wei & Luo, 2018; Bubeck et al., 2019). There are also many algorithms that is adaptive to the loss range, so-called 'scale-free' algorithms (De Rooij et al., 2014; Orabona & Pál, 2018; Hadiji & Stoltz, 2020). However, as mentioned above, to our knowledge, our work is the first to adapt to heavy-tail parameters.

2. Notations

We use $[N]$ to denote the integer set $\{1, 2, \cdots, N\}$. Let f be any strictly convex function defined on a convex set $\Omega \subseteq \mathbb{R}^K$. For $x, y \in \Omega$, if $\nabla f(x)$ exists, we denote the Bregman divergence induced by f as

$$D_f(y, x) \triangleq f(y) - f(x) - \langle \nabla f(x), y - x \rangle.$$

We use $f^*(y) \triangleq \sup_{x \in \mathbb{R}^K} \{\langle y, x \rangle - f(x)\}$ to denote the Fenchel conjugate of f. Denote the $K-1$-dimensional probability simplex by $\triangle_{[K]} = \{x \in \mathbb{R}_+^K \mid x_1 + x_2 + \cdots + x_K = 1\}$. We use $\mathbf{e}_i \in \triangle_{[K]}$ to denote the vector whose i-th coordinate is 1 and others are 0.

Let $\overline{f}$ denote the restriction of f on $\triangle_{[K]}$, i.e.,

$$\overline{f}(x) = \begin{cases} f(x), & x \in \triangle_{[K]} \\ \infty, & x \notin \triangle_{[K]} \end{cases}.$$

Let $\mathcal{E}$ be a random event, we use $\mathbb{1}[\mathcal{E}]$ to denote the indicator of $\mathcal{E}$, which equals 1 if $\mathcal{E}$ happens, and 0 otherwise.

3. Problem Setting

We now introduce our formulation of the heavy-tailed MAB problem. Formally speaking, there are $K \geq 2$ available arms indexed from 1 to K, and $T \geq 1$ time slots for the agent to make decisions sequentially. $\{\nu_{t,i}\}_{t \in [T], i \in [K]}$ are $T \times K$ probability distributions over real numbers, which are fixed before the game starts and unknown to the agent (i.e., obliviously adversely chosen). Instead of the usual assumption of bounded variance or even bounded range, we only assume that they are *heavy-tailed*, as follows.

Assumption 3.1 (Heavy-tailed Losses Assumption)**.** The α-th moment of all loss distributions $\{\nu_{t,i}\}$ are bounded by σ^α for some constants $1 < \alpha \leq 2$ and $\sigma > 0$, i.e.,

$$\mathop{\mathbb{E}}_{\ell \sim \nu_{t,i}} [|\ell|^\alpha] \leq \sigma^\alpha, \quad \forall t \in [T], i \in [K].$$

In this paper, we will discuss how to design algorithms for two different cases: α and σ are known before-hand or oblivious (i.e., fixed before-hand but unknown to the agent). We denote by $\mu_{t,i} \triangleq \mathbb{E}_{x \sim \nu_{t,i}}[x]$ the individual mean loss for each arm and $\mu_t \triangleq (\mu_{t,1}, \mu_{t,2}, \ldots, \mu_{t,K})$ the mean loss vector at time t, respectively.

At the beginning of each time slot t, the agent needs to choose an action $i_t \in [K]$. At the end of time slot t, the agent will receive and suffer a loss ℓ_{t,i_t}, which is guaranteed to be an independent sample from the distribution ν_{t,i_t}. The agent is allowed to make the decision i_t based on all history actions $i_1, \ldots, i_{t-1}$, all history feedback $\ell_{1,i_1}, \ldots, \ell_{t-1,i_{t-1}}$, and any amount of private randomness of the agent.

The objective of the agent is to minimize the total loss. Equivalently, the agent aims to minimize the following *pseudo-regret* defined by Bubeck & Slivkins (2012) (also referred to as the regret in this paper for simplicity):

Definition 3.2 (Pseudo-regret)**.** We define

$$\begin{aligned} \mathcal{R}_T &\triangleq \max_{i \in [K]} \mathbb{E}\left[\sum_{t=1}^T \ell_{t,i_t} - \sum_{t=1}^T \ell_{t,i}\right] \\ &= \max_{i \in [K]} \mathbb{E}\left[\sum_{t=1}^T \mu_{t,i_t} - \sum_{t=1}^T \mu_{t,i}\right] \end{aligned} \tag{1}$$

to be the **pseudo-regret** of an MAB algorithm, where the expectation is taken with respect to randomness from both the algorithm and the environment.

In the remaining of this paper, we will use $\mathcal{F}_t \triangleq \sigma(i_1, \cdots, i_t, \ell_{1,i_1}, \cdots, \ell_{t,i_t})$ to denote the natural filtration of an MAB algorithm execution.

3.1. Stochastically Constrained Environments

Definition 3.3 (Stochastic Environments)**.** If, for each arm $i \in [K]$, all T loss distributions $\nu_{1,i}, \nu_{2,i}, \ldots, \nu_{T,i}$ are identical, we call such environment a stochastic environment.

A more general setting is called *stochastically constrained adversarial* setting (Wei & Luo, 2018), defined as follows.

Definition 3.4 (Stochastically Constrained Adversarial Environments)**.** If, there exists an *optimal* arm $i^* \in [K]$ and mean gaps $\Delta_i \geq 0$ such that for all $t \in [T]$, we have $\mu_{t,i} - \mu_{t,i^*} \geq \Delta_i$ for all $i \neq i^*$, we call such environment a stochastically constrained adversarial environment.

It can be seen that stochastic problem instances are special cases of stochastically constrained adversarial instances. Hence, in this paper, we study this more general setting instead of stochastic cases. As in Zimmert & Seldin (2019), we make the following assumption.

Assumption 3.5 (Unique Optimal Arm Assumption)**.** In stochastically constrained adversarial environments, i^* is the unique best arm throughout the process, i.e.,

$$\Delta_i > 0, \quad \forall i \neq i^*.$$

Remark. The existence of a unique optimal arm is a common assumption in MAB and RL literature leveraging FTRL with Tsallis entropy regularizers (Zimmert & Seldin, 2019; Erez & Koren, 2021; Jin & Luo, 2020; Jin et al., 2021). Recently, Ito (2021) gave a new analysis of Tsallis-INF's logarithmic regret on stochastic MAB instances without this assumption. It is an interesting future work to figure out whether it is doable in our heavy-tailed losses setting.

3.2. Adversarial Environments

In contrast, an environment without any extra requirement is called an adversarial environment. We denote the best arm(s) in hindsight by i^*, i.e., the $i \in [K]$ that makes the expectation in Eq. (1) maximum. We make the following assumption on the losses of arm i^*.

Assumption 3.6 (Truncated Non-nagative Losses Assumption)**.** There exists an optimal arm i^* such that ℓ_{t,i^*} is *truncated non-negative* for all $t \in [T]$.

In the assumption, the truncated non-negative property is defined as follows.

Definition 3.7 (Truncated Non-negativity)**.** A random variable X is truncated non-negative, if for any $M \geq 0$,

$$\mathbb{E}\left[X \cdot \mathbb{1}[|X| > M]\right] \geq 0.$$

Remark. This truncated non-negative requirement is *strictly weaker* than the common non-negative losses assumption in MAB literature, especially works fitting in the FTRL framework (Auer et al., 2002; Zimmert & Seldin, 2019). Intuitively, truncated non-negativity forbids the random variable to hold too much mass on its negative part, but it can still have negative outcomes.

4. Static Algorithm: `HTINF`

In this section, we first present an algorithm achieving optimal regrets when α, σ are both known before-hand, and then extend it to the unknown α, σ case.

4.1. Known α, σ Case

For the case where both α and σ are known a-priori, we present a FTRL algorithm with the $\frac{1}{\alpha}$-Tsallis entropy function $\Psi(\mathbf{x}) = -\alpha \sum_{i=1}^{K} x_i^{1/\alpha}$ (Tsallis, 1988; Abernethy et al., 2015; Zimmert & Seldin, 2019) as the regularizer. We pick $\eta = t^{-1/\alpha}$ as the learning rate of the FTRL algorithm. Importance sampling is used to construct estimates $\hat{\ell}_t$ of the true loss feedback vector ℓ_t.

In this algorithm, to handle the heavy-tailed losses, we designed a novel *skipping* technique with action-dependent threshold $r_t \propto \eta_t^{-1} x_{t,i_t}^{1/\alpha}$ at time slot t, i.e., the agent simply discards those time slots with the absolute value of the loss feedback more than r_t. Note that this skipping criterion with dependency on i_t is properly defined, for it is checked *after* deciding the arm i_t and receiving the feedback. To decide x_t, the probability to pull each arm in a new time step, we pick the best mixed action x against the sum of all non-skipped estimated loss $\hat{\ell}_t$'s, in a regularized manner. The pseudo-code of the algorithm is presented in Algorithm 1.

The performance of Algorithm 1 is presented in the following Theorem 4.1. The proof is sketched in Section 5. For a detailed formal proof, see Appendix A.

Algorithm 1 `Heavy-Tail Tsallis-INF (HTINF)`

Input: Number of arms K, heavy-tail parameters α and σ
Output: Sequence of actions $i_1, i_2, \cdots, i_T \in [K]$

for $t = 1, 2, \cdots$ **do**
  Calculate policy with learning rate $\eta_t^{-1} = \sigma t^{1/\alpha}$; Pick the regularizer $\Psi(x) = -\alpha \sum_{i=1}^{K} x_i^{1/\alpha}$:

$$x_t \leftarrow \operatorname*{argmin}_{x \in \triangle_{[K]}} \left(\eta_t \sum_{s=1}^{t-1} \langle \hat{\ell}_s, x \rangle + \Psi(x) \right)$$

  Sample new action $i_t \sim x_t$.
  Calculate the skipping threshold $r_t \leftarrow \Theta_\alpha \eta_t^{-1} x_{t,i_t}^{1/\alpha}$ where $\Theta_\alpha = \min\{1 - 2^{-\frac{\alpha-1}{2\alpha-1}}, (2 - \frac{2}{\alpha})^{\frac{1}{2-\alpha}}\}$.
  Play according to i_t and observe loss feedback ℓ_{t,i_t}.
  if $|\ell_{t,i_t}| > r_t$ **then**
    $\hat{\ell}_t \leftarrow \mathbf{0}$.
  else
    Construct weighted importance sampling loss estimator $\hat{\ell}_{t,i} \leftarrow \frac{\ell_{t,i}}{x_{t,i}} \mathbb{1}[i = i_t], \forall i \in [K]$.
  end if
end for

Theorem 4.1 (Performance of `HTINF`)**.** *If Assumptions 3.1 and 3.6 hold, we have the following* ***best-of-both-worlds*** *style regret guarantees.*

1. *When the environment is adversarial, Algorithm 1 ensures regret bound*

$$\mathcal{R}_T \leq \mathcal{O}\left(\sigma K^{1-1/\alpha} T^{1/\alpha}\right).$$

2. *If the environment is stochastically constrained adversarial with a unique optimal arm i^*, i.e., Assumption 3.5 holds, then Algorithm 1 ensures*

$$\mathcal{R}_T \leq \mathcal{O}\left(\sigma^{\frac{\alpha}{\alpha-1}} \sum_{i \neq i^*} \Delta_i^{-\frac{1}{\alpha-1}} \log T\right)^{2}.$$

The $\mathcal{O}(\log T)$ instance-dependent bound in Theorem 4.1 is due to a property similar to the self-bounding property of $1/2$-Tsallis entropy (Zimmert & Seldin, 2019). For $\alpha < 2$, such properties of $1/\alpha$-Tsallis entropy do not automatically hold, they are made possible by our novel skipping mechanism with action-dependent threshold.

[2] In this big-O notation, we hide an $\exp(\mathcal{O}(\frac{1}{\alpha-1}))$ factor. Such factors also appear in prior upper-bounds and lower-bounds on heavy-tailed MAB, see e.g. (Bubeck et al., 2013) Theorem 1 and Theorem 3.

4.2. Extending to Unknown α, σ Case: `OptTINF`

The two hyper-parameters σ, α in Algorithm 3 are just set to the true heavy-tail parameters of the loss distributions when they are known before-hand. When the *distributions' heavy-tail parameters* α, σ are both unknown to the agent, we can prove that by directly running `HTINF` with algorithm *hyper-parameters* $\alpha = 2$ and $\sigma = 1$ (not necessarily equal to the true α, σ values) "optimistically" as in Algorithm 2, one can still achieve $\mathcal{O}(\log T)$ regret in stochastic case and sub-linear regret in adversarial case.

Algorithm 2 Optimistic `HTINF` (`OptTINF`)

Input: Number of arms K
Output: Sequence of actions $i_1, i_2, \cdots, i_T \in [K]$
1: Run `HTINF` (Algorithm 1) with hyper-parameters $\alpha = 2$ and σ=1.

The performance of Algorithm 2 is described below. As the analysis is quite similar to that of Algorithm 1, we postpone the formal proof to Appendix B.

Theorem 4.2 (Performance of `OptTINF`)**.** *If Assumptions 3.1 and 3.6 hold, the following two statements are valid.*

1. *In adversarial cases, Algorithm 2 achieves*

$$\mathcal{R}_T \leq \mathcal{O}\left(\sigma^\alpha K^{\frac{\alpha-1}{2}} T^{\frac{3-\alpha}{2}} + \sqrt{KT}\right).$$

2. *In stochastically constrained adversarial environments with a unique optimal arm i^* (Assumption 3.5), it ensures*

$$\mathcal{R}_T \leq \mathcal{O}\left(\sigma^{\frac{2\alpha}{\alpha-1}} \sum_{i \neq i^*} \Delta_i^{-\frac{3-\alpha}{\alpha-1}} \log T\right).$$

For both cases, σ and α in the regret bounds refer to the true heavy-tail parameters of the loss distributions.

Theorem 4.2 claims that when facing an instance with unknown $1 < \alpha < 2$, Algorithm 2 still guarantees $O(T^{\frac{3-\alpha}{2}})$ "no-regret" performance and $\mathcal{O}(\log T)$ instance-dependent regret upper-bound for stochastic instances.

