Data-dependent Bounds with T-Optimal Best-of-Both-Worlds Guarantees in Multi-Armed Bandits using Stability-Penalty Matching

Quan Nguyen MANHQUAN233@GMAIL.COM
University of Victoria[*]

Shinji Ito SHINJI@MIST.I.U-TOKYO.AC.JP
University of Tokyo and RIKEN AIP

Junpei Komiyama JUNPEI@KOMIYAMA.INFO
New York University and RIKEN AIP

Nishant A. Mehta NMEHTA@UVIC.CA
University of Victoria

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Existing data-dependent and best-of-both-worlds regret bounds for multi-armed bandits problems have limited adaptivity as they are either data-dependent but not best-of-both-worlds (BOBW), BOBW but not data-dependent or have sub-optimal $O(\sqrt{T \ln T})$ worst-case guarantee in the adversarial regime. To overcome these limitations, we propose real-time stability-penalty matching (SPM), a new method for obtaining regret bounds that are simultaneously data-dependent, best-of-both-worlds and T-optimal for multi-armed bandits problems. In particular, we show that real-time SPM obtains bounds with worst-case guarantees of order $O(\sqrt{T})$ in the adversarial regime and $O(\ln T)$ in the stochastic regime while simultaneously being adaptive to data-dependent quantities such as sparsity, variations, and small losses. Our results are obtained by extending the SPM technique for tuning the learning rates in the follow-the-regularized-leader (FTRL) framework, which further indicates that the combination of SPM and FTRL is a promising approach for proving new adaptive bounds in online learning problems.

Keywords: multi-armed bandits, adaptive bounds, best-of-both-worlds, stability-penalty matching

1. Introduction

The multi-armed bandits problem (Lai and Robbins, 1985; Auer et al., 2002a) is one of the most fundamental frameworks for modeling sequential decision making problems under limited feedback. In this problem, a learner sequentially interacts with the environment in T rounds. In round $t = 1, 2, \ldots$, the learner chooses an action I_t from a set of K available actions and observes a numerical feedback $\ell_{t,I_t} \in \mathbb{R}$. This ℓ_{t,I_t} is an element of a hidden vector $\ell_t \in \mathbb{R}^K$ chosen at the beginning of round t by an oblivious adversary. The performance of the learner is its *pseudo-regret*

$$R_T = \max_{a \in [K]} R_{T,a} = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \ell_{t,a}\right], \tag{1}$$

where $\mathbb{E}$ denote the expectation taken over all randomness from all T rounds. Existing works have constructed algorithms with *worst-case* regret bounds that hold under the assumption on

[*] The majority of this work was done when Quan Nguyen was at RIKEN AIP.

whether the adversary is adversarial (i.e., $(\ell_t)_t$ are arbitrary) or stochastic (i.e., $(\ell_t)_t$ are drawn i.i.d. from some distribution) (Lai and Robbins, 1985; Auer et al., 2002a,b), *best-of-both-worlds* (BOBW) bounds that have worst-case guarantees simultaneously for adversarial and stochastic adversaries (e.g. Bubeck and Slivkins, 2012; Zimmert and Seldin, 2021; Dann et al., 2023; Ito et al., 2024), or *data-dependent* bounds that are adaptive to the sequence $(\ell_t)_t$ (e.g. Wei and Luo, 2018; Bubeck et al., 2018; Ito, 2021; Ito et al., 2022; Tsuchiya et al., 2023). Despite this vast amount of literature on different types of worst-case and adaptive bounds for multi-armed bandits, we are not aware of any works that establish bounds that are *simultaneously* data-dependent, best-of-both-worlds *and* have optimal dependency on T. In particular, existing works suffer from at least one of three limitations: being data-dependent but not BOBW (Hazan and Kale, 2011; Bubeck et al., 2018; Wei and Luo, 2018), being BOBW but not data-dependent (Zimmert and Seldin, 2021; Dann et al., 2023) or having sub-optimal dependency on T (Hazan and Kale, 2011; Wei and Luo, 2018; Tsuchiya et al., 2023; Ito et al., 2024). In this work, we close this gap in the literature by introducing novel algorithms with regret bounds that are simultaneously BOBW, data-dependent and T-optimal.

All of our algorithms are established in the Follow-the-Regularized-Leader (FTRL) framework (see e.g. Lattimore and Szepesvári, 2020), in which the time-varying learning rates are tuned by the Stability-Penalty Matching (SPM) method. SPM was originally proposed by Ito et al. (2024) as a principled method for tuning learning rates in FTRL using both the *penalty* and *stability* terms. More specifically, in round t, our algorithms compute a probability vector

$$q_t = \operatorname*{arg\,min}_{x \in \Delta_K} \langle L_{t-1}, x \rangle + \phi_t(x), \tag{2}$$

where $\Delta_K = \{x \in \mathbb{R}_+^K : \sum_{i=1}^K x_i = 1\}$ denotes the K-dimensional simplex, $L_{t-1} \in \mathbb{R}^K$ is the estimated cumulative loss vector up to round $t-1$ and $\phi_t(x) : \Delta_K \to \mathbb{R}$ is the regularization function. We use the following specific form for ϕ_t:

$$\phi_t(x) = \beta_t f(x) + \gamma u(x), \tag{3}$$

where $f(x) : \Delta_K \to \mathbb{R}_-, u(x) : \Delta_K \to \mathbb{R}_+$ are convex, $\beta_t > 0$ is the learning rate in round t and γ is a constant. Then, the learner draws an arm $I_t \sim q_t$ according to q_t (or some $p_t \in \Delta_K$ derived from q_t) and computes an estimated loss vector $\hat{\ell}_t$. Let $D_t(x, y) = \phi_t(x) - \phi_t(y) - \langle \nabla \phi_t(y), x - y \rangle$ denote the Bregman divergence associated with ϕ_t. The standard analysis of FTRL (e.g. Lattimore and Szepesvári, 2020, Exercise 28.12) implies that

$$\begin{aligned} R_{T,a} &\lesssim \phi_{T+1}(e_a) - \phi_1(q_1) + \mathbb{E}\left[\sum_{t=1}^T (\phi_t(q_{t+1}) - \phi_{t+1}(q_{t+1})) + \sum_{t=1}^T (\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t))\right] \\ &\lesssim \gamma u(e_a) - \beta_1 f(q_1) + \underbrace{\mathbb{E}\left[\sum_{t=1}^T (\beta_{t+1} - \beta_t) h_{t+1}\right]}_{\text{penalty term}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{\beta_t}\right]}_{\text{stability term}} \end{aligned}$$

where e_a is the a-th vector in the standard basis of $\mathbb{R}^K$, h_{t+1} satisfies $(-f(q_{t+1})) \lesssim h_{t+1}$ and z_t satisfies $\beta_t \mathbb{E}[\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t)] \lesssim \mathbb{E}[z_t]$. SPM carefully chooses β_1, z_t and h_t so that $h_{t+1} \leq O(h_t)$ and sets the learning rate of the next round to be

$$\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}. \tag{4}$$

This makes $(\beta_{t+1}-\beta_t)h_{t+1}$ match with $\frac{z_t}{\beta_t}$ and implies $R_{T,a} \lesssim \gamma u(e_a) - \beta_1 f(q_1) + \mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{\beta_t}\right]$. An important insight in SPM is that by picking $f(x)$ and $u(x)$ appropriately, $\mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{\beta_t}\right]$ is naturally adaptive to the adversarial or stochastic nature of the environment (Ito et al., 2024). In our work, we will show that SPM can be made adaptive not only to the nature of the environment but also to the underlying structure of the sequence of losses such as sparsity and total variation.