5. Regret Analysis of `HTINF`

In this section, we sketch the analysis of Algorithm 1. By definition, we need to bound

$$\mathcal{R}_T(y) \triangleq \sum_{t=1}^{T} \mathbb{E}\left[\langle x_t - y, \mu_t \rangle\right] \quad (y \in \triangle_{[K]}) \tag{2}$$

for the *one-hot* vector $y \triangleq \mathbf{e}_{i^*}$. For any $t \in [T], i \in [K]$, let $\mu'_{t,i} \triangleq \mathbb{E}[\ell_{t,i} \mathbb{1}[|\ell_{t,i}| \leq r_t] \mid \mathcal{F}_{t-1}, i_t = i]$. For a given y, we decompose $\mathcal{R}_T(y)$ into two parts:

$$\begin{aligned}\mathcal{R}_T(y) &= \mathbb{E}\left[\sum_{t=1}^{T} \langle x_t - y, \mu_t - \mu'_t \rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} \langle x_t - y, \mu'_t \rangle\right] \\ &= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \langle x_t - y, \mu_t - \mu'_t \rangle\right]}_{\text{skipping gap}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \langle x_t - y, \hat{\ell}_t \rangle\right]}_{\text{FTRL error}}\end{aligned} \tag{3}$$

where the last step is due to $\mathbb{E}[\hat{\ell}_t \mid \mathcal{F}_{t-1}] = \mu'_t$. We call the first part the *skipping gap*, and the second, the *FTRL error*.

In the following sections, we will show that both parts can be controlled and transformed into expressions similar to the bounds with self-bounding properties in (Zimmert & Seldin, 2019), guaranteeing best-of-both-worlds style regret upper-bounds. Therefore, the design of `HTINF` and our new analysis generalizes the self-bounding property of (Zimmert & Seldin, 2019) from 1/2-Tsallis entropy regularizer to general α-Tsallis entropy regularizers where $1/2 \leq \alpha < 1$.

5.1. To Control the Skipping Gap

To control the skipping gap part, notice that for all $t \in [T], i \in [K]$, we can bound

$$\begin{aligned}\mu_{t,i} - \mu'_{t,i} &= \mathbb{E}\left[|\ell_{t,i}| \mathbb{1}[|\ell_{t,i}| > r_t] \mid \mathcal{F}_{t-1}, i_t = i\right] \\ &\leq \mathbb{E}\left[|\ell_{t,i}|^\alpha r_t^{1-\alpha} \mid \mathcal{F}_{t-1}, i_t = i\right] \\ &\leq \sigma^\alpha r_t^{1-\alpha} = \Theta_\alpha^{1-\alpha} \sigma t^{1/\alpha - 1} x_{t,i}^{1/\alpha - 1}\end{aligned}$$

where Θ_α is a factor in r_t and only dependent on α, as defined in Line 4 of Algorithm 1. Moreover, by Assumption 3.6, $\mu_{t,i^*} - \mu'_{t,i^*} \geq 0$ a.s. Summing over i and t gives

$$\begin{aligned}\sum_{t=1}^{T} \langle x_t - \mathbf{e}_{i^*}, \mu_t - \mu'_t \rangle &\leq \Theta_\alpha^{1-\alpha} \sigma \sum_{t=1}^{T} \sum_{i \neq i^*} t^{1/\alpha - 1} x_{t,i}^{1/\alpha} \\ &\leq 5\sigma \sum_{t=1}^{T} \sum_{i \neq i^*} t^{1/\alpha - 1} x_{t,i}^{1/\alpha} \qquad (4) \\ &\leq 10\sigma (T+1)^{1/\alpha} K^{1-1/\alpha}. \qquad (5)\end{aligned}$$

5.2. To Control the FTRL Error

For the FTRL error part, we follow the regular analysis for FTRL algorithms. Note that our skipping mechanism is equivalent to plugging in $\hat{\ell}_t = \mathbf{0}$ for all skipped time step t in a FTRL framework for MAB algorithms. Therefore, due to the definition that $\mathbb{E}[\hat{\ell}_t] = \mu'_t$, we can leverage most standard techniques on regret analysis of a FTRL algorithm and obtain following lemma.

Lemma 5.1 (FTRL Regret Decomposition)**.**

$$\sum_{t=1}^{T}\langle x_t - y, \hat{\ell}_t\rangle \le \underbrace{\sum_{t=1}^{T}(\eta_t^{-1} - \eta_{t-1}^{-1})\left(\Psi(y) - \Psi(x_t)\right)}_{\textit{Part (A)}} + \underbrace{\sum_{t=1}^{T}\eta_t^{-1} D_\Psi(x_t, z_t)}_{\textit{Part (B)}}$$

where

$$z_t \triangleq \nabla\Psi^*\left(\nabla\Psi(x_t) - \eta_t \mathbb{1}[|\ell_{t,i_t}| \le r_t](\hat{\ell}_t - \ell_{t,i_t}\mathbf{1})\right).$$

In Lemma 5.1, z_t is an intermediate action probability-like measure vector (which does not necessarily sum up to 1) during the FTRL algorithm. Here we leverage a trick of *drifting* the loss vectors (Wei & Luo, 2018) $\hat{\ell}'_t \triangleq \hat{\ell}_t - \ell_{t,i_t}\mathbf{1}$. Intuitively, one can see that feeding $\hat{\ell}'_t$ into a FTRL framework will produce exactly the same action sequence as $\hat{\ell}_t$.

We then divide this upper-bound in Lemma 5.1 into two parts, parts (A) and (B), and analyze them separately.

5.2.1. Bound for Part (A)

As y is an one-hot vector, we have $\Psi(y) = -\alpha$ for $\Psi(x) = -\alpha\sum_{i=1}^{K} x_i^{1/\alpha}$. Hence, each summand in part (A) becomes

$$\begin{aligned}&(\eta_t^{-1} - \eta_{t-1}^{-1})\left(-\alpha + \alpha\sum_{i=1}^{K} x_i^{1/\alpha}\right)\\ &\le 2\sigma\frac{1}{\alpha}t^{1/\alpha-1}\cdot\alpha\sum_{i\neq i^*} x_{t,i}^{1/\alpha}\end{aligned}$$

due to the concavity of $t^{1/\alpha}$ (Lemma E.6) and the fact that $x_{t,i} \le 1$. This further implies

$$\text{(A)} \le \sum_{t=1}^{T} 2\sigma t^{1/\alpha-1}\sum_{i\neq i^*} x_{t,i}^{1/\alpha} \tag{6}$$

$$\le 4\sigma(T+1)^{1/\alpha}K^{1-1/\alpha}. \tag{7}$$

5.2.2. Bound for Part (B)

We can bound the expectation of each summand in part (B) as the following lemma states.

Lemma 5.2. *Algorithm 1 ensures*

$$\mathbb{E}[\eta_t^{-1}D_\Psi(x_t, z_t) \mid \mathcal{F}_{t-1}] \le 8\sigma t^{1/\alpha-1}\sum_{i\neq i^*} x_{t,i}^{1/\alpha} \tag{8}$$

$$\le 8\sigma t^{1/\alpha-1}K^{1-1/\alpha}. \tag{9}$$

5.3. Combining All Parts

In order to derive the claimed regret upper-bounds in Theorem 4.1, it suffices to plug in the bounds for the terms in Eq. (3) and Lemma 5.1.

Adversarial Case (Statement 1 in Theorem 4.1): To obtain an instance-independent bound for the expected total pseudo-regret $\mathcal{R}_T$, we can plug inequalities (5), (7) and (9) into Eq. (3) to obtain

$$\mathcal{R}_T \le 30\sigma K^{1-1/\alpha}(T+1)^{1/\alpha}.$$

Stochastically Constrained Adversarial Case (Statement 2 in Theorem 4.1): To obtain an instance-dependent bound for $\mathcal{R}_T$, we leverage the arm-pulling probability $\{x_t\}$ dependent bounds (6) and (8) for the FTRL part of $\mathcal{R}_T$. After plugging them together with (4) into (3), we see that

$$\mathcal{R}_T \le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*}\underbrace{15\sigma\left(\frac{1}{t}\right)^{1-1/\alpha}x_{t,i}^{1/\alpha}}_{\triangleq s_{t,i}}\right]. \tag{10}$$

We further apply the inequality of arithmetic and geometric means (AM-GM inequality) to $s_{t,i}$, as

$$\begin{aligned}s_{t,i} &= \left(\frac{\alpha\Delta_i}{2}x_{t,i}\right)^{\frac{1}{\alpha}}\left[\left(\frac{\alpha\Delta_i}{2}\right)^{-\frac{1}{\alpha-1}}\left(\frac{30\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}}\frac{1}{t}\right]^{\frac{\alpha-1}{\alpha}}\\ &\le \frac{\Delta_i}{2}x_{t,i} + \frac{\alpha-1}{\alpha}\left(\frac{\alpha}{2}\right)^{-\frac{1}{\alpha-1}}\left(\frac{30\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}}\Delta_i^{-\frac{1}{\alpha-1}}\frac{1}{t}.\end{aligned}$$

By noticing the fact that $\sum_{t\in[T]}\sum_{i\neq i^*}\Delta_i\,\mathbb{E}[x_{t,i}] \le \mathcal{R}_T$ (Lemma E.7), Eq. (10) solves to

$$\begin{aligned}\mathcal{R}_T &\le \frac{2\alpha-2}{\alpha}\left(\frac{\alpha}{2}\right)^{-\frac{1}{\alpha-1}}\left(\frac{30\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}}\\ &\quad\cdot\sum_{i\neq i^*}\Delta_i^{-\frac{1}{\alpha-1}}\ln(T+1)\\ &= \exp\left(\mathcal{O}\left(\frac{1}{\alpha-1}\right)\right)\sigma^{\frac{\alpha}{\alpha-1}}\sum_{i\neq i^*}\Delta_i^{-\frac{1}{\alpha-1}}\ln(T+1).\end{aligned}$$

6. Adaptive Algorithm: `AdaTINF`

In this section, our main goal is to achieve minimax optimal regret bounds for adversarial settings, without any knowledge about α, σ. Instead of estimating α and σ explicitly, which can be challenging, our key idea is to leverage a trade-off relationship between Part (A) and Part (B) in the FTRL error part (defined in Lemma 5.1), to balance the two parts dynamically.

To achieve a balance, we use a *doubling trick* to tune the learning rates and skipping thresholds, which has been adopted in the literature to design adaptive algorithms (see, e.g., Wei & Luo (2018)). The formal procedure of AdaTINF is given in Algorithm 3, with the crucial differences between Algorithm 1 highlighted in blue texts.

It can be seen as HTINF equipped with a multiplier to both learning rates and skipping thresholds, maintained at running time, as

$$\eta_t^{-1} = \lambda_t \sqrt{t}, \quad r_t = \lambda_t \Theta_2 \sqrt{t} \sqrt{x_{t,i_t}}, \quad \forall 1 \le t \le T,$$

where λ_t is the doubling magnitude for the t-th time slot.

Algorithm 3 Adaptive Tsallis-INF (AdaTINF)

Input: Number of arms K, time horizon T
Output: Sequence of actions $i_1, i_2, \cdots, i_T \in [K]$

1: Initialize $J \leftarrow 0$, $S_0 \leftarrow 0$
2: **for** $t = 1, 2, \cdots$ **do**
3: $\quad \lambda_t \leftarrow 2^J$
4: $\quad$ Calculate policy with learning rate $\eta_t^{-1} = \lambda_t \sqrt{t}$ and regularizar $\Psi(x) = -2\sum_{i=1}^K x_i^{1/2}$:

$$x_t \leftarrow \operatorname*{argmin}_{x \in \triangle_{[K]}} \left(\eta_t \sum_{s=1}^{t-1} \langle \hat{\ell}_s, x \rangle + \Psi(x) \right)$$

5: $\quad$ Decide action $i_t \sim x_t$, calculate $r_t \leftarrow \lambda_t (1 - 2^{-1/3}) \sqrt{t} \sqrt{x_{t,i_t}}$.
6: $\quad$ Play according to i_t and observe loss feedback ℓ_{t,i_t}.
7: $\quad$ **if** $|\ell_{t,i_t}| > r_t$ **then**
8: $\quad\quad \hat{\ell}_t \leftarrow \mathbf{0}$
9: $\quad\quad c_t \leftarrow \ell_{t,i_t}$
10: $\quad$ **else**
11: $\quad\quad$ Construct weighted importance sampling loss estimator $\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{x_{t,i}} \mathbb{1}[i = i_t], \forall i \in [K]$.
12: $\quad\quad c_t \leftarrow 2\eta_t x_{t,i_t}^{-1/2} \ell_{t,i_t}^2$
13: $\quad$ **end if**
14: $\quad S_J \leftarrow S_J + c_t$
15: $\quad$ **if** $2^J \sqrt{K(T+1)} < S_J$ **then**
16: $\quad\quad J \leftarrow \max\{J+1, \lceil \log_2(c_t / \sqrt{K(T+1)}) \rceil + 1\}$
17: $\quad\quad S_J \leftarrow c_t$
18: $\quad$ **end if**
19: **end for**

We briefly explain our design. Suppose, initially, all λ_t's are set to a same number $\lambda > 1$ instead of 1. Then, part (A) will become approximately λ times bigger than that under HTINF, while the expected value of part (B) will be scaled by a factor $\lambda^{1-\alpha} < 1$. In other words, increasing λ enlarges part (A) but makes part (B) smaller. Therefore, if we can estimate parts (A) and (B), we can keep them of roughly the same magnitude, by doubling λ whenever (A) becomes smaller than (B).

As Eq. (4) and (5) are similar to Eq. (8) and (9), the skipping gap can be treated similarly to (B). Therefore, we also take it into consideration in the doubling-balance mechanism. Due to the future-dependent Eq. (6) is hard to estimate, we use the looser Eq. (7) to represent part (A). This stops Algorithm 3 from enjoying an $\mathcal{O}(\log T)$-style gap-dependent regret. However, it can still guarantee a minimax optimal regret in general case, as described in Theorem 6.1.

Theorem 6.1 (Performance of AdaTINF)**.** *If Assumptions 3.1 and 3.6 hold, Algorithm 3 ensures a regret of*

$$\mathcal{R}_T \le \mathcal{O}(\sigma K^{1-1/\alpha} T^{1/\alpha} + \sqrt{KT}),$$

which is ***minimax optimal****.*

The proof is sketched in Section 7, while the formal version is deferred to Appendix C.

Remark. Though T is assumed to be known in Algorithm 3, the assumption can be removed via another doubling trick without effect to the order of the total regret. Check Appendix D for more details.

7. Analysis of AdaTINF

Since the crucial learning rate multiplier λ_t is maintained by an adaptive doubling trick, in the analysis, we will group time slots with equal λ_t's into *epochs*. For $j \ge 0$, $\mathcal{T}_j \triangleq \{t \in [T] \mid \lambda_t = 2^j\}$ are the indices of time slots belonging to epoch j. Further denote the first step in epoch j by $\gamma_j \triangleq \min\{t \in \mathcal{T}_j\}$ and the last one by $\tau_j \triangleq \max\{t \in \mathcal{T}_j\}$. Without loss of generality, assume no doubling happened at slot T, then the final value of J in Algorithm 3 is just the index of the last non-empty epoch.