1.1. Main Contributions and Techniques

Throughout the paper, we will write $O(\square \ln(T), \square\sqrt{T})$ to denote a BOBW bound that holds for stochastic and adversarial regimes, respectively, where $\square$ contains problem-dependent terms. The original SPM method (Ito et al., 2024) used $z_t = \Omega(\beta_t \mathbb{E}_{I_t}[\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t)])$, where $\mathbb{E}_{I_t}$ denotes an expectation taken over I_t. Because only one out of K arms is observable in each round t, this in-expectation form of z_t inevitably requires taking the trivial bounds (e.g. 1) of the losses into its computation, thus limiting its adaptivity to $(\ell_t)_t$. Our work overcomes this limitation by setting

$$z_t = \Omega(\beta_t(\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t))). \tag{5}$$

We call this *real-time SPM*, since z_t depends on the observed arm I_t. The main technical challenge is now z_t can be very large since it grows with $\mathrm{poly}(\frac{1}{p_{t,I_t}})$. At the same time, we need to limit the amount of explicit exploration to obtain a BOBW bound for stochastic bandits. Table 1 summarizes our main results, showing that real-time SPM can be controlled effectively to give BOBW *and* data-dependent bounds with optimal dependency on T. Our results also hold for the more general adversarial regime with self-bounding constraint setting (Zimmert and Seldin, 2021). Appendix A gives a more detailed discussion on related works. Our paper is organized as follows:

- Section 2 introduces the real-time SPM method and states Lemma 2, a key technical lemma for bounding $\mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{\beta_t}\right]$. While the original analysis of SPM (Ito et al., 2024) relies on having a small $\max_{t\in[T]} z_t$ and thus cannot be applied to real-time SPM, our Lemma 2 instead shows that real-time SPM incurs an additional regret of at most $O\left(\max_{t\in[T]} \frac{z_t}{\beta_t} \ln \sum_{t=1}^T \frac{z_t}{h_t}\right)$. Moreover, both $\frac{z_t}{\beta_t}$ and $\frac{z_t}{h_t}$ can be effectively controlled by appropriate choices of $\phi_t(x)$.
- Section 3 considers the bandits problems with signed sparse losses, where $\ell_{t,i} \in [-1, 1]$ and $\|\ell_t\|_0 \le S$. We show that using α-Tsallis entropy and log-barrier functions in place of f and g in (3) leads to an $O\left(\frac{(K^{1-\alpha}-1)S^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}}, \left(\sqrt{\frac{(K^{1-\alpha}-1)S^\alpha T}{\alpha(1-\alpha)}}\right)\right)$. This bound is T-optimal and improves upon the best known bound for this setting established by Tsuchiya et al. (2023). When S is known, we show that the adversarial bound is improved to $O(\sqrt{ST\ln(K/S)})$, resolving an open question in Kwon and Perchet (2016). Furthermore, we prove a near-matching lower bound for problems in which the sparsity constraint holds in expectation.
- Section 4 considers problems with small total variation Q (defined in Section 1.2) and presents a new algorithm obtaining a $O\left(\frac{(K-1)^{1-\alpha}K^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}}, \sqrt{Q\ln(K)}\right)$ BOBW bound. In the adversarial regime, the $O(\sqrt{Q\ln(K)})$ bound matches the best known bound in Bubeck et al. (2018) while having the advantage of not requiring knowledge of Q.

- Section 5 introduces a new SPM method called coordinate-wise SPM, which maintain arm-dependent learning rates $\beta_{t,i}$ and performs real-time SPM on each arm separately. We show that coordinate-wise SPM achieves a BOBW bound with order $O\left(\frac{1}{\alpha(1-\alpha)}\sum_{i\neq i^*}\frac{\ln(T)}{\Delta_i}\right)$ in stochastic bandits and $O\left(\min\left\{\sqrt{K\ln(T)\min(Q_\infty, L^*, T-L^*)}, K^{\frac{\alpha}{2}}\sqrt{KT}\right\}\right)$ in adversarial bandits, where Q_∞ and L^* are ℓ_∞-norm total variation and total loss of the best arm, respectively (see Section 1.2 for their formal definitions).

Table 1: Summary of data-dependent results in existing and ours works. The three blocks of rows show bounds dependent on sparsity S, total variation Q and a combination of Q_∞ and L^*, respectively (formal definitions are in Section 1.2). We use $H^*_\infty = \min(Q_\infty, L^*, T-L^*)$. "$T$-opt BOBW" denote whether a bound is BOBW and T-optimal. "Param-free" denote whether a bound requires knowledge of the data-dependent quantities.

Algorithms	Stochastic	Adversarial	T-opt BOBW?	Param-free?
Bubeck et al. (2018)	–	$\sqrt{ST\ln K}$	×	×
Tsuchiya et al. (2023)	$\frac{S\ln(T)\ln(KT)}{\Delta_{\min}}$	$\sqrt{ST\ln T\ln K}$	×	✓
Theorem 3	$\frac{S\ln T\ln K}{\Delta_{\min}}$	$\sqrt{ST\ln K}$	✓	✓
Hazan and Kale (2011)	–	$\sqrt{Q\ln T\ln K}$	×	✓
Bubeck et al. (2018)	–	$\sqrt{Q\ln K}$	×	×
Theorem 7	$\frac{K\ln T}{\Delta_{\min}}$	$\sqrt{Q\ln K}$	✓	✓
Wei and Luo (2018)	$\frac{K\ln T}{\Delta_{\min}}$	$\sqrt{KL^*\ln T}$	×	✓
Ito (2021)	$\sum_{i\neq i^*}\frac{\ln T}{\Delta_i}$	$\sqrt{K\min(Q_\infty, L^*)\ln T}$	×	✓
Ito et al. (2022)	$\sum_{i\neq i^*}(\frac{\sigma_i^2}{\Delta_i}+1)\ln T$	$\sqrt{KH^*_\infty\ln T}$	×	✓
Theorem 9	$\sum_{i\neq i^*}\frac{\ln T}{\Delta_i}$	$\min(\sqrt{KH^*_\infty\ln T}, \sqrt{K^{1+\alpha}T})$	✓	✓

1.2. Problem Setup

For an integer N, let $[N] = \{1, 2, \ldots, N\}$ denote the set of integers from 1 to N. We study the multi-armed bandits problem (Lai and Robbins, 1985; Auer et al., 2002a) in which a learner is given K arms and interacts with the environment in T rounds. In each round t, an adversary selects a hidden vector $\ell_t = (\ell_{t,1}, \ell_{t,2}, \ldots, \ell_{t,K})^\top$. The learner chooses one arm $I_t \in [K]$ and observes its loss ℓ_{t,I_t}. We assume $\left|\ell_{t,i}\right| \leq 1$ for all $t \in [T], i \in [K]$. The learner aims to minimize its regret R_T over T rounds, defined by Equation (1).

We are interested in developing learning algorithms with provable upper bounds on R_T that hold simultaneously for two regimes: adversarial (Auer et al., 2002a) and adversarial with a (Δ, C, T) self-bounding constraint (Zimmert and Seldin, 2021). In the *adversarial regime*, no assumption is made on how the adversary generates $(\ell_t)_{t\in[T]}$. The adversarial regime with a (Δ, C, T) self-bounding constraint (Zimmert and Seldin, 2021) is given below.