We will first show

$$\mathcal{R}_T \le \mathcal{O}\left(\mathbb{E}[2^J] \sqrt{KT} \right).$$

As defined in the pseudo-code, let

$$c_t \triangleq 2\eta_t x_{t,i_t}^{-1/2} \ell_{t,i_t}^2 \mathbb{1}[|\ell_{t,i_t}| \le r_t] + \ell_{t,i_t} \mathbb{1}[|\ell_{t,i_t}| > r_t].$$

According to the condition to enter a new epoch (Line 15 in Algorithm 3), for all $0 \le j < J$, if $\mathcal{T}_j$ is non-empty, τ_j will cause $S_j > 2^j \sqrt{K(T+1)}$. Hence, we have the following conditions:

$$\mathbb{1}[\gamma_j > 1] c_{\gamma_j - 1} + \sum_{t \in \mathcal{T}_j \setminus \{\tau_j\}} c_t \le 2^j \sqrt{K(T+1)}, \quad (11)$$

$$\sum_{t \in \mathcal{T}_j} c_t > 2^{j-1} \sqrt{K(T+1)}. \quad (12)$$

For $j = J$, as no doubling has happened after that, we have

$$\mathbb{1}[\gamma_J > 1] c_{\gamma_J - 1} + \sum_{t \in \mathcal{T}_J} c_t \le 2^J \sqrt{K(T+1)}. \quad (13)$$

Similar to Eq. (3) used in Section 5, we begin with the following decomposition of $\mathcal{R}_T(y)$ for $y = \mathbf{e}_{i^*}$:

$$\mathcal{R}_T(y) = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\langle x_t - y, \mu_t - \mu'_t\rangle\right]}_{\mathcal{R}_T^s} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\langle x_t - y, \hat{\ell}_t\rangle\right]}_{\mathcal{R}_T^f} \tag{14}$$

where $\mu'_{t,i} \triangleq \mathbb{E}[\ell_{t,i}\mathbb{1}[|\ell_{t,i}| \le r_t] \mid \mathcal{F}_{t-1}, i_t = i]$. We still call $\mathcal{R}_T^s$ the skipping gap and $\mathcal{R}_T^f$ the FTRL error.

According to Lemma 5.1, we have

$$\begin{aligned}\mathcal{R}_T^f &\le \mathbb{E}\left[\underbrace{\eta_T \max_{x\in\Delta_{[K]}}\Psi(x)}_{\text{Part (A)}}\right] + \mathbb{E}\left[\underbrace{\sum_{t=1}^{T}\eta_t^{-1}D_\Psi(x_t, z_t)}_{\text{Part (B)}}\right] \\ &\le \mathbb{E}[2^J]\sqrt{K(T+1)} + \mathbb{E}\left[\sum_{t=1}^{T}\eta_t^{-1}D_\Psi(x_t, z_t)\right] \end{aligned} \tag{15}$$

Similar to Algorithm 1, we can show $D_\Psi(x_t, z_t) \le 2\eta_t x_{t,i_t}^{-1/2}\ell_{t,i_t}^2\mathbb{1}[|\ell_{t,i_t}| \le r_t]$ for all $t \in [T]$. Moreover, by Assumption 3.6, $\mathcal{R}_T^s \le \mathbb{E}[\sum_{t=1}^{T}\ell_{t,i_t}\mathbb{1}[|\ell_{t,i_t}| > r_t]]$. Therefore, with the help of Eq. (11) and (13), we have

$$\begin{aligned}&\mathcal{R}_T^s + \mathbb{E}[\text{Part (B)}] \\ &\le \mathbb{E}\left[\sum_{t=1}^{T}c_t\right] \\ &\le \mathbb{E}\left[2^{J+1}\sqrt{K(T+1)}\right]. \end{aligned} \tag{16}$$

Combining Eq. (14), (15) and (16) gives

$$\mathcal{R}_T \le \mathbb{E}[2^J]\cdot 3\sqrt{K(T+1)},$$

Therefore, it remains to bound $\mathbb{E}[2^J]$. When $J = 0$, there is nothing to do. Otherwise, consider the second to last non-empty epoch, $\mathcal{T}_{J'}$. The condition to enter a new epoch also guarantees that $2^J\sqrt{K(T+1)} \le 2^{J'+1}\sqrt{K(T+1)} + 4c_{\tau_{J'}}$. Applying Eq. (12) to $J' < J$, we obtain

$$\mathbb{1}[J \ge 1](2^{J'})^\alpha\sqrt{K(T+1)} \le (2^{J'})^{\alpha-1}2\sum_{t\in\mathcal{T}_{J'}}c_t, \tag{17}$$

After appropriate relaxing the RHS of Eq. (17) and taking expectation of both sides, it solves to the following upper-bound for $\mathbb{E}[\mathbb{1}[J \ge 1]2^{J'}]$:

Lemma 7.1. *Algorithm 3 guarantees that*

$$\mathbb{E}[\mathbb{1}[J \ge 1]2^{J'}] \le 28\sigma K^{1/2-1/\alpha}(T+1)^{1/\alpha-1/2}.$$

Moreover, we can obtain a bound for $\mathbb{E}[c_{\tau_{J'}}]$ stated as follows:

Lemma 7.2. *Algorithm 3 guarantees that*

$$\begin{aligned}&\mathbb{E}[\mathbb{1}[J \ge 1]c_{\tau_{J'}}] \\ &\le 0.1\,\mathbb{E}[\mathbb{1}[J \ge 1]2^{J'}\sqrt{T}] + 4\,\mathbb{E}\left[\max_{t\in[T]}|\ell_{t,i_t}|\right].\end{aligned}$$

Using the fact that $\mathbb{E}[\max_{t\in[T]}|\ell_{t,i_t}|] \le \sigma T^{1/\alpha}$ (Lemma E.3), we conclude that Algorithm 3 has the regret guarantee of

$$\begin{aligned}\mathcal{R}_T &\le 3\,\mathbb{E}[2^J\sqrt{K(T+1)}] \\ &\le 3\sqrt{K(T+1)} + 204\sigma K^{1-1/\alpha}(T+1)^{1/\alpha} + 12\sigma T^{1/\alpha}.\end{aligned}$$

8. Conclusion

We propose `HTINF`, a novel algorithm achieving the optimal instance-dependent regret bound for the stochastic heavy-tailed MAB problem, and the optimal instance-independent regret bound for a more general adversarial setting, without extra logarithmic factors. We also propose `AdaTINF`, which can achieve the same optimal instance-independent regret even when prior knowledge on heavy-tailed parameters α, σ are absent. Our work shows that the FTRL (or OMD) technique can be a powerful tool for designing heavy-tailed MAB algorithm, leading to novel theoretical results that have not been achieved by UCB algorithms.

It is an interesting future work to figure out whether it is possible to design a best-of-both-worlds algorithm without knowning the actual heavy-tail distribution parameters α and σ.

Acknowledgment

This work is supported by the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2020AAA0108400 and 2020AAA0108403.

References

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

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

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 322–331. IEEE, 1995.

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

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

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

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59 (11):7711–7717, 2013.

Bubeck, S., Li, Y., Luo, H., and Wei, C.-Y. Improved path-length regret bounds for bandits. In *Conference On Learning Theory*, pp. 508–528. PMLR, 2019.

De Rooij, S., Van Erven, T., Grünwald, P. D., and Koolen, W. M. Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research*, 15(1):1281–1316, 2014.

Erez, L. and Koren, T. Best-of-all-worlds bounds for online learning with feedback graphs. *arXiv preprint arXiv:2107.09572*, 2021.

Garivier, A. and Cappé, O. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pp. 359–376. JMLR Workshop and Conference Proceedings, 2011.

Hadiji, H. and Stoltz, G. Adaptation to the range in k-armed bandits. *arXiv preprint arXiv:2006.03378*, 2020.

Ito, S. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, pp. 2552–2583. PMLR, 2021.

Jin, T. and Luo, H. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in neural information processing systems*, 33: 16557–16566, 2020.

Jin, T., Huang, L., and Luo, H. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *Advances in Neural Information Processing Systems*, 34, 2021.

Kagrecha, A., Nair, J., and Jagannathan, K. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 11272–11281, 2019.

Lattimore, T. Optimally confident ucb: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*, 2015.

Lee, K., Yang, H., Lim, S., and Oh, S. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. *Advances in Neural Information Processing Systems*, 33:8452–8462, 2020.

Lu, S., Wang, G., Hu, Y., and Zhang, L. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, pp. 4154–4163. PMLR, 2019.

Luo, H. and Schapire, R. E. Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory*, pp. 1286–1304. PMLR, 2015.

Medina, A. M. and Yang, S. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pp. 1642–1650. PMLR, 2016.

Orabona, F. and Pal, D. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29:577–585, 2016.

Orabona, F. and Pál, D. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018.

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

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

Shao, H., Yu, X., King, I., and Lyu, M. R. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31, 2018.

Tsallis, C. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487, 1988.

Vakili, S., Liu, K., and Zhao, Q. Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *IEEE Journal of Selected Topics in Signal Processing*, 7(5):759–767, 2013.

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

Xue, B., Wang, G., Wang, Y., and Zhang, L. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 2936–2942, 2021.

Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 467–475. PMLR, 2019.

Supplementary Materials: Proofs and Discussions

A Formal Analysis of `HTINF` (Algorithm 1) **12**

A.1 Main Theorem . . . 12

A.2 Proof when Bounding $\mathcal{R}_T^s$ (*the skipped part*) . . . 15

A.3 Proof when Bounding $\mathcal{R}_T^f$ (*the FTRL part*) . . . 15

B Formal Analysis of `OptTINF` (Algorithm 2) **18**

B.1 Main Theorem . . . 18

B.2 Proof when Bounding $\mathcal{R}_T^s$ (*the skipped part*) . . . 20

B.3 Proof when Bounding $\mathcal{R}_T^f$ (*the FTRL part*) . . . 20

C Formal Analysis of `AdaTINF` (Algorithm 3) **21**

C.1 Main Theorem . . . 21

C.2 Proof when Reducing $\mathcal{R}_T$ to $\mathbb{E}[2^J]$ . . . 24

C.3 Proof when Bounding $\mathbb{E}[2^J]$ . . . 24

D Removing Dependency on Time Horizon T in Algorithm 3 **25**

E Auxiliary Lemmas **26**

E.1 Probability Lemmas . . . 26

E.2 Arithmetic Lemmas . . . 26

E.3 Lemmas on the FTRL Framework for MAB Algorithm Design . . . 27

A. Formal Analysis of `HTINF` (Algorithm 1)

A.1. Main Theorem

In this section, we present a formal proof of Theorem 4.1. For the sake of accuracy, we state the regret guarantees without using any big-Oh notations, as follows (which directly implies Theorem 4.1).

Theorem A.1 (Regret Guarantee of Algorithm 1)**.** *If Assumptions 3.1 and 3.6 hold, i.e., the environment is heavy-tailed with parameters α and σ, and there is an optimal arm whose all losses are truncated non-negative. Then Algorithm 1 guarantees:*

1. *The regret is no more than*

$$\mathcal{R}_T \leq 30\sigma K^{1-1/\alpha}(T+1)^{1/\alpha},$$

   no matter the environment is stochastic or adversarial.

2. *Furthermore, if the environment is stochastically constrained with a unique best arm i^*, i.e., Assumption 3.5 holds, then it, in addition, enjoys a regret bound of*

$$\mathcal{R}_T \leq \frac{2\alpha-2}{\alpha}\left(\frac{\alpha}{2}\right)^{-\frac{1}{\alpha-1}}\left(\frac{30\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}}\sum_{i\neq i^*}\Delta_i^{-\frac{1}{\alpha-1}}\ln(T+1).$$

Proof. Define $\mu'_{t,i} \triangleq \mathbb{E}[\ell_{t,i}\mathbb{1}[|\ell_{t,i}| \le r_t] \mid \mathcal{F}_{t-1}, i_t = i]$. For the given $y = \mathbf{e}_{i^*} \in \triangle_{[K]}$, consider the regret of the algorithm with respect to policy y, defined and decomposed as

$$\mathcal{R}_T(y) \triangleq \sum_{t=1}^{T} \mathbb{E}[\langle x_t - y, \mu_t\rangle] = \mathbb{E}\left[\sum_{t=1}^{T}\langle x_t - y, \mu_t - \mu'_t\rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T}\langle x_t - y, \mu'_t\rangle\right] \triangleq \mathcal{R}^s_T(y) + \mathcal{R}^f_T(y),$$

which we called the *skipped part* and *FTRL part*. For simplicity, we abbreviate the parameter (y) for $\mathcal{R}^s_T$ and $\mathcal{R}^f_T$.

As defined in Algorithm 1, $\hat{\ell}_t$ is set to 0 when $|\ell_{t,i_t}| > r_t$. Hence, by the property of weighted importance sampling estimator (Lemma E.8; note that it is applied to the truncated loss with mean $\mu'_{t,i}$), $\mathbb{E}[\hat{\ell}_{t,i} \mid \mathcal{F}_{t-1}] = \mu'_{t,i}$

$$\mathcal{R}^f_T = \mathbb{E}\left[\sum_{t=1}^{T}\langle x_t - y, \hat{\ell}_t\rangle\right].$$

For the first term, $\mathcal{R}^s_T$, we can bound it using the following two lemmas, whose proof are propounded to next subsection.

Lemma A.2. *For any $1 \le t \le T$ and $i \in [K]$, we have*

$$\mu_{t,i} - \mu'_{t,i} \le \Theta_\alpha^{1-\alpha}\sigma t^{1/\alpha-1}x_{t,i}^{1/\alpha-1},$$

where Θ_α is a constant used in Algorithm 1 that only depends on α.