Definition 1 *(Adversarial regime with a self-bounding constraint) For $T \geq 1, \Delta \in [0,1]^K$ and $C \geq 0$, the problem is in adversarial regime with a (Δ, C, T) self-bounding constraint if the regret of any algorithm at time T satisfies $R_T \geq \sum_{t=1}^T \sum_{i=1}^K \Delta_i \mathbb{P}(I_t = i) - C$.*

As noted in Zimmert and Seldin (2021), the stochastic bandits setting (Lai and Robbins, 1985) satisfies Definition 1. We also use the common assumption that there exists an optimal arm i^* such that $\Delta_i > 0$ for all $i \neq i^*$, that is, the optimal arm is unique. Let $\Delta_{\min} = \min_{i \in [K]} \{\Delta_i : \Delta_i > 0\}$.

We focus on obtaining bounds that are adaptive not only to the adversary's regime but also to the data-dependent properties of the loss sequence $(\ell_t)_{t \in [T]}$. The following data-dependent quantities are considered in our work.

- **Sparsity of losses** (Kwon and Perchet, 2016). All loss vectors have at most $1 \leq S \leq K$ non-zero elements, i.e., $\|\ell_t\|_0 \leq S$, where S is unknown.
- **Variation of losses** (Hazan and Kale, 2011; Ito et al., 2022) The total variation of the sequence $(\ell_t)_t$ is $Q = \sum_{t=1}^T \left\| \ell_t - \frac{1}{T} \sum_{s=1}^T \ell_s \right\|_2^2$. The ℓ_∞-norm total variation is $Q_\infty = \min_{\bar{\ell} \in [0,1]^K} \sum_{t=1}^T \left\| \ell_t - \bar{\ell} \right\|_\infty^2$.
- **Best-arm loss**. For non-negative losses, we consider the cumulative loss of the best arm $L_* = \min_{i \in [K]} \sum_{t=1}^T \ell_{t,i}$.

2. Stability-Penalty Matching with Real-Time Stability Term

Let $\tilde{p} = \min(1-p, p)$ for $p \in [0,1]$. We use the notation $f \lesssim g$ to denote $f = O(g)$. To obtain data-dependent bounds using SPM, we use SPM where the stability term is a function of the *observed* loss, i.e., z_t satisfies Equation (5). Note that z_t grows with ℓ_{t,I_t}^2 and $\frac{1}{p_{t,I_t}}$. The benefit of this real-time z_t is that data-dependent quantities such as sparsity and total variation naturally come out of $\mathbb{E}[z_t]$. For example, in Algorithm 1 for bandits with sparse losses $\|\ell_t\|_0 \leq S$, we use $z_t = O(\tilde{p}_{t,I_t}^{2-\alpha} \frac{\ell_{t,I_t}^2}{p_{t,I_t}^2})$ for some $\alpha \in (0,1)$. It follows that $\mathbb{E}[z_t] = O(\sum_{i:\ell_{t,i} \neq 0} \tilde{p}_{t,i}^{1-\alpha} \ell_{t,i}^2) \leq O(S^\alpha)$, leading to the $O(\sqrt{ST \ln K})$ bound. The main challenge in using the real-time z_t is the value of z_t can be unbounded whenever p_{t,I_t} is very small. It follows that $z_{\max} = \max_{t \in [T]} z_t$ can be unbounded, which makes it difficult to apply existing techniques in Ito et al. (2024, Lemma 10) that bounds $\mathbb{E}[\sum_{t=1}^T \frac{z_t}{\beta_t}]$ by a quantity that grows with $\mathbb{E}[z_{\max}]$. We resolve this challenge by using the following technical lemma.

Lemma 2 *For any $T \geq 1, z_{1:T} \geq 0, h_{1:T} > 0$ and a sequence $\beta_{1:T}$ defined by Equation (4), let $F(z_{1:T}, h_{1:T}) = \sum_{t=1}^T \frac{z_t}{\beta_t}$ and $G(z_{1:T}, h_{1:T}) = \sum_{t=1}^T \frac{z_t}{\sqrt{\sum_{s=1}^t \frac{z_s}{h_s}}}$. We have*

$$F(z_{1:T}, h_{1:T}) \lesssim G(z_{1:T}, h_{1:T}) + \left(\max_{t \in [T]} \frac{z_t}{\beta_t} \right) \ln \left(\sum_{t=1}^T \frac{z_t}{h_t} \right). \tag{6}$$

Proof (Sketch) Our proof extends from the proof of Ito et al. (2024, Lemma 10). Similar to the proof of Ito et al. (2024, Lemma 10), we define a new sequence $\beta'_t = \sqrt{\beta_1^2 + 2 \sum_{s=1}^{t-1} \frac{z_s}{h_s}}$ and consider the

set of rounds $E = \{t \in [T] : \beta'_{t+1} \geq \sqrt{2}\beta'_t\}$. The complement of E is $E^c = [T] \setminus E$. We have

$$F(z_{1:T}, h_{1:T}) = \underbrace{\sum_{t \in E^c} \frac{z_t}{\beta_t}}_{(a)} + \underbrace{\sum_{t \in E} \frac{z_t}{\beta_t}}_{(b)},$$

where (a) is bounded by $G(z_{1:T}, h_{1:T})$ as in Ito et al. (2024, Lemma 10), and (b) is bounded by

$$(b) \leq \left(\max_{t \in [T]} \frac{z_t}{\beta_t}\right) |E| \leq \left(\max_{t \in [T]} \frac{z_t}{\beta_t}\right) \log_{\sqrt{2}} \left(\frac{\beta'_{T+1}}{\beta'_1}\right) \lesssim \left(\max_{t \in [T]} \frac{z_t}{\beta_t}\right) \ln \left(\sum_{t=1}^{T} \frac{z_t}{h_t}\right),$$

where the second inequality is from the fact that β'_t is multiplied by at least $\sqrt{2}$ after every round in E; thus there can be at most $\log_{\sqrt{2}} \frac{\beta'_{T+1}}{\beta'_1}$ such multiplications. ■

Lemma 2 implies that if (I) the sum $\mathbb{E}[\sum_{t=1}^{T} \frac{z_t}{h_t}]$ grows with $\mathrm{poly}(T)$ and (II) $\max_t \frac{z_t}{\beta_t}$ is small, then $\mathbb{E}[F(z_{1:T}, h_{1:T})]$ grows dominantly with $\mathbb{E}[G(z_{1:T}, h_{1:T})]$ plus an $O(\ln(T))$ term. Hence, we can safely ignore other terms and focus only on bounding $G(z_{1:T}, h_{1:T})$. The proof of Ito et al. (2024, Lemma 10) already showed that

$$G(z_{1:T}, h_{1:T}) \lesssim \min \left\{ \sqrt{\ln(T) \sum_{t=1}^{T} h_t z_t} + \sqrt{\frac{1}{T} h_{\max} \sum_{t=1}^{T} z_t}, \sqrt{h_{\max} \sum_{t=1}^{T} z_t} \right\}. \tag{7}$$

In the rest of the paper, we will show that different choices of the (hybrid) regularization function lead to specific forms of z_t and h_t such that not only do both conditions (I) and (II) hold but also they imply BOBW data-dependent bounds with optimal dependency on T from (7).

3. Application I: BOBW Bounds for Bandits with Sparse Losses

We consider the multi-armed bandits setting with sparse losses (Kwon and Perchet, 2016), in which the loss vector $\ell_t \in [-1, 1]^K$ has at most S non-zero elements, i.e., $\max_{t \in [T]} \|\ell_t\|_0 \leq S$. Note that S is unknown to the learner. Let $\psi_{TE}(p) = \frac{1}{\alpha}(1 - \sum_{i=1}^{K} p_i^\alpha)$ be the α-Tsallis entropy with some $\alpha \in (0, 1)$ and $\psi_{LB}(p) = -\sum_{i=1}^{K} \ln(p_i)$ be the log-barrier function. Our approach for this setting is in Algorithm 1, in which we use the hybrid regularizer $\phi_t(p) = \beta_t \psi_{TE}(p) + \gamma \psi_{LB}(p)$ to obtain

$$q_t = \arg\min_{p \in \Delta_K} \{\langle L_{t-1}, p \rangle + \beta_t \psi_{TE}(p) + \gamma \psi_{LB}(p)\},$$

Then, we mix q_t with $\frac{1}{T}$-uniform exploration to obtain the sampling probability p_t, i.e., $p_t = \left(1 - \frac{K}{T}\right) q_t + \frac{1}{T}\mathbf{1}$. The learning rates $(\beta_t)_t$ are set by the SPM rule by Ito et al. (2024) as

$$\beta_1 = \frac{8K}{1 - \alpha}, \ \beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}, \tag{8}$$

where

$$z_t = \min \left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min(p_{t,I_t}, 1 - p_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2, \beta_t \frac{18d^2}{\gamma} \ell_{t,I_t}^2 \right), \ h_t = (-\psi_{TE}(p_t)), \tag{9}$$

Input: $K \geq 3, T \geq 4K, \alpha \in (0,1), \beta_1 = \frac{8K}{1-\alpha}, \gamma = \max(6, 48\sqrt{\frac{\alpha}{1-\alpha}}), d = 2.$
Initialize $L_{0,i} = 0$ for $i \in [K]$
for *each round* $t = 1, \ldots, T$ **do**
  Compute $q_t = \arg\min_{p \in \Delta_K} \langle L_{t-1}, p \rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha) \right) - \gamma \sum_{i=1}^K \ln(p_i)$
  Compute $p_t = \left(1 - \frac{K}{T}\right) q_t + \frac{1}{T}\mathbf{1}$
  Draw $I_t \sim p_t$ and observe ℓ_{t,I_t}
  Compute loss estimate $\hat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{1}\{I_t = i\}}{p_{t,i}}$ and update $L_{t,i} = L_{t-1,i} + \hat{\ell}_{t,i}$
  Compute $z_t = \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min(p_{t,I_t}, 1 - p_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{\beta_t 18 d^2}{\gamma} \ell_{t,I_t}^2 \right)$
  Compute $h_t = \left(\frac{1}{\alpha}(\sum_{i=1}^K p_{t,i}^\alpha - 1) \right)$
  Compute $\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}$
end

Algorithm 1: Real-time SPM with hybrid regularization for losses in $[-1, 1]$.

and $\gamma = \max(6, 48\sqrt{\frac{\alpha}{1-\alpha}}), d = 2$. Note that $\beta_1 \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ for $\omega = 2$. The following theorem states the BOBW bounds of Algorithm 1.

Theorem 3 *For any $K \geq 4, T \geq 4k$, Algorithm 1 guarantees the following bounds simultaneously*

- *In the adversarial regime:*

$$R_T \leq O\left(\sqrt{\frac{(K^{1-\alpha} - 1) S^\alpha T}{\alpha(1-\alpha)}} \right) \tag{10}$$

- *In the adversarial regime with a self-bounding constraint:*

$$R_T \leq O\left(\frac{(K-1)^{1-\alpha} S^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}} + \sqrt{C \frac{(K-1)^{1-\alpha} S^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}}} + \sqrt{\frac{(K-1)^{1-\alpha} S^\alpha}{\alpha(1-\alpha)}} \right) \tag{11}$$

In Appendix G, we show that by setting $\alpha = 1 - \frac{1}{2\ln(K)}$, we obtain $\frac{(K^{1-\alpha}-1)S^\alpha}{\alpha(1-\alpha)} \lesssim S^\alpha \ln(K)$ and $\frac{(K-1)^{1-\alpha}S^\alpha}{\alpha(1-\alpha)} \lesssim S^\alpha \ln(K)$. In the adversarial regime, Theorem 3 recovers the $O(\sqrt{ST\ln(K)})$ bound in Bubeck et al. (2018) and Tsuchiya et al. (2023) while still being S-agnostic. In the adversarial regime with a self-bounding constraint, the bound becomes $O(\frac{S\ln(K)\ln(T)}{\Delta_{\min}})$ which has an optimal dependency on T. To the best of our knowledge, Theorem 3 is the first result for bandits with sparse signed losses that is simultaneously S-agnostic, T-optimal and BOBW. Also, our approach is more computationally efficient than that of Tsuchiya et al. (2023) as we do not need to solve any additional optimization problems to compute the learning rates $(\beta_t)_t$.

Remark 4 *When S is known, then $\frac{(K^{1-\alpha}-1)S^\alpha}{\alpha(1-\alpha)}$ can be further bounded by $6S\ln(\frac{K}{S})$. Consider only the case where S is sufficiently small so that $e^2 S \leq K$ (the other direction trivially leads to*

$O(\frac{S}{\alpha(1-\alpha)})$. *Letting* $\alpha = 1 - \frac{1}{\ln(K/S)}$, *then* $(\frac{K-1}{S})^{1-\alpha} \leq (\frac{K}{S})^{1-\alpha} = e$. *Since* $\ln(K/S) \geq 2$, *we have* $\alpha \geq \frac{1}{2}$. *Therefore,*

$$\frac{(K^{1-\alpha}-1)S^\alpha}{\alpha(1-\alpha)} \leq \frac{K^{1-\alpha}S^\alpha}{\alpha(1-\alpha)} = S\left(\frac{K}{S}\right)^{1-\alpha}\frac{\ln(K/S)}{\alpha} = \frac{eS\ln(K/S)}{\alpha} \leq 6S\ln(K/S).$$

This result shows that an $O(\sqrt{ST\ln(K/S)})$ *upper bound is attainable even for signed losses, which resolves an open question posed in (Kwon and Perchet, 2016, Remark 12).*

Remark 5 *In Appendix F, we also show that Algorithm 1 can be applied in the related setting of adversarial sleeping bandits and obtain a regret bound that matches the best known bound in Nguyen and Mehta (2024) despite using fewer assumptions.*

3.1. Proof Sketch for Theorem 3

As mentioned in Section 2, we first show that z_t and h_t in (9) satisfy the two conditions (I) $\mathbb{E}[\sum_{t=1}^T \frac{z_t}{h_t}] = O(\mathrm{poly}(T))$ and (II) $\max_t \frac{z_t}{\beta_t}$ is small. The second condition is straightforward from the definition of z_t, γ and d, since $\frac{z_t}{\beta_t} \leq \frac{18d^2}{\gamma} \leq 6d^2 = 24$ which is a constant. To see that (I) is true, note that h_t is fixed with respect to I_t. Hence,

$$\mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{h_t}\right] = \mathbb{E}\left[\sum_{t=1}^T \frac{\mathbb{E}_{I_t}[z_t]}{h_t}\right] \leq T\mathbb{E}\left[\frac{\max_{t\in[T]}\mathbb{E}_{I_t}[z_t]}{\min_{t\in[T]} h_t}\right].$$