Lemma A.3. *If i^* is an optimal arm whose loss feedback are all truncated non-negative, then for $y = \mathbf{e}_{i^*}$, we have*

$$\mathcal{R}^s_T(y) \le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*} x_{t,i}(\mu_{t,i} - \mu'_{t,i})\right].$$

Therefore, for $y = \mathbf{e}_{i^*}$ we have

$$\begin{aligned}
\mathcal{R}^s_T &\le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*}\Theta_\alpha^{1-\alpha}\sigma t^{1/\alpha-1}x_{t,i}^{1/\alpha}\right] \\
&\overset{(a)}{\le} \mathbb{E}\left[5\sum_{t=1}^{T}\sum_{i\neq i^*}\sigma t^{1/\alpha-1}x_{t,i}^{1/\alpha}\right] \qquad (18)\\
&\overset{(b)}{\le} 5\alpha\sigma K^{1-1/\alpha}(T+1)^{1/\alpha}, \qquad (19)
\end{aligned}$$

where step (a) is due to $\Theta_\alpha^{1-\alpha} \le \Theta_2^{-1} \le 5$ and (b) applies Lemma E.4 and Lemma E.5.

Now consider the second term, $\mathcal{R}^f_T$. Consider the vector $\hat{\ell}'_t \triangleq \mathbb{1}[|\ell_{t,i_t}| \le r_t](\hat{\ell}_t - \ell_{t,i_t}\mathbf{1})$. Note that $\langle\hat{\ell}'_t, x\rangle = \langle\hat{\ell}_t, x\rangle - \mathbb{1}[|\ell_{t,i_t}| \le r_t]\ell_{t,i_t}$ for any vector $x \in \triangle_{[K]}$, so a FTRL algorithm fed with loss vector $\hat{\ell}'_t$ with produce exactly the same action sequence as another instance fed with $\hat{\ell}_t$ (as constant terms will never affect the choice of the argmax operator over the simplex). Therefore, we can apply Lemma E.9 with loss vectors as $\hat{\ell}'_t$, yielding

$$\sum_{t=1}^{T}\langle x_t - y, \hat{\ell}_t\rangle = \sum_{t=1}^{T}\langle x_t - y, \hat{\ell}'_t\rangle \le \underbrace{\sum_{t=1}^{T}\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right)\left(\Psi(y) - \Psi(x_t)\right)}_{\text{Part (A)}} + \underbrace{\sum_{t=1}^{T}\eta_t^{-1}D_\Psi(x_t, z_t)}_{\text{Part (B)}} \qquad (20)$$

where $z_t \triangleq \nabla\Psi^*(\nabla\Psi(x_t) - \eta_t\hat{\ell}'_t) = \nabla\Psi^*\left(\nabla\Psi(x_t) - \eta_t\mathbb{1}[|\ell_{t,i_t}| \le r_t](\hat{\ell}_t - \ell_{t,i_t}\mathbf{1})\right)$.

Now consider the first term $\sum_{t=1}^{T}(\eta_t^{-1} - \eta_{t-1}^{-1})(\Psi(y) - \Psi(x_t))$, which is denoted by (A) for simplicity. We have

Lemma A.4. *For part (A), Algorithm 1 ensures the following inequality for any one-hot vector $y \in \triangle_{[K]}$:*

$$\mathbb{E}[(A)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[(\eta_t^{-1} - \eta_{t-1}^{-1})(\Psi(y) - \Psi(x_t)) \mid \mathcal{F}_{t-1}\right]\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} 2\sigma t^{1/\alpha - 1} \sum_{i \neq i^*} x_{t,i}^{1/\alpha}\right], \tag{21}$$

which further implies

$$\mathbb{E}[(A)] \leq \sum_{t=1}^{T} 2\sigma t^{1/\alpha - 1} K^{1 - 1/\alpha}. \tag{22}$$

For the second term, denoted by (B), we have

Lemma A.5 (Restatement of Lemma 5.2)**.** *For Part (B), Algorithm 1 ensures*

$$\mathbb{E}[(B)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[\eta_t^{-1} D_\Psi(x_t, z_t) \mid \mathcal{F}_{t-1}]\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} 8\sigma t^{1/\alpha - 1} \sum_{i \neq i^*} x_{t,i}^{1/\alpha}\right], \tag{23}$$

which further implies

$$\mathbb{E}[(B)] \leq \sum_{t=1}^{T} 8\sigma t^{1/\alpha - 1} K^{1 - 1/\alpha}. \tag{24}$$

Hence, for general cases, due to Equations (22) and (24) we have

$$\mathcal{R}_T^f = \mathbb{E}[(\mathrm{A})] + \mathbb{E}[(\mathrm{B})] \leq \sum_{t=1}^{T} 10\sigma K^{1 - 1/\alpha} \sum_{t=1}^{T} t^{1/\alpha - 1} \leq 10\alpha\sigma K^{1 - 1/\alpha} (T+1)^{1/\alpha},$$

where the last inequality comes from Lemma E.5. Therefore, taking (18) into consideration, we have:

$$\mathcal{R}_T = \mathcal{R}_T^s + \mathcal{R}_T^f \leq 15\alpha\sigma K^{1 - 1/\alpha} (T+1)^{1/\alpha} \leq 30\sigma K^{1 - 1/\alpha} (T+1)^{1/\alpha}.$$

Now, for stochastically constrained adversarial case with unique best arm i^* throughout the process, due to Equations (18) (21) and (23), we have

$$\mathcal{R}_T = \mathcal{R}_T^s + \mathbb{E}[(\mathrm{A})] + \mathbb{E}[(\mathrm{B})] \leq \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} \underbrace{15\sigma t^{1/\alpha - 1} x_{t,i}^{1/\alpha}}_{\triangleq s_{t,i}}\right].$$

We can then write

$$\begin{aligned} s_{t,i} &= \left(\frac{\alpha}{2} \Delta_i x_{t,i}\right)^{1/\alpha} \left[\left(\frac{\alpha}{2}\right)^{-\frac{1}{\alpha-1}} \left(\frac{30\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} \Delta_i^{-\frac{1}{\alpha-1}} \frac{1}{t}\right]^{\frac{\alpha-1}{\alpha}} \\ &\leq \frac{\Delta_i}{2} x_{t,i} + \frac{\alpha - 1}{\alpha} \left(\frac{\alpha}{2}\right)^{-\frac{1}{\alpha-1}} \left(\frac{30\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} \Delta_i^{-\frac{1}{\alpha-1}} \frac{1}{t} \end{aligned}$$

where the last step uses the inequality of arithmetic and geometric means $a^{1/\alpha} b^{1-1/\alpha} \leq \frac{1}{\alpha} a + \left(1 - \frac{1}{\alpha}\right) b$. Therefore

$$R_T \leq \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} \frac{\Delta_i}{2} x_{t,i}\right] + \sum_{t=1}^{T} \sum_{i \neq i^*} \frac{\alpha - 1}{\alpha} \left(\frac{\alpha}{2}\right)^{-\frac{1}{\alpha-1}} \left(\frac{30\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} \Delta_i^{-\frac{1}{\alpha-1}} \frac{1}{t}$$

$$\leq \frac{1}{2}R_T + \sum_{i\neq i^*} \frac{\alpha-1}{\alpha}\left(\frac{\alpha}{2}\right)^{-\frac{1}{\alpha-1}}\left(\frac{30\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} \Delta_i^{-\frac{1}{\alpha-1}} \ln(T+1) \tag{25}$$

where the last step uses Lemma E.7. Equation (25) then solves to

$$\mathcal{R}_T \leq \sum_{i\neq i^*} \frac{2\alpha-2}{\alpha}\left(\frac{\alpha}{2}\right)^{-\frac{1}{\alpha-1}}\left(\frac{30\sigma}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} \Delta_i^{-\frac{1}{\alpha-1}} \ln(T+1),$$

as claimed. □

Proof of Theorem 4.1. It is a direct consequence of the theorem above. □

A.2. Proof when Bounding $\mathcal{R}_T^s$ (*the skipped part*)

Proof of Lemma A.2. Starting from the definition of $\mu'_{t,i}$ and $\mu_{t,i}$, we can write

$$\begin{aligned}
\mu_{t,i} - \mu'_{t,i} &= \mathbb{E}\left[\ell_{t,i_t} \mid \mathcal{F}_{t-1}, i_t = i\right] - \mathbb{E}\left[\ell_{t,i_t} \cdot \mathbb{1}[|\ell_{t,i_t}| \leq r_t] \mid \mathcal{F}_{t-1}, i_t = i\right] \\
&= \mathbb{E}\left[\ell_{t,i_t} \cdot \mathbb{1}[|\ell_{t,i_t}| > r_t] \mid \mathcal{F}_{t-1}, i_t = i\right] \\
&\leq \mathbb{E}\left[|\ell_{t,i_t}| \cdot \mathbb{1}[|\ell_{t,i_t}| > r_t] \mid \mathcal{F}_{t-1}, i_t = i\right] \\
&\leq \mathbb{E}\left[|\ell_{t,i_t}|^\alpha r_t^{1-\alpha} \cdot \mathbb{1}[|\ell_{t,i_t}| > r_t] \mid \mathcal{F}_{t-1}, i_t = i\right] \\
&\leq \mathbb{E}\left[|\ell_{t,i_t}|^\alpha r_t^{1-\alpha} \mid \mathcal{F}_{t-1}, i_t = i\right] \\
&\overset{\text{(a)}}{=} \mathbb{E}\left[|\ell_{t,i}|^\alpha \Theta_\alpha^{1-\alpha}\sigma^{1-\alpha} t^{\frac{1-\alpha}{\alpha}} x_{t,i}^{\frac{1-\alpha}{\alpha}} \mid \mathcal{F}_{t-1}\right] \\
&\leq \sigma\Theta_\alpha^{1-\alpha} t^{\frac{1-\alpha}{\alpha}} x_{t,i}^{\frac{1-\alpha}{\alpha}}
\end{aligned}$$

where in step (a) we plug in $r_t = \Theta_\alpha \eta_t^{-1} x_{t,i_t}^{1/\alpha}$. □

Proof of Lemma A.3. Recall that $\mu_{t,i} - \mu'_{t,i} = \mathbb{E}\left[\ell_{t,i} \cdot \mathbb{1}[|\ell_{t,i}| > r_t] \mid \mathcal{F}_{t-1}, i_t = i\right]$, hence according to our assumption that ℓ_{t,i^*} is truncated non-negative (Assumption 3.6), we have $\mu_{t,i^*} - \mu'_{t,i^*} \geq 0$ a.s., thus when $y = \mathbf{e}_{i^*}$,

$$(x_{t,i^*} - y) \cdot (\mu_{t,i} - \mu_{t,i^*}) = (x_{t,i^*} - 1) \cdot (\mu_{t,i} - \mu_{t,i^*}) \leq 0.$$

Therefore

$$\begin{aligned}
\mathcal{R}_T^s(y) &= \mathbb{E}\left[\sum_{t=1}^T \langle x_t - y, \mu_t - \mu'_t\rangle\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^T \sum_{i\neq i^*} (x_{t,i} - y_i) \cdot \left(\mu_{t,i} - \mu'_{t,i}\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \sum_{i\neq i^*} x_{t,i}\left(\mu_{t,i} - \mu'_{t,i}\right)\right],
\end{aligned}$$

as claimed. □

A.3. Proof when Bounding $\mathcal{R}_T^f$ (*the FTRL part*)

For our purpose, we need a technical lemma stating that the components of z_t are at most a constant times larger than x_t's components.

Lemma A.6. *For any $t \in [T]$ and $i \in [K]$, Algorithm 1 guarantees that*

$$z_{t,i} \leq 2^{\frac{\alpha}{2\alpha-1}} x_{t,i}$$

where $z_t \triangleq \nabla\Psi^*(\nabla\Psi(x_t) - \eta_t \mathbb{1}[|\ell_{t,i_t}| \leq r_t](\hat{\ell}_t - \ell_{t,i_t}\mathbf{1}))$.

Proof. If $|\ell_{t,i_t}| > r_t$, then $x_t = z_t$. Otherwise, we denote $\nabla\Psi(x_t)$ by x_t^*, and denote $\nabla\Psi(z_t)$ by z_t^*, then we have $-x_{t,i}^* = x^{-\frac{\alpha-1}{\alpha}}$ and

$$\begin{aligned} z_{t,i}^* &= x_{t,i}^* - \eta_t \hat{\ell}_{t,i} + \eta_t \ell_{t,i_t} \\ &= \begin{cases} x_{t,i}^* - \eta_t \frac{\ell_{t,i}}{x_{t,i}} + \eta_t \ell_{t,i} & i = i_t \\ x_{t,i^*} + \eta_t \ell_{t,i_t} & i \neq i_t. \end{cases} \end{aligned}$$

If $i = i_t$, we have

$$-z_{t,i}^* \geq -x_{t,i}^* - \eta_t \frac{|\ell_{t,i}|}{x_{t,i}} = x_{t,i}^{-\frac{\alpha-1}{\alpha}} - \eta_t \frac{|\ell_{t,i}|}{x_{t,i}} \geq x_{t,i}^{-\frac{\alpha-1}{\alpha}} - \Theta_\alpha x_{t,i}^{\frac{1-\alpha}{\alpha}},$$

where the last step is due to $|\ell_{t,i_t}| \leq r_t$ and our choice of r_t in Algorithm 1. Thus

$$z_{t,i} = (-z_{t,i}^*)^{-\frac{\alpha}{\alpha-1}} \leq x_{t,i}(1-\Theta_\alpha)^{-\frac{\alpha}{\alpha-1}} \leq 2^{\frac{\alpha}{2\alpha-1}} x_{t,i}$$

where the last step is because $\Theta_\alpha \leq 1 - 2^{-\frac{\alpha-1}{2\alpha-1}}$.