Then, the condition (I) follows from Lemma 11, which shows that $h_t \geq \frac{1-\alpha}{4\alpha}T^{-\alpha}$ and $E_{I_t}[z_t] \leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)}S^\alpha$. Jensen's inequality implies that $\mathbb{E}\left[\ln\left(\sum_{t=1}^T \frac{z_t}{h_t}\right)\right] \leq \ln\left(\mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{h_t}\right]\right)$. Combining this with Lemma 11, we conclude that $\mathbb{E}[F(z_{1:T}, h_{1:T})]$ grows dominantly with $\mathbb{E}[G(z_{1:T}, h_{1:T})]$. The last part of the proof is showing that plugging z_t and h_t from (9) into (7) yields the desired bounds. In the adversarial regime, the bound (10) follows directly from (7), Lemma 11, $h_t \leq \frac{K^{1-\alpha}-1}{\alpha}$ and Jensen's inequality $\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$. In the adversarial regime with a self-bounding constraint, we can prove (11) by first showing that

$$\mathbb{E}[h_t z_t] \leq \frac{(6d)^{2-\alpha}}{\alpha(1-\alpha)\Delta_{\min}}(K-1)^{1-\alpha}S^\alpha\mathbb{E}\left[\sum_{i=1}^K p_{t,i}\Delta_i\right].$$

and then following the same argument as in Ito et al. (2024).

3.2. A Lower Bound for Problems with Soft Sparsity Constraint

It remains an open question whether the BOBW bounds in Theorem 3 are tight under the hard constraint $\|\ell_t\|_0 \leq S$. This hard-constraint problem belongs to a broader class of settings with a more relaxed constraint, in which there exists an $\alpha \in (0,1)$ and $1 \leq U \leq K^\alpha$ such that for all $t \in [T]$,

$$\mathbb{E}\left[\left(\sum_{i=1}^K |\ell_{t,i}|^{2/\alpha}\right)^\alpha\right] \leq U. \tag{12}$$

In other words, the sparsity constraint holds in expectation. Obviously, the hard-constraint setting with $\|\ell_t\|_0 \leq S$ satisfies (12) for any $\alpha \in (0,1)$ and $U = S^\alpha$. Moreover, by using the same Algorithm 1 and straightforward modifications in its proof, we can obtain the corresponding $O(\frac{K^{1-\alpha}U}{\alpha(1-\alpha)\Delta_{\min}}\ln T)$ and $O(\sqrt{\frac{K^{1-\alpha}}{\alpha(1-\alpha)}UT})$ BOBW bounds for stochastic and adversarial regimes, respectively. The following theorem, whose proof is in Section C, shows near-matching lower bounds for problems with soft sparsity constraint defined in (12).

Theorem 6 *(Instance-Dependent Lower Bound) For any consistent algorithm, for any $\Delta \in (0,1), K \geq 4, \alpha \in (0,1)$ and $1 \leq U \leq \frac{K^\alpha}{4}$, there exists a K-armed stochastic bandit instance with $\Delta_{\min} = \Delta$ and loss distribution satisfying (12) such that*

$$\lim_{T\to\infty} \frac{R_T}{\ln(T)} = \Omega\left(\frac{K^{1-\alpha}U}{\Delta}\right).$$

(Minimax Lower Bound) For any algorithm, for any $K \geq 4, \alpha \in (0,1)$ and $U \leq K^\alpha$, there exists an adversarial bandit instance with K arms and loss distribution satisfying (12) such that

$$R_T = \Omega(\sqrt{K^{1-\alpha}UT}).$$

4. Application II: $\sqrt{Q\ln(K)}$ Upper Bound with Unknown Q using Optimistic FTRL

In this section, we propose a new approach for obtaining a BOBW $O(\frac{K\ln T}{\Delta_{\min}}, \sqrt{Q\ln K})$-bound with unknown Q. For ease of exposition, we assume losses are in $[0,1]$ and note that the analysis can be easily extended for losses in $[-1,1]$. The new approach is based on applying real-time SPM on the Optimistic FTRL framework (Rakhlin and Sridharan, 2013), and then combining with the Reservoir Sampling algorithm (Hazan and Kale, 2011). In principle, our algorithm follows the same framework as Hazan and Kale (2011); Bubeck et al. (2018) where the learner maintains a reservoir $\mathbb{S}_i$ of observed losses for each arm $i \in [K]$ and then uses the estimated mean $m_{t,i} = \tilde{\mu}_{t,i}$ of these reservoirs as the optimistic vector m_t in Optimistic FTRL. In each round t, the learner chooses to perform either a reservoir sampling step for updating the reservoir $\mathbb{S}_i$, or a FTRL learning step for minimizing the regret. When the FTRL learning step is performed in round t, the vector q_t is computed by

$$q_t = \arg\min_{x\in\Delta_K} \langle m_t + L_{t-1}, x\rangle + \beta_t\left(\frac{1}{\alpha}(1-\sum_{i=1}^K x_i^\alpha)\right) - \gamma\sum_{i=1}^K \ln(x_i). \tag{13}$$

Similar to Algorithm 1, the sampling probability vector p_t is obtained by mixing with $\frac{1}{T}$, i.e, $p_t = \left(1-\frac{K}{T}\right)q_t + \frac{1}{T}\mathbf{1}$. After an arm $I_t \sim p_t$ is drawn, the loss estimates are $\hat{\ell}_{t,i} = m_{t,i} + \frac{(\ell_{t,i}-m_{t,i})\mathbb{1}\{I_t=i\}}{p_{t,i}}$. The learning rates $(\beta_t)_t$ are computed by real-time SPM, with z_t and h_t defined as

$$z_t = \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)}\tilde{p}_{t,I_t}^{2-\alpha}(\hat{\ell}_{t,I_t}-m_{t,I_t})^2, \frac{\beta_t 18d^2}{\gamma}(\ell_{t,I_t}-m_{t,I_t})^2\right), \quad h_t = \frac{1}{\alpha}\left(\sum_{i=1}^K p_{t,i}^\alpha - 1\right).$$

The full procedure is given in Algorithm 4 in Appendix D. The following theorem states the BOBW bound for this approach.