If $i \neq i_t$, we have $-z_{t,i}^* \geq -x_{t,i}^* - \Theta_\alpha x_{t,i_t}^{1/\alpha} \geq x_{t,i}^{-\frac{\alpha-1}{\alpha}} - \Theta_\alpha$, thus

$$z_{t,i} = (-z_{t,i}^*)^{-\frac{\alpha}{\alpha-1}} \leq x_{t,i}(1 - \Theta_\alpha x_{t,i}^{\frac{\alpha-1}{\alpha}})^{-\frac{\alpha}{\alpha-1}} \leq x_{t,i}(1-\Theta_\alpha)^{-\frac{\alpha}{\alpha-1}} \leq 2^{\frac{\alpha}{2\alpha-1}} x_{t,i}.$$

Combining two cases together gives our conclusion. □

Proof of Lemma A.4. By definition, for any $t \in [T]$, one-hot $y \in \triangle_{[K]}$ and $x_t \in \triangle_{[K]}$, we have

$$\eta_t^{-1} - \eta_{t-1}^{-1} = \sigma\left(t^{1/\alpha} - (t-1)^{1/\alpha}\right) \overset{\text{(a)}}{\leq} \sigma \frac{1}{\alpha}(t-1)^{1/\alpha - 1} \overset{\text{(b)}}{\leq} 2\sigma \frac{1}{\alpha} t^{1/\alpha - 1},$$

where (a) comes from Lemma E.6 and (b) comes from the fact that $t \geq 1$ and $\frac{1}{\alpha} - 1 \geq -\frac{1}{2}$. Moreover, by definition of $\Psi(x) = -\alpha \sum_{i=1}^K x_i^{1/\alpha}$, we have

$$\Psi(y) - \Psi(x) = \alpha \sum_{i=1}^K x_i^{1/\alpha} - \alpha \sum_{i=1}^K y_i^{1/\alpha} = \alpha \sum_{i=1}^K x_i^{1/\alpha} - \alpha \leq \alpha \sum_{i \neq i^*} x_{t,i}^{1/\alpha}$$

from the assumption that y is an one-hot vector. Therefore, we have

$$\mathbb{E}[(\mathrm{A})] = \sum_{t=1}^T [(\eta_t^{-1} - \eta_{t-1}^{-1})(\Psi(y) - \Psi(x_t)) \mid \mathcal{F}_{t-1}] \leq \mathbb{E}\left[\sum_{t=1}^T \sum_{i \neq i^*} 2\sigma t^{1/\alpha - 1} x_{t,i}^{1/\alpha}\right],$$

which further implies (by Lemma E.4)

$$\mathbb{E}[(\mathrm{A})] = \sum_{t=1}^T [(\eta_t^{-1} - \eta_{t-1}^{-1})(\Psi(y) - \Psi(x_t)) \mid \mathcal{F}_{t-1}] \leq \sum_{t=1}^T 2\sigma t^{1/\alpha - 1} K^{1 - 1/\alpha}.$$

□

Proof of Lemma A.5 (and also Lemma 5.2). Consider a summand before taking expectation, i.e., $\eta_t^{-1} D_\Psi(x_t, z_t)$. Let $f(x) = -\alpha x^{1/\alpha}$, we then have

$$\eta_t^{-1} D_\Psi(x_t, z_t) \overset{(a)}{=} \eta_t^{-1} D_{\Psi^*}(\nabla\Psi(z_t), \nabla\Psi(x_t))$$

$$
\begin{aligned}
&= \Psi^*(\nabla\Psi(z_t)) - \Psi^*(\nabla\Psi(x_t)) - \langle x_t, \nabla\Psi(z_t) - \nabla\Psi(x_t)\rangle \\
&\overset{(b)}{\le} \eta_t^{-1}\sum_{i=1}^{K}\frac{1}{2}\max\{f''(x_{t,i})^{-1}, f''(z_{t,i})^{-1}\}\cdot\eta_t^2(\hat{\ell}_{t,i}-\ell_{t,i_t})^2 \\
&\le \eta_t^{-1}\sum_{i=1}^{K}\frac{\alpha}{2(\alpha-1)}\max\{x_{t,i}, z_{t,i}\}^{2-1/\alpha}\eta_t^2(\hat{\ell}_{t,i}-\ell_{t,i_t})^2 \\
&\le \eta_t^{-1}\sum_{i=1}^{K}\frac{\alpha}{2(\alpha-1)}(2^{\frac{\alpha}{2\alpha-1}})^{2-1/\alpha}x_{t,i}^{2-1/\alpha}\eta_t^2(\hat{\ell}_{t,i}-\ell_{t,i_t})^2 \\
&= \frac{\alpha}{\alpha-1}\eta_t\sum_{i=1}^{K}x_{t,i}^{2-1/\alpha}(\hat{\ell}_{t,i}-\ell_{t,i_t})^2 \\
&= \frac{\alpha}{\alpha-1}\eta_t\ell_{t,i_t}^2\sum_{i=1}^{K}x_{t,i}^{2-1/\alpha}\left(1-\frac{\mathbb{1}[i_t=i]}{x_{t,i_t}}\right)^2 \\
&\le \frac{\alpha}{\alpha-1}\eta_t r_t^{2-\alpha}|\ell_{t,i_t}|^\alpha\sum_{i=1}^{K}x_{t,i}^{2-1/\alpha}\left(1-\frac{\mathbb{1}[i_t=i]}{x_{t,i_t}}\right)^2 \\
&\overset{(c)}{\le} \frac{\alpha}{\alpha-1}t^{1/\alpha-1}\sigma^{1-\alpha}\Theta_\alpha^{2-\alpha}|\ell_{t,i_t}|^\alpha x_{t,i_t}^{2/\alpha-1}\sum_{i=1}^{K}x_{t,i}^{2-1/\alpha}\left(1-\frac{\mathbb{1}[i_t=i]}{x_{t,i_t}}\right)^2 \\
&\overset{(d)}{\le} 2t^{1/\alpha-1}\sigma^{1-\alpha}|\ell_{t,i_t}|^\alpha x_{t,i_t}^{2/\alpha-1}\sum_{i=1}^{K}x_{t,i}^{2-1/\alpha}\left(1-\frac{\mathbb{1}[i_t=i]}{x_{t,i_t}}\right)^2
\end{aligned}
$$

where step (a) is due to the duality property of Bregman divergences, step (b) regards $D_{\Psi^*}(\cdot,\cdot)$ as a second-order Lagrange remainder. step (c) plugs in $\eta_t^{-1} = \sigma t^{1/\alpha}$ and $r_t = \Theta_\alpha\eta_t^{-1}x_{t,i_t}^{1/\alpha}$, thus $\eta_t r_t^{2-\alpha} = t^{1/\alpha-1}\sigma^{1-\alpha}\Theta_\alpha^{2-\alpha}x_{t,i_t}^{2/\alpha-1}$. Step (d) uses $\Theta_\alpha \le (2-\frac{2}{\alpha})^{\frac{1}{2-\alpha}}$ and thus $\Theta_\alpha^{2-\alpha} \le 2\cdot\frac{\alpha-1}{\alpha}$.

After taking expectations, we get

$$
\begin{aligned}
\mathbb{E}\left[\eta_t^{-1}D_\Psi(x_t,z_t)\mid\mathcal{F}_{t-1}\right] &\le 2t^{1/\alpha-1}\sigma\sum_{i=1}^{K}x_{t,i}^{2/\alpha}\left[\underbrace{\sum_{j=1}^{K}x_{t,j}^{2-1/\alpha}}_{\le 1\le x_{t,i}^{-1/\alpha}} - 2x_{t,i}^{1-1/\alpha} + x_{t,i}^{-1/\alpha}\right] \\
&\le 2\sigma t^{1/\alpha-1}\cdot 2\left[-\sum_{i=1}^{K}x_{t,i}^{1+1/\alpha} + \sum_{i=1}^{K}x_{t,i}^{1/\alpha}\right] \\
&= 4\sigma t^{1/\alpha-1}\sum_{i=1}^{K}x_{t,i}^{1/\alpha}(1-x_{t,i}) \\
&\le 8\sigma t^{1/\alpha-1}\sum_{i\neq i^*}x_{t,i}^{1/\alpha},
\end{aligned}
$$

where the last step is due to the fact that $1-x_{t,i^*} = \sum_{i\neq i^*}x_{t,i} \le \sum_{i\neq i^*}x_{t,i}^{1/\alpha}$ and $1-x_{t,i}\le 1$ for any $i\neq i^*$. After applying Lemma E.4, we get

$$
\mathbb{E}\left[\eta_t^{-1}D_\Psi(x_t,z_t)\mid\mathcal{F}_{t-1}\right] \le 8\sigma t^{1/\alpha-1}K^{1-1/\alpha}.
$$

Hence, we have

$$
\mathbb{E}\left[\sum_{t=1}^{T}\eta_t^{-1}D_\Psi(x_t,z_t)\right] = \sum_{t=1}^{T}\mathbb{E}\left[\eta_t^{-1}D_\Psi(x_t,z_t)\mid\mathcal{F}_{t-1}\right] \le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*}8\sigma t^{1/\alpha-1}x_{t,i}^{1/\alpha}\right]
$$

$$\leq \sum_{t=1}^{T} 8\sigma t^{1/\alpha-1} K^{1-1/\alpha}.$$

□

B. Formal Analysis of `OptTINF` (Algorithm 2)

B.1. Main Theorem

In this section, we present a formal proof of Theorem 4.2. We still state a regret guarantee without any big-Oh notation first.

Theorem B.1 (Regret Guarantee of Algorithm 2)**.** *If Assumptions 3.1 and 3.6 hold, Algorithm 2 enjoys:*

1. *For adversarial environments, the regret is bounded by*

$$\mathcal{R}_T \leq 26\sigma^\alpha K^{\frac{\alpha-1}{2}} (T+1)^{\frac{3-\alpha}{2}} + 4\sqrt{K(T+1)}.$$

2. *Moreover, if the environment is stochastically constrained with a unique best arm i^* (Assumption 3.5), then Algorithm 2 enjoys*

$$\begin{aligned}\mathcal{R}_T \leq\ & 2 \times 4^{\frac{3-\alpha}{\alpha-1}} 5^{\frac{2}{\alpha-1}} \sigma^{\frac{2\alpha}{\alpha-1}} \sum_{i \neq i^*} \Delta_i^{\frac{\alpha-3}{\alpha-1}} \ln(T+1) \\ & + \frac{32\sigma}{\alpha-1} \sum_{i \neq i^*} \Delta_i^{-1} \ln(T+1) \\ & + 2 \times 8^{\frac{2}{\alpha-1}} 4^{\frac{3-\alpha}{\alpha-1}} \sigma^{\frac{2\alpha}{\alpha-1}} \sum_{i \neq i^*} \Delta_i^{\frac{\alpha-3}{\alpha-1}} \ln(T+1).\end{aligned}$$

Proof. In Algorithm 2, when the parameters are set as $\alpha = 2$ and $\sigma = 1$, we have $\eta_t^{-1} = \sqrt{t}$ and $r_t = \Theta_2 \sqrt{t} \sqrt{x_{t,i_t}}$ where $\Theta_2 = 1 - 2^{-1/3}$ is an absolute constant. From now on, to avoid confusion, we use α, σ only to denote the real (hidden) parameters of the environment, instead of the parameters of the algorithm.

Following the proof of Theorem 4.1 in Appendix A, we still decompose $\mathcal{R}_T(y)$ for $y = \mathbf{e}_{i^*}$ into $\mathcal{R}_T^s$ and $\mathcal{R}_T^f$, as follows.

$$\mathcal{R}_T(y) \triangleq \sum_{t=1}^{T} \mathbb{E}[\langle x_t - y, \mu_t \rangle] = \mathbb{E}\left[\sum_{t=1}^{T} \langle x_t - y, \mu_t - \mu_t' \rangle\right] + \mathbb{E}\left[\sum_{t=1}^{T} \langle x_t - y, \mu_t' \rangle\right] \triangleq \mathcal{R}_T^s + \mathcal{R}_T^f.$$

Following the analysis of Algorithm 1, we have the following lemma.

Lemma B.2. *For the given $y = \mathbf{e}_{i^*} \in \triangle_{[K]}$, Algorithm 2 ensures*

$$\mathcal{R}_{\mathcal{T}}^s \leq \mathbb{E}\left[5\sigma^\alpha \sum_{t=1}^{T} \sum_{i \neq i^*} t^{1/2-\alpha/2} x_{t,i}^{(3-\alpha)/2}\right],$$

which further implies

$$\mathcal{R}_{\mathcal{T}}^s \leq 5\sigma^\alpha \sum_{t=1}^{T} t^{\frac{1-\alpha}{2}} K^{\frac{\alpha-1}{2}}.$$

We continue our analysis by bounding the FTRL part, $\mathcal{R}_T^f$. As in Appendix A, we also decompose it into two parts from Lemma E.9:

$$\sum_{t=1}^{T} \langle x_t - y, \hat{\ell}_t \rangle \leq \underbrace{\sum_{t=1}^{T} \left(\eta_t^{-1} - \eta_{t-1}^{-1}\right) \left(\Psi(y) - \Psi(x_t)\right)}_{\text{Part (A)}} + \underbrace{\sum_{t=1}^{T} \eta_t^{-1} D_\Psi(x_t, z_t)}_{\text{Part (B)}},$$

where $z_t \triangleq \nabla\Psi^*(\nabla\Psi(x_t) - \eta_t \mathbb{1}[|\ell_{t,i_t}| \le r_t](\hat{\ell}_t - \ell_{t,i_t}\mathbf{1}))$. For Part (A), from Lemma A.4, we have (recall that Ψ is now $\frac{1}{2}$-Tsallis entropy)

Lemma B.3. *For part (A), for any one-hot vector $y \in \triangle_{[K]}$, Algorithm 2 ensures*

$$\mathbb{E}[(A)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[(\eta_t^{-1} - \eta_{t-1}^{-1})(\Psi(y) - \Psi(x_t)) \mid \mathcal{F}_{t-1}]\right] \le \mathbb{E}\left[\sum_{t=1}^{T} 2t^{-1/2} \sum_{i \neq i^*} x_{t,i}^{1/2}\right],$$

which further implies

$$\mathbb{E}[(A)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[(\eta_t^{-1} - \eta_{t-1}^{-1})(\Psi(y) - \Psi(x_t)) \mid \mathcal{F}_{t-1}]\right] \le \sum_{t=1}^{T} 2t^{-1/2} K^{1/2}.$$

For part (B), we have

Lemma B.4. *For part (B), Algorithm 2 ensures*

$$\mathbb{E}[(B)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[\eta_t^{-1} D_\Psi(x_t, z_t) \mid \mathcal{F}_{t-1}]\right] \le \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} 8\Theta_2^{2-\alpha}\sigma^\alpha t^{\frac{1-\alpha}{2}} x_{t,i}^{(3-\alpha)/2}\right],$$

which further implies

$$\mathbb{E}[(B)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[\eta_t^{-1} D_\Psi(x_t, z_t) \mid \mathcal{F}_{t-1}]\right] \le \sum_{t=1}^{T} 8\Theta_2^{2-\alpha}\sigma^\alpha t^{\frac{1-\alpha}{2}} K^{\frac{\alpha-1}{2}}.$$

Therefore, for adversarial case (i.e., the first statement), we have

$$\begin{aligned}
\mathcal{R}_T = \mathcal{R}_T^s + \mathcal{R}_T^f &\le \mathcal{R}_T^s + \mathbb{E}[(\mathrm{A})] + \mathbb{E}[(\mathrm{B})] \\
&\le 13\sigma^\alpha \sum_{t=1}^{T} t^{\frac{1-\alpha}{2}} K^{\frac{\alpha-1}{2}} + 2\sum_{t=1}^{T} t^{-1/2} K^{1/2} \\
&\le 26\sigma^\alpha K^{\frac{\alpha-1}{2}} (T+1)^{\frac{3-\alpha}{2}} + 4\sqrt{K(T+1)},
\end{aligned}$$

where the last step uses Lemma E.5.

Moreover, for the stochastically constrained case with a unique best arm $i^* \in [K]$, with the help of AM-GM inequality, we bound each of $\mathcal{R}_T^s$, $\mathbb{E}[(\mathrm{A})]$ and $\mathbb{E}[(\mathrm{B})]$ by

$$\begin{aligned}
\mathcal{R}_T^s &\le \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} \left(5^{\frac{2}{\alpha-1}} \sigma^{\frac{2\alpha}{\alpha-1}} \left(\frac{\Delta_i}{4}\right)^{-\frac{3-\alpha}{\alpha-1}} \frac{1}{t}\right)^{\frac{\alpha-1}{2}} \left(\frac{\Delta_i}{4} x_{t,i}\right)^{\frac{3-\alpha}{2}}\right] \\
&\le \frac{\alpha-1}{2} 4^{\frac{3-\alpha}{\alpha-1}} 5^{\frac{2}{\alpha-1}} \sigma^{\frac{2\alpha}{\alpha-1}} \sum_{i \neq i^*} \Delta_i^{\frac{\alpha-3}{\alpha-1}} \ln(T+1) + \frac{3-\alpha}{2} \frac{\mathcal{R}_T}{4}, \\
\mathbb{E}[(\mathrm{A})] &\le \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} \left(4\sigma \left(\frac{\Delta_i}{4}\right)^{-1} \frac{1}{t}\right)^{1/2} \left(\frac{\Delta_i}{4} x_{t,i}\right)^{1/2}\right] \\
&\le \frac{1}{2} \cdot 16\sigma \sum_{i \neq i^*} \Delta_i^{-1} \ln(T+1) + \frac{1}{2} \frac{\mathcal{R}_T}{4}, \\
\mathbb{E}[(\mathrm{B})] &\le \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} \left(8^{\frac{2}{\alpha-1}} \sigma^{\frac{2\alpha}{\alpha-1}} \left(\frac{\Delta_i}{4}\right)^{-\frac{3-\alpha}{\alpha-1}} \frac{1}{t}\right)^{\frac{\alpha-1}{2}} \left(\frac{\Delta_i}{4} x_{t,i}\right)^{\frac{3-\alpha}{2}}\right]
\end{aligned}$$

$$\leq \frac{\alpha-1}{2} 8^{\frac{2}{\alpha-1}} 4^{\frac{3-\alpha}{\alpha-1}} \sigma^{\frac{2\alpha}{\alpha-1}} \sum_{i \neq i^*} \Delta_i^{\frac{\alpha-3}{\alpha-1}} \ln(T+1) + \frac{3-\alpha}{2} \frac{\mathcal{R}_T}{4}.$$

Therefore, we have

$$\begin{aligned}(1 - \frac{(2-\alpha)+1+(3-\alpha)}{2}\frac{1}{4})\mathcal{R}_T = \frac{\alpha-1}{4}\mathcal{R}_T \leq \frac{\alpha-1}{2} 4^{\frac{3-\alpha}{\alpha-1}} 5^{\frac{2}{\alpha-1}} \sigma^{\frac{2\alpha}{\alpha-1}} \sum_{i \neq i^*} \Delta_i^{\frac{\alpha-3}{\alpha-1}} \ln(T+1) \\ + \frac{1}{2} \cdot 16\sigma \sum_{i \neq i^*} \Delta_i^{-1} \ln(T+1) \\ + \frac{\alpha-1}{2} 8^{\frac{2}{\alpha-1}} 4^{\frac{3-\alpha}{\alpha-1}} \sigma^{\frac{2\alpha}{\alpha-1}} \sum_{i \neq i^*} \Delta_i^{\frac{\alpha-3}{\alpha-1}} \ln(T+1),\end{aligned}$$

which gives our result. □

Proof of Theorem 4.2. It is a direct consequence of the theorem above. □

B.2. Proof when Bounding $\mathcal{R}_T^s$ (*the skipped part*)

Proof of Lemma B.2. For any $t \in [T]$ and $i \in [K]$, we can bound between the difference between the loss mean, $\mu_{t,i}$, and the truncated loss mean, $\mu'_{t,i}$, as

$$\begin{aligned}\mu_{t,i} - \mu'_{t,i} &= \mathbb{E}[\ell_{t,i} \mathbb{1}[|\ell_{t,i} > r_t|] \mid \mathcal{F}_{t-1}, i_t = i] \leq \mathbb{E}[|\ell_{t,i}|^\alpha r_t^{1-\alpha} \cdot \mathbb{1}[|\ell_{t,i}| > r_t] \mid \mathcal{F}_{t-1}, i_t = i] \\ &\leq \mathbb{E}[|\ell_{t,i}|^\alpha r_t^{1-\alpha} \mid \mathcal{F}_{t-1}, i_t = i] \leq \sigma^\alpha \Theta_2^{1-\alpha} t^{\frac{1-\alpha}{2}} x_{t,i}^{\frac{1-\alpha}{2}}.\end{aligned}$$

Hence, we have

$$\begin{aligned}\mathcal{R}_T^s = \sum_{t=1}^T \mathbb{E}\left[\langle x_t - y, \mu_t - \mu'_t \rangle\right] &\leq \mathbb{E}\left[\sigma^\alpha \Theta_2^{1-\alpha} \sum_{t=1}^T \sum_{i \neq i^*} t^{1/2-\alpha/2} x_{t,i}^{1/2-\alpha/2} \cdot x_{t,i}\right] \\ &\leq \mathbb{E}\left[5\sigma^\alpha \sum_{t=1}^T \sum_{i \neq i^*} t^{1/2-\alpha/2} x_{t,i}^{3/2-\alpha/2}\right],\end{aligned}$$

where the last step uses $\Theta_2^{1-\alpha} \leq \Theta_2^{-1} \leq 5$. It further gives, by Lemma E.4, that

$$\mathcal{R}_T^s \leq 5\sigma^\alpha \sum_{t=1}^T t^{1/2-\alpha/2} K^{\alpha/2-1/2}.$$

□

B.3. Proof when Bounding $\mathcal{R}_T^f$ (*the FTRL part*)

Proof of Lemma B.3. This is just a restatement of Lemma A.4. □

Proof of Lemma B.4. We simply follow the proof of Lemma A.5, except for some slight modifications (instead of the previous lemma, we cannot directly modify all α's to 2, as the second moment of ℓ_{t,i_t} may not exist). The first few steps are exactly the same, giving

$$\begin{aligned}\eta_t^{-1} D_\Psi(x_t, z_t) &\leq \frac{2}{2-1} \eta_t r_t^{2-\alpha} |\ell_{t,i_t}|^\alpha \sum_{i=1}^K x_{t,i}^{2-1/2} \left(1 - \frac{\mathbb{1}[i_t = i]}{x_{t,i_t}}\right)^2 \\ &\leq 2\left(t^{1/2}\right)^{-1} \Theta_2^{2-\alpha} \left(t^{1/2}\right)^{2-\alpha} x_{t,i_t}^{\frac{2-\alpha}{2}} |\ell_{t,i_t}|^\alpha \sum_{i=1}^K x_{t,i}^{2-1/2} \left(1 - \frac{\mathbb{1}[i_t = i]}{x_{t,i_t}}\right)^2\end{aligned}$$

$$= 2\Theta_2^{2-\alpha}|\ell_{t,i_t}|^\alpha t^{\frac{1-\alpha}{2}} x_{t,i_t}^{\frac{2-\alpha}{2}} \sum_{i=1}^K x_{t,i}^{2-1/2} \left(1 - \frac{\mathbb{1}[i_t = i]}{x_{t,i_t}}\right)^2.$$

After taking expectations, we have

$$\begin{aligned}
\mathbb{E}[\eta_t^{-1} D_\Psi(x_t, z_t) \mid \mathcal{F}_{t-1}] &\le 2\Theta_2^{2-\alpha}\sigma^\alpha t^{\frac{1-\alpha}{2}} \sum_{i=1}^K x_{t,i}^{2-\alpha/2} \left[\underbrace{\sum_{j=1}^K x_{t,j}^{3/2}}_{\le 1 \le x_{t,i}^{-1/2}} - 2x_{t,i}^{1/2} + x_{t,i}^{-1/2} \right] \\
&\le 4\Theta_2^{2-\alpha}\sigma^\alpha t^{\frac{1-\alpha}{2}} \left[\sum_{i=1}^K x_{t,i}^{3/2-\alpha/2} - \sum_{i=1}^K x_{t,i}^{5/2-\alpha/2} \right] \\
&= 4\Theta_2^{2-\alpha}\sigma^\alpha t^{\frac{1-\alpha}{2}} \left[\sum_{i=1}^K x_{t,i}^{3/2-\alpha/2} (1 - x_{t,i}) \right] \\
&\le 8\Theta_2^{2-\alpha}\sigma^\alpha t^{\frac{1-\alpha}{2}} \sum_{i \ne i^*} x_{t,i}^{3/2-\alpha/2}
\end{aligned}$$

Therefore, we have

$$\mathbb{E}\left[\sum_{t=1}^T \eta_t^{-1} D_\Psi(x_t, z_t)\right] \le \mathbb{E}\left[\sum_{t=1}^T \sum_{i \ne i^*} 8\Theta_2^{2-\alpha}\sigma^\alpha t^{\frac{1-\alpha}{2}} x_{t,i}^{(3-\alpha)/2}\right],$$

which further gives

$$\mathbb{E}\left[\sum_{t=1}^T \eta_t^{-1} D_\Psi(x_t, z_t)\right] \le \sum_{t=1}^T 8\Theta_2^{2-\alpha}\sigma^\alpha t^{\frac{1-\alpha}{2}} K^{\frac{\alpha-1}{2}}$$

by Lemma E.4. □

C. Formal Analysis of `AdaTINF` (Algorithm 3)

C.1. Main Theorem

We again begin with a regret guarantee without any big-Oh notations.

Theorem C.1 (Regret Guarantee of Algorithm 3)**.** *If Assumptions 3.1 and 3.6 hold, Algorithm 3 ensures*

$$\mathcal{R}_T \le 3\sqrt{K(T+1)} + 204\sigma K^{1-1/\alpha}(T+1)^{1/\alpha} + 12\sigma T^{1/\alpha}.$$

Proof. As defined in the text, we group time slots with equal λ_t's into epochs, as

$$\mathcal{T}_j \triangleq \{t \in [T] \mid \lambda_t = 2^j\}, \quad \forall j \ge 0.$$

For any non-empty $\mathcal{T}_j$'s, denote the first and last time slot of $\mathcal{T}_j$ by

$$\gamma_j \triangleq \min\{t \in \mathcal{T}_j\}, \tau_j \triangleq \max\{t \in \mathcal{T}_j\}.$$

Then, without loss of generality, assume that no doubling has happened for time slot T. Otherwise, one can always add a virtual time slot $t = T+1$ with $\ell_{t,i} = 0$ for all i . Therefore, we have $\mathcal{T}_J \ne \varnothing$ where J is the final value of variable J defined in the code.

We adopt the notation of c_t as defined in Algorithm 3:

$$c_t = 2\eta_t x_{t,i_t}^{-1/2} \ell_{t,i_t^2} \mathbb{1}[|\ell_{t,i_t}| \le r_t] + \ell_{t,i_t} \mathbb{1}[|\ell_{t,i_t}| > r_t].$$

Moreover, from the doubling criterion of Algorithm 3, for each non-empty epoch, we have the following lemma.

Lemma C.2. *For any $0 \le j < J$ such that $\mathcal{T}_j \neq \varnothing$, we have*

$$\mathbb{1}[\gamma_j > 1] c_{\gamma_j - 1} + \sum_{t \in \mathcal{T}_j \setminus \{\tau_j\}} c_t \le 2^j \sqrt{K(T+1)}, \tag{26}$$

$$\sum_{t \in \mathcal{T}_j} c_t > 2^{j-1} \sqrt{K(T+1)}, \tag{27}$$

Moreover, for $j = J$ (recall that $\mathcal{T}_J \neq \varnothing$), we have

$$\mathbb{1}[\gamma_j > 1] c_{\gamma_j - 1} + \sum_{t \in \mathcal{T}_j} c_t \le 2^J \sqrt{K(T+1)}. \tag{28}$$

Similar to previous analysis, we define $\mu'_{t,i} = \mathbb{E}[\ell_{t,i} \mathbb{1}[|\ell_{t,i}| \le r_t] \mid \mathcal{F}_{t-1}, i_t = i]$ and decompose the regret $\mathcal{R}_T(y)$ as follows

$$\mathcal{R}_T(y) = \mathbb{E}\left[\sum_{t=1}^T \langle x_t - y, \mu_t - \mu'_t \rangle\right] + \mathbb{E}\left[\sum_{t=1}^T \langle x_t - y, \hat{\ell}_t \rangle\right] \triangleq \mathcal{R}_T^s + \mathcal{R}_T^f.$$

According to Lemma A.3, we have

$$\begin{aligned} R_T^s &\le \mathbb{E}\left[\sum_{t=1}^T \sum_{i=1}^K x_{t,i} (\mu_{t,i} - \mu'_{t,i})\right] \\ &= \mathbb{E}\left[\sum_{t=1}^T \sum_{i=1}^K \ell_{t,i_t} \cdot \mathbb{1}[|\ell_{t,i_t}| > r_t]\right]. \end{aligned} \tag{29}$$

Furthermore, due to the properties of weighted importance sampling estimator (as in Appendix A, $\mathbb{E}[\hat{\ell}_{t,i} \mid \mathcal{F}_{t-1}] = \mu'_{t,i}$), we have

$$\mathcal{R}_T^f = \mathbb{E}\left[\sum_{t=1}^T \langle x_t - y, \hat{\ell}_t \rangle\right].$$

We can then apply Lemma E.9 to $\mathcal{R}_T^f$, giving

$$\sum_{t=1}^T \langle x_t - y, \hat{\ell}_t \rangle \le \eta_T \max_{x \in \triangle_{[K]}} \Psi(x) + \sum_{t=1}^T \eta_t^{-1} D_\Psi(x_t, z_t)$$

where $z_t \triangleq \nabla \Psi^*(\nabla \Psi(x_t) - \eta_t \hat{\ell}_t)$. The first term is simply within $2^J \sqrt{KT}$. For the second term, we have the following property similar to Lemma A.5:

Lemma C.3. *Algorithm 3 guarantees that for any $t \in [T]$,*

$$\eta_t^{-1} D_\Psi(x_t, z_t) \le 2 \eta_t x_{t,i_t}^{3/2} \hat{\ell}_{t,i_t}^2$$

where $z_t \triangleq \nabla \Psi^*(\nabla \Psi(x_t) - \eta_t \hat{\ell}_t)$.