Theorem 7 *Algorithm 4 (in Appendix D) guarantees the following bounds simultaneously*

- *In the adversarial regime:*

$$R_T \leq O\left(\sqrt{\frac{(K^{1-\alpha}-1)Q}{\alpha(1-\alpha)}}\right). \tag{14}$$

- *In the adversarial regime with a self-bounding constraint:*

$$R_T \leq O\left(\frac{(K-1)^{1-\alpha}K^{\alpha}\ln(T)}{\alpha(1-\alpha)\Delta_{\min}} + \sqrt{C\frac{(K-1)^{1-\alpha}K^{\alpha}\ln(T)}{\alpha(1-\alpha)\Delta_{\min}}} + \sqrt{\frac{(K-1)^{1-\alpha}K^{\alpha}}{\alpha(1-\alpha)}}\right) \tag{15}$$

Proof (Sketch) Our analysis follows from the analysis of Algorithm 1 and the observation by Hazan and Kale (2011) that the reservoir sampling steps only add an $O(\ln(T)^2)$ amount to the regret bound. As a result, the bound (15) for adversarial regime with a self-bounding constraint follows almost identically to that of Algorithm 1. For the bound (14) in the adversarial regime, the total variation Q naturally comes out of $\sum_{t=1}^{T} z_t$ as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} z_t\right] &\lesssim \frac{1}{1-\alpha}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\ell_{t,i}-m_{t,i})^2\right] = \frac{1}{1-\alpha}\mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t-\tilde{\mu}_t\|_2^2\right] \qquad (\text{since } m_t=\tilde{\mu}_t)\\
&\leq \frac{1}{1-\alpha}\left(\mathbb{E}\left[\sum_{t=1}^{T}\|\ell_t-\mu_T\|_2^2\right] + \mathbb{E}\left[\sum_{t=1}^{T}\|\tilde{\mu}_t-\mu_t\|_2^2\right]\right)\\
&\leq \frac{1}{1-\alpha}\left(Q + \sum_{t=1}^{T}\frac{Q}{t\ln(T)}\right) \leq O\left(\frac{Q}{1-\alpha}\right),
\end{aligned}$$

where the second inequality follows from triangle inequality and Lemma 10 in Hazan and Kale (2011), the third inequality is by Lemma 11 in Hazan and Kale (2011), and the last inequality is due to $\sum_{t=1}^{T}\frac{1}{t\ln T} \leq O(1)$. Together with (7) and $h_{\max} \leq \frac{K^{1-\alpha}-1}{\alpha}$, this implies (14). ■

Remark 8 *While existing works (Hazan and Kale, 2011; Bubeck et al., 2018) require either the knowledge of Q or sophisticated doubling tricks to estimate Q, our Algorithm 4 does not require such knowledge or any tricks. When $\alpha \to 1$, the bound in (14) becomes $O(\sqrt{Q\ln(K)})$. This bound matches the best known upper bound in Bubeck et al. (2018) and never exceeds $O(\sqrt{TK\ln(K)})$ in the worst case, all while simultaneously having a T-optimal best-of-both-worlds guarantee.*

5. Coordinate-Wise Stability-Penalty Matching

We further generalize the SPM framework by introducing a new technique called coordinate-wise SPM (`CoWSPM`). As the name suggests, `CoWSPM` maintains separate learning rate $\beta_{t,i}$, stability term

Input: $K \geq 3, T \geq 4K, \alpha \in (0,1), \beta_1 = \frac{8K}{1-\alpha}\mathbf{1}, \gamma = \max(6, 48\sqrt{\frac{\alpha}{1-\alpha}}), d = 2.$
Initialize $L_{0,i} = 0$ for $i \in [K]$
for *each round* $t = 1, \ldots, T$ **do**
  Compute $m_t \in [0,1]^K$ where $m_{t,i} = \frac{1}{1+\sum_{s=1}^{t-1} \mathbb{1}\{I_s = i\}} \left(\frac{1}{2} + \sum_{s=1}^{t-1} \mathbb{1}\{I_s = i\}\ell_{t,i}\right)$
  Compute q_t by Equation (13)
  Compute $p_t = (1 - \frac{K}{T})q_t + \frac{1}{T}\mathbf{1}$
  Draw $I_t \sim p_t$ and observe ℓ_{t,I_t}
  Compute loss estimate $\hat{\ell}_{t,i} = m_{t,i} + \frac{(\ell_{t,i} - m_{t,i})\mathbb{1}\{I_t = i\}}{p_{t,i}}$ and update $L_{t,i} = L_{t-1,i} + \hat{\ell}_{t,i}$
  Compute $z_{t,i} = \mathbb{1}\{i = I_t\}(\ell_{t,I_t} - m_{t,I_t})^2 \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min\left\{p_{t,I_t}^{-\alpha}, \frac{1-p_{t,I_t}}{p_{t,I_t}^2}\right\}, \frac{\beta_{t,i}18d^2}{\gamma}\right)$
  Compute $h_{t,i} = \frac{1}{\alpha}p_{t,i}^{\alpha}$
  Compute $\beta_{t+1,i} = \beta_{t,i} + \frac{z_{t,i}}{\beta_{t,i}h_{t,i}}$
end

Algorithm 2: Coordinate-wise SPM with hybrid regularization for losses in $[0,1]$.

$z_{t,i}$ and penalty term $h_{t,i}$ for each arm $i \in [K]$. In each round t, `CoWSPM` updates the learning rates for each arm using the SPM update formula (4), i.e.,

$$\beta_{t+1,i} = \beta_{t,i} + \frac{z_{t,i}}{\beta_{t,i}h_{t,i}}. \tag{16}$$

Obviously, if $(z_{t,i})_{i\in[K]}$ and $(h_{t,i})_{i\in[K]}$ take the same values across all arms, then this approach recovers Algorithm 1. Instead, we adopt a different approach where $z_{t,i} = 0$ for all $i \neq I_t$ so that only the learning rate β_{t,I_t} of the observed arm I_t is updated in round t. The full procedure of `CoWSPM` is given in Algorithm 2, which uses the Optimistic FTRL framework with

$$\phi_t(x) = \sum_{i=1}^{K} \beta_{t,i}\left(\frac{-x_i^{\alpha}}{\alpha} + (1-x_i)\ln(1-x_i) + x_i\right) - \gamma\sum_{i=1}^{K}\ln(x_i). \tag{17}$$

This regularization function contains not only the α-Tsallis entropy, but also a part of the Shannon entropy and a linear term. The addition of these terms into the regularizer has been done in Ito et al. (2022) in order to have a regret bound containing the quantity $\tilde{p}_{t,i} = \min(p_{t,i}, 1-p_{t,i})$ for the stochastic setting. This technique has a similar impact in our work, where it allows us to bound $z_{t,i}$ by a quantity containing $\tilde{p}_{t,i}$. However, while we use the same technique to introduce $\tilde{p}_{t,i}$ into our bounds, our analysis develops fundamentally different technical lemmas from that of Ito et al. (2022) in order to use this new regularizer in the real-time SPM framework. Next, to ensure that only β_{t,I_t} is updated, we set $z_{t,i}$ by

$$z_{t,i} = \mathbb{1}\{i = I_t\}(\ell_{t,I_t} - m_{t,I_t})^2 \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min\left\{p_{t,I_t}^{-\alpha}, \frac{1-p_{t,I_t}}{p_{t,I_t}^2}\right\}, \frac{\beta_{t,i}18d^2}{\gamma}\right), \tag{18}$$

so that $z_{t,I_t} \geq 0$ and $z_{t,i} = 0$ for $i \neq I_t$. The following theorem states the BOBW data-dependent bound of Algorithm 2, whose full proof is given in Appendix E. The proof sketch outlines the main technical challenges in the analysis of Algorithm 2.