Thus we have

$$\begin{aligned} \mathcal{R}_T^f &\le \mathbb{E}[2^J] \sqrt{KT} + \mathbb{E}\left[\sum_{t=1}^T 2 \eta_t x_{t,i_t}^{3/2} \hat{\ell}_{t,i_t}^2\right] \\ &= \mathbb{E}[2^J] \sqrt{KT} + \mathbb{E}\left[\sum_{t=1}^T 2 \eta_t x_{t,i_t}^{-1/2} \ell_{t,i_t}^2 \cdot \mathbb{1}[|\ell_{t,i_t}| \le r_t]\right]. \end{aligned} \tag{30}$$

Combining Eq. (30) and (30), we can see

$$\mathcal{R}_T \le \mathbb{E}[2^J]\sqrt{KT} + \mathbb{E}\left[\sum_{t=1}^{T}\left(\ell_{t,i_t}\cdot\mathbb{1}[|\ell_{t,i_t}|>r_t] + 2\eta_t x_{t,i_t}^{-1/2}\ell_{t,i_t}^2\cdot\mathbb{1}[|\ell_{t,i_t}|\le r_t]\right)\right]$$
$$= \mathbb{E}[2^J]\sqrt{KT} + \mathbb{E}\left[\sum_{t=1}^{T} c_t\right].$$

Summing up Equation (26) for all non-empty epoch $j < J$ and Equation (28), we get

$$\sum_{t=1}^{T} c_t = \sum_{j=0}^{J}\sum_{t\in\mathcal{T}_j} c_t \le \sum_{j=0}^{J} 2^j\sqrt{K(T+1)} \le 2^{J+1}\sqrt{K(T+1)},$$

and we can conclude

$$\mathcal{R}_T \le \mathbb{E}[2^J]\cdot 3\sqrt{K(T+1)}.$$

It remains to bound $\mathbb{E}[2^J]$. When $J \ge 1$, there are at least two non-empty epochs. Let J' be the index of the second last epoch. The doubling condition of Algorithm 3 further reduce the task to bound 2^J into bounding $2^{J'}$ and $c_{\tau_{J'}}$, as the following lemma states.

Lemma C.4. *Algorithm 3 guarantees that, when $J \ge 1$, we have*

$$2^J\sqrt{K(T+1)} \le 2^{J'+1}\sqrt{K(T+1)} + 4c_{\tau_{J'}}. \tag{31}$$

We can derive the following expectation bound for both $2^{J'}$ and $c_{\tau_{J'}}$:

Lemma C.5 (Restatement of Lemma 7.1)**.** *Algorithm 3 guarantees that*

$$\mathbb{E}[\mathbb{1}[J\ge 1]2^{J'}] \le 28\sigma K^{1/2-1/\alpha}(T+1)^{1/\alpha-1/2}. \tag{32}$$

Lemma C.6 (Restatement of Lemma 7.2)**.** *Algorithm 3 guarantees that*

$$\mathbb{E}[\mathbb{1}[J\ge 1]c_{\tau_{J'}}] \le 0.1\,\mathbb{E}[\mathbb{1}[J\ge 1]2^{J'}\sqrt{T}] + \mathbb{E}\left[\max_{t\in[T]}|\ell_{t,i_t}|\right]. \tag{33}$$

Applying Lemma E.3 and Equation (32) to Eqation (33), we get

$$\mathbb{E}[\mathbb{1}[J\ge 1]c_{\tau_{J'}}] \le 3\sigma K^{1/2-1/\alpha}(T+1)^{1/\alpha} + \sigma T^{1/\alpha}.$$

Plugging this into Equation (31), we get

$$\mathbb{E}\left[\mathbb{1}[J\ge 1]2^J\sqrt{K(T+1)}\right] \le 68\sigma K^{1-1/\alpha}(T+1)^{1/\alpha} + 4\sigma T^{1/\alpha},$$

and thus

$$\mathbb{E}\left[2^J\sqrt{K(T+1)}\right] \le 68\sigma K^{1-1/\alpha}(T+1)^{1/\alpha} + 4\sigma T^{1/\alpha} + \sqrt{K(T+1)},$$

$$\mathcal{R}_T \le 3\,\mathbb{E}\left[2^J\sqrt{K(T+1)}\right] \le 204\sigma K^{1-1/\alpha}(T+1)^{1/\alpha} + 12\sigma T^{1/\alpha} + 3\sqrt{K(T+1)}.$$

□

Proof of Theorem 6.1. It is a direct consequence of the theorem above. □

C.2. Proof when Reducing $\mathcal{R}_T$ to $\mathbb{E}[2^J]$

Proof of Lemma C.2. It suffices to notice that in Algorithm 3, during a particular epoch j, when the doubling condition at Line 15 evaluates to true, the current value of the variable S_j is $1[\gamma_j > 1]c_{\gamma_j - 1} + \sum_{t \in \mathcal{T}_j} c_t$, thus

$$1[\gamma_j > 1]c_{\gamma_j - 1} + \sum_{t \in \mathcal{T}_j} c_t > 2^j \sqrt{K(T+1)}.$$

When $\gamma_j = 1$ (or equivalently, $j = 0$), Equation (27) automatically holds. Otherwise Line 16 guarantees that $j \geq \lceil \log_2(c_{\tau_{\gamma_j - 1}} / \sqrt{K(T+1)}) \rceil + 1$, hence $2^{j-1}\sqrt{K(T+1)} \geq c_{\tau_{\gamma_j - 1}}$. We will have

$$2^{j-1}\sqrt{K(T+1)} + \sum_{t \in \mathcal{T}_j} c_t > 2^j \sqrt{K(T+1)},$$

which also solves to Equation (27).

When the doubling condition at Line 15 evaluates to false for the last time, the value of S_j is $1[\gamma_j > 1]c_{\gamma_j - 1} + \sum_{t \in \mathcal{T}_j \setminus \{\tau_j\}} c_t$. At this time we have $S_j \leq 2^j \sqrt{K(T+1)}$, hence Equation (26) and (28) hold. □

Proof of Lemma C.3. It is exactly the same calculation we did in Lemma A.5, the only difference is that $\hat{\ell}_t$ does not come with a $-\ell_{t,i_t}$ drift. □

C.3. Proof when Bounding $\mathbb{E}[2^J]$

Proof of Lemma C.4. According to Line 16 of Algorithm 3, J, J' and $c_{\tau_{J'}}$ satisfy

$$J = \max\left\{J' + 1, \lceil \log_2(c_{\tau_{J'}} / \sqrt{K(T+1)}) \rceil + 1\right\},$$

thus

$$2^J \leq \max\left\{2 \cdot 2^{J'}, 4 \cdot c_{\tau_{J'}} / \sqrt{K(T+1)}\right\}$$

and

$$\begin{aligned} 2^J \sqrt{K(T+1)} &\leq \max\left\{2 \cdot 2^{J'} \sqrt{K(T+1)}, 4 \cdot c_{\tau_{J'}}\right\} \\ &\leq 2 \cdot 2^{J'} \sqrt{K(T+1)} + 4 \cdot c_{\tau_{J'}}. \end{aligned}$$

□

Proof of Lemma C.5. Applying Eq. (27) to $j = J' < J$, we get

$$\mathbb{1}[J \geq 1](2^{J'})^\alpha \sqrt{K(T+1)} \leq (2^{J'})^{\alpha - 1} 2 \sum_{t \in \mathcal{T}_{J'}} c_t \tag{34}$$

We further upper-bound the RHS of (34) by enlarging the summation range to $[T]$. Specifically, let $\tilde{\eta}_t = 2^{-J'} t^{-1/2}$, $\tilde{r}_t = 2^{J'} \Theta_2 \sqrt{t x_{t,i_t}}$. Define the summands by

$$\begin{aligned} \tilde{c}_t &\triangleq 2\tilde{\eta}_t x_{t,i_t}^{-1/2} \ell_{t,i_t}^2 \mathbb{1}[|\ell_{t,i_t}| \leq \tilde{r}_t] + \ell_{t,i_t} \mathbb{1}[|\ell_{t,i_t}| > \tilde{r}_t] \\ &\leq 2\tilde{\eta}_t x_{t,i_t}^{-1/2} |\ell_{t,i_t}|^\alpha \tilde{r}_t^{2-\alpha} + |\ell_{t,i_t}|^\alpha \tilde{r}_t^{1-\alpha} \\ &\leq (2\tilde{\eta}_t \tilde{r}_t^{2-\alpha} x_{t,i_t}^{-1/2} + \tilde{r}_t^{1-\alpha}) \cdot |\ell_{t,i_t}|^\alpha \\ &= (2\Theta_2^{2-\alpha} + \Theta_2^{1-\alpha}) \cdot 2^{(1-\alpha)J'} t^{\frac{1-\alpha}{2}} x_{t,i_t}^{\frac{1-\alpha}{2}} |\ell_{t,i_t}|^\alpha \\ &\leq (2 + \Theta_2^{-1}) \cdot 2^{(1-\alpha)J'} t^{\frac{1-\alpha}{2}} x_{t,i_t}^{\frac{1-\alpha}{2}} |\ell_{t,i_t}|^\alpha \\ &\leq 7 \cdot 2^{(1-\alpha)J'} t^{\frac{1-\alpha}{2}} x_{t,i_t}^{\frac{1-\alpha}{2}} |\ell_{t,i_t}|^\alpha. \end{aligned} \tag{35}$$

We see that the definition in Eq. (35) coincides with c_t for $t \in \mathcal{T}_{J'}$. Thus, the RHS of (34) is no more than

$$14 \sum_{t=1}^{T} t^{\frac{1-\alpha}{2}} x_{t,i_t}^{\frac{1-\alpha}{2}} |\ell_{t,i_t}|^{\alpha}$$

Taking expectation on both sides of (34), we get

$$\mathbb{E}\left[\mathbb{1}[J \geq 1](2^{J'})^{\alpha}\right] \sqrt{K(T+1)} \leq 28\sigma^{\alpha} K^{\frac{\alpha-1}{2}} (T+1)^{\frac{3-\alpha}{2}},$$

which gives $\mathbb{E}[\mathbb{1}[J \geq 1](2^{J'})^{\alpha}] \leq 28\sigma^{\alpha} K^{\alpha/2-1}(T+1)^{1-\alpha/2}$. By Jensen's inequality,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}[J \geq 1] 2^{J'}\right] &\leq \left(\mathbb{E}\left[\mathbb{1}[J \geq 1]^{\alpha} \left(2^{J'}\right)^{\alpha}\right]\right)^{1/\alpha} \\ &\leq 28\sigma K^{1/2-1/\alpha}(T+1)^{1/\alpha-1/2}.\end{aligned}$$

□

Proof of Lemma C.6. We can do the calculation

$$\begin{aligned}c_{\tau_{J'}} &= 2\eta_{\tau_{J'}} x_{\tau_{J'},i_{\tau_{J'}}}^{-1/2} \ell_{\tau_{J'},i_{\tau_{J'}}}^2 \mathbb{1}[|\ell_{\tau_{J'},i_{\tau_{J'}}}| \leq r_{\tau_{J'}}] + \ell_{\tau_{J'},i_{\tau_{J'}}} \mathbb{1}[|\ell_{\tau_{J'},i_{\tau_{J'}}}| > r_{\tau_{J'}}] \\ &\leq 2\eta_{\tau_{J'}} x_{\tau_{J'},i_{\tau_{J'}}}^{-1/2} r_{\tau_{J'}}^2 + \max_{t\in[T]} |\ell_{t,i_t}| \\ &= 2^{J'} \cdot 2\Theta_2^2 \sqrt{\tau_{J'}} x_{\tau_{J'},i_{\tau_{J'}}}^{1/2} + \max_{t\in[T]} |\ell_{t,i_t}| \\ &\leq 0.1 \cdot 2^{J'} \sqrt{T} + \max_{t\in[T]} |\ell_{t,i_t}|.\end{aligned}$$

□

D. Removing Dependency on Time Horizon T in Algorithm 3

To remove the dependency of T, we leverage the following doubling trick, which is commonly used for unknown T's (Auer et al., 1995; Besson & Kaufmann, 2018). This gives our More Adaptive `AdaTINF` algorithm, which we called `Ada`^{2}`TINF`.

Algorithm 4 More Adaptive `AdaTINF` (`Ada`^{2}`TINF`)

Input: Number of arms K
Output: Sequence of actions $i_1, i_2, \cdots, i_T \in [K]$

Initialize $T_0 \leftarrow 1, S \leftarrow 0$
for $t = 1, 2, \cdots$ **do**
  if $t \geq S$ **then**
    $T_0 \leftarrow 2T_0$, $S \leftarrow S + T_0 - 1$
    Initialize a new `AdaTINF` instance (Algorithm 3) with parameters K and $T_0 - 1$
  end if
  Run current `AdaTINF` instance for one time slot, act what it acts and feed it with the feedback ℓ_{t,i_t}
end for

Theorem D.1 (Regret Guarantee of Algorithm 4)**.** *Under the same assumptions of Theorem 6.1, i.e., Assumptions 3.1 and 3.6 hold,* `Ada`^{2}`TINF` *(Algorithm 4) ensures*

$$\mathcal{R}_T \leq 600\sigma K^{1-1/\alpha}(T+1)^{1/\alpha}.$$

Proof. We divide the time horizon T into several *super-epochs*, each with length $T_0 - 1 = 2^1 - 1, 2^2 - 1, 2^3 - 1, \cdots$. For each of the super-epoch, as we restarted the whole process, we can regard each of them as a independent execution of

AdaTINF. Therefore, by Theorem 6.1, for an super-epoch from t_0 to $t_0 + T_0 - 2$, we have

$$\mathbb{E}\left[\sum_{t=t_0}^{t_0+T_0-2} \langle x_t - \mathbf{e}_{i^*}, \mu_t \rangle\right] = \mathcal{R}_{T_0-1} \le 300\sigma K^{1-1/\alpha} T_0^{1/\alpha}.$$

Therefore, the total regret is bounded by

$$\mathcal{R}_T \le \sum_{T_0 = 2^1-1, 2^2-1, \cdots, 2^{\lceil \log_2(T+1) \rceil}-1} 300\sigma K^{1-1/\alpha}(T_0+1)^{1/\alpha} \le 600\sigma K^{1-1/\alpha} T^{1/\alpha},$$

as desired. □

E. Auxiliary Lemmas

E.1. Probability Lemmas

Lemma E.1. *For a non-negative random variable X whose α-th moment exists and a constant $c > 0$, we have*

$$\Pr\{X \ge c\} \le \frac{\mathbb{E}[X^\alpha]}{c^\alpha}$$

Proof. As both X, c are non-negative, $\Pr\{X \ge c\} = \Pr\{X^\alpha \ge c^\alpha\} \le \frac{\mathbb{E}[X^\alpha]}{c^\alpha}$ by Markov's inequality. □

Lemma E.2. *For a random variable Y with q-th moment $\mathbb{E}[|Y|^q]$ bounded by σ^q (where $q \in [1, 2]$), its p-th moment $\mathbb{E}[|Y|^p]$ is also bounded by σ^p if $1 \le p \le q$.*

Proof. As the function $f\colon x \mapsto x^\alpha$ is convex for any $\alpha \ge 1$, by Jensen's inequality, we have $f(\mathbb{E}[X]) \le \mathbb{E}[f(X)]$ for any random variable X. Hence, by picking $X = |Y|^p$ and $\alpha = \frac{q}{p}$, we have $(\mathbb{E}[|Y|^p])^{q/p} \le \mathbb{E}[(|Y|^p)^{q/p}] = \mathbb{E}[|Y|^q] \le \sigma^q$, so $\mathbb{E}[|Y|^p] \le \sigma^p$ for any $1 \le p \le q$. □

Lemma E.3. *For n independent random variables $X_1, X_2, \cdots, X_n$, each with α-th moment ($1 < \alpha \le 2$) bounded by σ^α, i.e., $\mathbb{E}_{x_i \sim X_i}[|x_i|^\alpha] \le \sigma^\alpha$ for all $1 \le i \le n$, we have*

$$\mathop{\mathbb{E}}_{x_1 \sim X_1, x_2 \sim X_2, \cdots, x_n \sim X_n} \left[\max_{1 \le i \le n} |x_i|\right] \le \sigma n^{1/\alpha}.$$

Proof. By Jensen's inequality, we have (here, $\mathbf{x} \sim \mathbf{X}$ denotes $x_1 \sim X_1, x_2 \sim X_2, \cdots, x_n \sim X_n$)

$$\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbf{X}}\left[\max_{1\le i\le n}|x_i|\right]\right)^\alpha \le \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbf{X}}\left[\left(\max_{1\le i\le n}|x_i|\right)^\alpha\right] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbf{X}}\left[\max_{1\le i\le n}|x_i|^\alpha\right] \le \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbf{X}}\left[\sum_{i=1}^n |x_i|^\alpha\right] = \sum_{i=1}^n \mathop{\mathbb{E}}_{x_i\sim X_i}[|x_i|^\alpha] \le n\sigma^\alpha,$$

which gives $\mathbb{E}_{\mathbf{x}\sim\mathbf{X}}\left[\max_{1\le i\le n}|x_i|\right] \le \sigma n^{1/\alpha}$. □

E.2. Arithmetic Lemmas

Lemma E.4. *For any $x \in \triangle_{[K]}$ (i.e., $\sum_{i=1}^K x_i = 1$), we have*

$$\sum_{i=1}^K x_i^t \le K^{1-t}$$

for $\frac{1}{2} \le t < 1$.

Proof. By Hölder's inequality $\|fg\|_1 \le \|f\|_p \|g\|_q$, we have $\sum_{i=1}^K x_i^t \le (\sum_{i=1}^K (x_i^t)^{1/t})^t (\sum_{i=1}^K 1^q)^{1/q} = K^{1-t}$ by picking $p = \frac{1}{t}$ and $q = \frac{1}{1-t}$. □

Lemma E.5. *For any positive integer n, we have*

$$\sum_{i=1}^{n} \frac{1}{i} \le \ln(n+1).$$

Moreover, for any $-1 < t < 0$, we have

$$\sum_{i=1}^{n} i^t \le \frac{(n+1)^{t+1}}{t+1}.$$

Proof. If $t = -1$, we have $\sum_{i=1}^{n} i^t \le \int_1^{(} n+1) \frac{\mathrm{d}x}{x} = \ln n$. If $t > -1$, we have $\sum_{i=1}^{n} i^t \le \int_0^{n+1} x^t \,\mathrm{d}x = \frac{(n+1)^{t+1}}{t+1}$. □

Lemma E.6. *For any $x \ge 1$ and $q \in (0,1)$, we have*

$$x^q - (x-1)^q \le q(x-1)^{q-1}.$$

Proof. Consider the function f defined by $x \mapsto x^q$. We have $f''(x) = q(q-1)x^{q-2} \le 0$ for $x \ge 0$ and $q \in (0,1)$. Hence, $f(x)$ is concave for $x \ge 0$ and $q \in (0,1)$. Therefore, by properties of concave functions, we have $f(x) \le f(x-1) + f'(x-1)(x-(x-1)) = f(x-1) + q(x-1)^{q-1}$ for any $x \ge 1$ and $q \in (0,1)$, which gives $x^q - x^{q-1} \le q(x-1)^{q-1}$. □

E.3. Lemmas on the FTRL Framework for MAB Algorithm Design

Lemma E.7. *For any algorithm that plays action $i_t \sim x_t$ where $\{x_t\}_{t=1}^T$ can be regarded as a stochastic process adapted to the natural filtration $\{\mathcal{F}_t\}_{t=0}^T$, its regret, in a stochastically constrained adversarial environment with unique best arm $i^* \in [K]$, is lower-bounded by*

$$\mathcal{R}_T \ge \sum_{t\in[T]} \sum_{i\ne i^*} \Delta_i \,\mathbb{E}[x_{t,i} \mid \mathcal{F}_{t-1}].$$

Proof. By definition of $\mathcal{R}_T$ and Δ_i, we have

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \langle x_t - \mathbf{e}_{i^*}, \mu_t\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\ne i^*} x_{t,i}\mu_{t,i} - (1-x_{t,i^*})\mu_{t,i^*}\right] = \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i\ne i^*} x_{t,i}(\mu_{t,i} - \mu_{t,i^*})\right],$$

which is exactly $\sum_{t\in[T]}\sum_{i\ne i^*} \Delta_i \,\mathbb{E}[x_{t,i} \mid \mathcal{F}_{t-1}]$. □

Lemma E.8 (Property of Weighted Importance Sampling Estimator)**.** *For any distribution $x \in \triangle_{[K]}$ and loss vector $\ell \in \mathbb{R}^K$ sampled from a distribution $\nu \in \triangle_{\mathbb{R}^k}$, if we pulled an arm i according to x, then the weighted importance sampler $\tilde{\ell}(j) \triangleq \frac{\ell(j)}{x_j}\mathbb{1}[i=j]$ gives an unbiased estimate of $\mathbb{E}[\ell]$, i.e.,*

$$\mathop{\mathbb{E}}_{i\sim x}\left[\tilde{\ell}(j)\right] = \mathbb{E}[\ell(j)], \quad \forall 1 \le j \le K.$$

Proof. As the adversary is oblivious (or even stochastic),

$$\mathop{\mathbb{E}}_{i\sim x}\left[\tilde{\ell}(j)\right] = \sum_{i=1}^{K} \frac{\mathbb{E}[\ell(j)]}{x_j}\mathbb{1}_{i=j} \cdot \Pr\{\mathbb{1}_i\} = \Pr\{i=j\}\frac{\mathbb{E}[\ell(j)]}{x_j} = \mathbb{E}[\ell(j)],$$

for any $1 \le j \le K$. □

Lemma E.9 (FTRL Regret Decomposition)**.** *For any FTRL algorithm, i.e., the action x_t for any $t \in [T]$ is decided by $\operatorname{argmin}_{x\in\triangle_{[K]}}(\eta_t \sum_{1\le s\le t}\langle \hat{\ell}_s, x\rangle + \Psi(x))$, where η_t is the learning rate, $\hat{\ell}_s$ is some arbitrary vector and $\Psi(x)$ is a convex regularizer, we have*

$$\sum_{t=1}^{T}\langle x_t - y, \hat{\ell}_t\rangle \le \sum_{t=1}^{T}(\eta_t^{-1} - \eta_{t-1}^{-1})(\Psi(y) - \Psi(x_t)) + \sum_{t=1}^{T}\eta_t^{-1} D_\Psi(x_t, z_t)$$

for any $y \in \triangle_{[K]}$, where $z_t \triangleq \nabla\Psi^*(\nabla\Psi(x_t) - \eta_t\hat{\ell}_t)$.

Proof. Let $\hat{L}_t \triangleq \sum_{s=1}^{t}\hat{\ell}_s$, we then have

$$
\begin{aligned}
\sum_{t=1}^{T}\langle x_t - y, \hat{\ell}_t\rangle &= \sum_{t=1}^{T} -\eta_t^{-1}\langle x_t, -\eta_t\hat{\ell}_t\rangle + \langle y, -\hat{L}_T\rangle \\
&= \sum_{t=1}^{T}\eta_t^{-1}\left[\overline{\Psi}^*(-\eta_t\hat{L}_t) - \overline{\Psi}^*(-\eta_t\hat{L}_{t-1}) - \langle x_t, -\eta_t\hat{\ell}_t\rangle\right] \\
&\quad + \sum_{t=1}^{T}\left[\eta_t^{-1}\overline{\Psi}^*(-\eta_t\hat{L}_{t-1}) - \eta_t^{-1}\overline{\Psi}^*(-\eta_t\hat{L}_t)\right] + \langle y, -\hat{L}_T\rangle \\
&= \sum_{t=1}^{T}\eta_t^{-1}D_{\overline{\Psi}^*}(-\eta_t\hat{L}_t, -\eta_t\hat{L}_{t-1}) + \sum_{t=1}^{T}\left[\eta_t^{-1}\overline{\Psi}^*(-\eta_t\hat{L}_{t-1}) - \eta_t^{-1}\overline{\Psi}^*(-\eta_t\hat{L}_t)\right] + \langle y, -\hat{L}_T\rangle \\
&= \sum_{t=1}^{T}\eta_t^{-1}D_{\Psi}(x_t, \nabla\overline{\Psi}^*(-\eta_t\hat{L}_t)) + \sum_{t=1}^{T}\left[\eta_t^{-1}\overline{\Psi}^*(-\eta_t\hat{L}_{t-1}) - \eta_t^{-1}\overline{\Psi}^*(-\eta_t\hat{L}_t)\right] + \langle y, -\hat{L}_T\rangle \\
&\overset{(a)}{\leq} \sum_{t=1}^{T}\eta_t^{-1}D_{\Psi}(x_t, \nabla\Psi^*(-\eta_t\hat{L}_t)) + \sum_{t=1}^{T}\left[\eta_t^{-1}\overline{\Psi}^*(-\eta_t\hat{L}_{t-1}) - \eta_t^{-1}\overline{\Psi}^*(-\eta_t\hat{L}_t)\right] + \langle y, -\hat{L}_T\rangle \\
&= \sum_{t=1}^{T}\eta_t^{-1}D_{\Psi}(x_t, z_t) + \sum_{t=1}^{T}\left[\eta_t^{-1}\overline{\Psi}^*(-\eta_t\hat{L}_{t-1}) - \eta_t^{-1}\overline{\Psi}^*(-\eta_t\hat{L}_t)\right] + \langle y, -\hat{L}_T\rangle \\
&\overset{(b)}{=} \sum_{t=1}^{T}\eta_t^{-1}D_{\Psi}(x_t, z_t) + \sum_{t=1}^{T-1}\left[\langle x_t, -\hat{L}_{t-1}\rangle - \eta_t^{-1}\Psi(x_t) - \sup_{x\in\Delta_{[K]}}\left\{\langle x, -\hat{L}_t\rangle - \eta_t^{-1}\Psi(x)\right\}\right] \\
&\quad + \langle x_T, -\hat{L}_{T-1}\rangle - \eta_T^{-1}\Psi(x_T) - \sup_{x\in\Delta_{[K]}}\left\{\langle x, -\hat{L}_T\rangle - \eta_T^{-1}\Psi(x)\right\} + \langle y, -\hat{L}_T\rangle \\
&\leq \sum_{t=1}^{T}\eta_t^{-1}D_{\Psi}(x_t, z_t) + \sum_{t=1}^{T-1}\left[\langle x_t, -\hat{L}_{t-1}\rangle - \eta_t^{-1}\Psi(x_t) - \langle x_{t+1}, -\hat{L}_t\rangle + \eta_t^{-1}\Psi(x_{t+1})\right] \\
&\quad + \langle x_T, -\hat{L}_{T-1}\rangle - \eta_T^{-1}\Psi(x_T) - \sup_{x\in\Delta_{[K]}}\left\{\langle x, -\hat{L}_T\rangle - \eta_T^{-1}\Psi(x)\right\} + \langle y, -\hat{L}_T\rangle \\
&= \sum_{t=1}^{T}\eta_t^{-1}D_{\Psi}(x_t, z_t) + \sum_{t=1}^{T}(\eta_{t-1}^{-1} - \eta_t^{-1})\Psi(x_t) - \sup_{x\in\Delta_{[K]}}\left\{\langle x, -\hat{L}_T\rangle - \eta_T^{-1}\Psi(x)\right\} + \langle y, -\hat{L}_T\rangle \\
&= \sum_{t=1}^{T}\eta_t^{-1}D_{\Psi}(x_t, z_t) + \sum_{t=1}^{T}(\eta_{t-1}^{-1} - \eta_t^{-1})\Psi(x_t) \\
&\quad - \sup_{x\in\Delta_{[K]}}\left\{\langle x, -\hat{L}_T\rangle - \eta_T^{-1}\Psi(x)\right\} + \langle y, -\hat{L}_T\rangle - \eta_T^{-1}\Psi(y) + \eta_T^{-1}\Psi(y) \\
&\leq \sum_{t=1}^{T}\eta_t^{-1}D_{\Psi}(x_t, z_t) + \sum_{t=1}^{T}(\eta_{t-1}^{-1} - \eta_t^{-1})\Psi(x_t) + \eta_T^{-1}\Psi(y) \\
&= \sum_{t=1}^{T}\eta_t^{-1}D_{\Psi}(x_t, z_t) + \sum_{t=1}^{T}(\eta_t^{-1} - \eta_{t-1}^{-1})(\Psi(y) - \Psi(x_t))
\end{aligned}
$$

where step (a) is due to the Pythagoras property of Bregman divergences, and in step (b) we just plugged in the definition of $\overline{\Psi}^*$ in Ψ. □