Theorem 9 *For any $K \geq 4, T \geq 4k$, CoWSPM (Algorithm 2) with $\alpha \in (0,1)$ guarantees the following bounds simultaneously*

- *In the adversarial regime:*

$$R_T \lesssim \min\left\{\sqrt{K\ln(T)\min(Q_\infty, L^*, T-L^*)}, K^{\frac{\alpha}{2}}\sqrt{KT}\right\}.$$

- *In the stochastic regime:*

$$R_T \lesssim \frac{1}{\alpha(1-\alpha)}\sum_{i\neq i^*}\frac{\ln(T)}{\Delta_i}.$$

Proof (Sketch) Intuitively, coordinate-wise SPM consists of K separate real-time SPM processes, one for each arm. Similar to Ito et al. (2022), we find that this more refined approach enables deriving a bound (for the adversarial regime) that is adaptive to simultaneously different data-dependent quantities such as Q_∞ and L^*. However, having separate learning rates introduces several new technical challenges. First, the analysis developed for Algorithm 1 that bounds $q_{t+1,i} = O(q_{t,i})$ for all $i \in [K]$ no longer applies because in each round t, the learning rates $(\beta_{t,i})_{i\in[K]}$ can be arbitrarily different from each other. The CoWSPM algorithm resolves this by using $\beta_{t+1,i} = \beta_{t,i}$ for $i \neq I_T$ so that it only need $q_{t+1,i} = O(q_{t,i})$ to hold for $i = I_t$ since

$$\phi_t(q_{t+1}) - \phi_{t+1}(q_{t+1}) = \sum_{i=1}^{K}(\beta_{t+1,i}-\beta_{t,i})(-f(q_{t+1,i})) = (\beta_{t+1,I_t}-\beta_{t,I_t})f(q_{t+1,I_t}).$$

The second and also more important challenge is that even if $q_{t+1,i} = O(q_{t,i})$, the naive decomposition of the α-Tsallis entropy into its coordinate-wise form $-\psi_{TE}(x) = \frac{1}{\alpha}\sum_{i=1}^{K}(x_i^\alpha - x_i)$ and then assigning $h_{t,i} = \frac{1}{\alpha}(x_i^\alpha - x_i)$ does *not* guarantee that $h_{t+1,i} = O(h_{t,i})$. This is because the function $x \mapsto x^\alpha - x$ gets arbitrarily close to 0 when x gets close to 1. This prompts a different choice for $h_{t,i}$ rather than $-f(p_{t,i})$. Algorithm 2 uses $h_{t,i} = \frac{1}{\alpha}p_{t,i}^\alpha$, which is monotonically increasing and ensures that $h_{t+1,I_t} = O(h_{t,I_t})$ for $q_{t+1,i} = O(q_{t,i})$. This choice of $h_{t,i}$ is justified by the technical Lemma 30, which states that $(x-1)\ln(1-x) \leq x^\alpha$ for any $x, \alpha \in [0,1]$.

Finally, we prove that with $z_{t,i}$ defined in (18), the product $\mathbb{E}[h_{t,i}z_{t,i}]$ is upper bounded by a quantity containing $\tilde{p}_{t,i}$ and thus an $O(\sum_{i\neq i^*}\frac{\ln T}{\Delta_i})$ regret bound holds for stochastic bandits. This is handled by Lemma 34, which shows that $\mathbb{E}_{I_t}\left[z_{t,i}\right] \leq 2\min(p_{t,i}, 1-p_{t,i})$. ■

Remark 10 *Theorem 9 holds for all $\alpha \in (0,1)$. In particular, for $\alpha \neq \frac{1}{2}$, we do not require any additional assumptions such as the Δ_i being known in order to get the T-optimal BOBW bound. This is a major difference compared to the Tsallis-INF algorithm (Zimmert and Seldin, 2021). On the other hand, the adversarial bound in Theorem 9 has an extra factor $\sqrt{K^\alpha}$. It is unclear to us whether this extra factor is a fundamental limitation of CoWSPM or an artifact of our analysis.*

6. Conclusion and Future Works

We developed real-time SPM, an extension of the SPM method originally developed for obtaining best-of-both-worlds bounds in bandits problems. We showed that real-time SPM algorithms achieve novel bounds that are simultaneously best-of-both-worlds, data-dependent and have optimal dependency on T in both stochastic and adversarial regimes. Our bounds also have optimal dependency on the data-dependent quantities such as sparsity or total variation of the loss sequence without knowing them nor using sophisticated estimation tricks. Future work includes applying real-time SPM on other bandits problems, such as contextual linear bandits, and making real-time SPM adaptive towards other challenging data-dependent quantities like ℓ_1 and ℓ_2-norm path-length bounds.

Acknowledgments

QN is supported by RIKEN AIP OSC program and partially by an NSERC Discovery Grant. SI was supported by JSPS KAKENHI Grant Number JP25K03184. NM is partially supported by an NSERC Discovery Grant. We also thank Taira Tsuchiya and Junya Honda for helpful discussions.

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, May 2002a. ISSN 1573-0565. doi: 10.1023/A:1013689704352.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 42.1–42.23, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL https://proceedings.mlr.press/v23/bubeck12b.html.

Sébastien Bubeck, Michael Cohen, and Yuanzhi Li. Sparsity, variance and curvature in multi-armed bandits. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 111–127. PMLR, 07–09 Apr 2018. URL https://proceedings.mlr.press/v83/bubeck18a.html.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Chris Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5503–5570. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/dann23a.html.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(35):1287–1311, 2011. URL http://jmlr.org/papers/v12/hazan11a.html.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of Thirty Fourth Conference on Learning Theory*, 2021.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1421–1422. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/ito22a.html.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adaptive learning rate for follow-the-regularized-leader: Competitive analysis and best-of-both-worlds. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 2522–2563. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/ito24a.html.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine Learning*, 80(2–3):245–272, sep 2010. ISSN 0885-6125. doi: 10.1007/s10994-010-5178-7. URL https://doi.org/10.1007/s10994-010-5178-7.

Joon Kwon and Vianney Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *Journal of Machine Learning Research*, 17(227):1–32, 2016. URL http://jmlr.org/papers/v17/15-503.html.

T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Quan M Nguyen and Nishant Mehta. Near-optimal per-action regret bounds for sleeping bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR, 2024.

Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2023. URL http://arxiv.org/abs/1912.13213.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 993–1019, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL https://proceedings.mlr.press/v30/Rakhlin13.html.

Igal Sason. On reverse Pinsker inequalities. *arXiv preprint arXiv:1503.07118*, 2015.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=J3taqrzyyA.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1263–1291. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/wei18a.html.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

Appendix A. Related Works

Due to the vast literature on BOBW and data-dependent bounds in various bandits learning settings, this sections presents only the most relevant works in multi-armed bandits. A more comprehensive list of related works can be found in Ito et al. (2024); Tsuchiya et al. (2023) and references therein.

Best-of-both-worlds bounds. The BOBW bounds in our paper are derived using the SPM method for tuning learning rates in the FTRL framework, originally proposed in Ito et al. (2024). For stochastic bandits, our $O(\frac{K \ln T}{\Delta_{\min}})$bound in Sections 3 and 4 matches that of Wei and Luo (2018); Ito et al. (2024), and our $O(\sum_{i \neq i^*} \frac{\ln T}{\Delta_i})$ bound in Section 5 matches that of Zimmert and Seldin (2021); Ito (2021). Both of these bounds are looser than the $O(\sum_{i \neq i^*} \frac{\sigma_i^2 \ln T}{\Delta_i})$ in Ito et al. (2022) obtained by a more specialized approach, where σ_i^2 is the variance of the losses of a sub-optimal arm i. However, except for Ito et al. (2024), these existing works have an $O(\sqrt{T \ln T})$ worst-case bound for adversarial bandits, which contains an extra $\ln T$ factor compared to our work. Our BOBW bound also have data-dependent guarantees, which is an advantage over Ito et al. (2024). For bandits with sparse losses, Tsuchiya et al. (2023) similarly obtained bounds that are both BOBW and dependent on the sparsity constraint; however their bounds contain extra factors of $\ln(KT)$ in stochastic bandits and $\sqrt{\ln T}$ in adversarial bandits compared to our results.

Data-dependent bounds. We study the following data-dependent quantities: sparsity of losses, total variations and small losses. For bandits with sparse negative losses where $\|\ell_t\| \leq S$ and S is unknown, our $O(\frac{S \ln T}{\Delta_{\min}}, \sqrt{ST \ln(K)})$ BOBW bound is the first S-agnostic and T-optimal BOBW bound for this setting, which improves upon on the bound of Tsuchiya et al. (2023) and matches the best known bound for adversarial bandits in Bubeck et al. (2018). When the total variations Q, Q_∞ and/or the loss of the best arm L^* (defined in Section 1.2) are small, our algorithms are based on the optimistic FTRL (OFTRL) framework similar to Hazan and Kale (2011); Bubeck et al. (2018); Ito et al. (2022). The dependency on Q, Q_∞ and L^* in our results match the best known bounds in these works, while our BOBW bounds have an optimal $O(\square \ln T, \square \sqrt{T})$ dependency on T. Particularly, our coordinate-wise real-time SPM algorithm in Section 5 can be seen as a T-optimal variant of the algorithm in Ito et al. (2022), which share the idea of using separate learning rates for each arm.

Appendix B. Proofs for Section 3

B.1. Proof for Theorem 3

Let $D_{TE}(p, q)$ and $D_{LB}(p, q)$ denote the Bregman divergences induced by the α-Tsallis entropy and the log-barrier function, respectively. Let $D_t(p, q) = \beta_t D_{TE}(p, q) + \gamma D_{LB}(p, q)$ denote the Bregman divergence induced by the hybrid regularizer $\phi_t(p) = \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha)\right) - \gamma \sum_{i=1}^K \ln(p_i)$.

Let $\hat{\ell}_t = \begin{bmatrix} \hat{\ell}_{t,1} \\ \hat{\ell}_{t,2} \\ \cdots \\ \hat{\ell}_{t,K} \end{bmatrix}$ be the estimated loss vector at time t. We state the following three stability lemmas, whose proofs are in Section B.2 and Section B.3.

Lemma 11 *For any $t \in [T]$, Algorithm 1 guarantees*

$$h_t \geq \frac{1-\alpha}{4\alpha} T^{-\alpha} \quad \textit{and} \quad E_{I_t}[z_t] \leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} S^\alpha.$$

Lemma 12 *For any $t \in [T]$, Algorithm 1 guarantees*

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min \left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)} \min(p_{t,I_t}, 1-p_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma} \ell_{t,I_t}^2 \right). \tag{19}$$

Note that in Lemma 12, the right-hand side is exactly $\frac{z_t}{\beta_t}$.

Lemma 13 *For any $t \in [T]$, Algorithm 1 guarantees that for all $i \in [K]$,*

$$q_{t+1,i} \leq 3d q_{t,i} \leq 6d p_{t,i}. \tag{20}$$

Moreover, this implies that $(-\psi_{TE}(q_{t+1})) \leq 3d(-\psi_{TE}(q_t)) \leq 6d(-\psi_{TE}(p_t))$.

Proof (Of Theorem 3) Next, let

$$\Phi_t(p) = \beta_t \psi_{TE}(p) + \gamma \psi_{LB}(p) \tag{21}$$

be the time-varying regularizer in Algorithm 1. For any $a \in [K]$, define

$$u_a = \left(1 - \frac{K}{T}\right) e_a + \frac{1}{T}\mathbf{1}.$$

The pseudo-regret with respect to arm a is

$$\begin{aligned}
R_{T,a} &= \mathbb{E}[\sum_{t=1}^{T} \langle \ell_t, p_t - e_a \rangle] \\
&= \mathbb{E}[\sum_{t=1}^{T} \langle \ell_t, q_t - u_a \rangle] + \mathbb{E}[\sum_{t=1}^{T} \langle \ell_t, p_t - q_t \rangle] + \mathbb{E}[\sum_{t=1}^{T} \langle \ell_t, u_a - e_a \rangle] \\
&\leq \mathbb{E}[\sum_{t=1}^{T} \langle \ell_t, q_t - u_a \rangle] + 4K \\
&= \mathbb{E}[\sum_{t=1}^{T} \langle \hat{\ell}_t, q_t - u_a \rangle] + 4K,
\end{aligned}$$

where the inequality is from $\langle \ell_t, p_t - q_t \rangle = \frac{1}{T} \sum_{i=1}^{K} \ell_{t,i}(1 - K q_{t,i}) \leq \frac{2K}{T}$ and $\langle \ell_t, u_a - e_a \rangle \leq \frac{2K}{T}$, and the last equality is from $\mathbb{E}[\hat{\ell}_t] = \ell_t$. By the standard analysis of FTRL with time-varying regularizer (Lattimore and Szepesvári, 2020), we have

$$\begin{aligned}
\sum_{t=1}^{T} \langle \hat{\ell}_t, q_t - u_a \rangle \leq \Phi_{T+1}(u_a) - \min_{p \in \Delta_K} \Phi_1(p) + \sum_{t=1}^{T} \Phi_t(q_{t+1}) - \Phi_{t+1}(q_{t+1}) \\
+ \sum_{t=1}^{T} \langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t)
\end{aligned}$$

$$
\begin{aligned}
&= \Phi_{T+1}(u_a) - \min_{p\in\Delta_K} \Phi_1(p) + \sum_{t=1}^{T}(\beta_{t+1}-\beta_t)(-\psi_{TE}(q_{t+1})) \\
&\quad + \sum_{t=1}^{T}\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \\
&\le \Phi_{T+1}(u_a) - \min_{p\in\Delta_K} \Phi_1(p) + 6d\left(\sum_{t=1}^{T}(\beta_{t+1}-\beta_t)h_t + \sum_{t=1}^{T}\frac{z_t}{\beta_t}\right) \\
&= \Phi_{T+1}(u_a) - \min_{p\in\Delta_K} \Phi_1(p) + 24\sum_{t=1}^{T}\frac{z_t}{\beta_t} \\
&\le \gamma\psi_{LB}(u_a) - \beta_1 \min_{p\in\Delta_K} \psi_{TE}(p) + 24\sum_{t=1}^{T}\frac{z_t}{\beta_t} \\
&\le \gamma K\ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha}-1) + 24\sum_{t=1}^{T}\frac{z_t}{\beta_t},
\end{aligned}
$$

where the second inequality is from Lemma 12 and Lemma 13, the second equality is from the update rule $(\beta_{t+1}-\beta_t)h_t = \frac{z_t}{\beta_t}$, the third inequality is due to $\psi_{TE}(p) \le 0$ and $\psi_{LB}(p) > 0$ for all $p \in \Delta_K$, and the last inequality is due to $(u_a)_i \ge \frac{1}{T}$.

BOUNDING $\mathbb{E}[\sum_{t=1}^{T}\frac{z_t}{\beta_t}]$

Note that we should not directly apply the SPM bound based on $z_{\max} = \max_{t\in[T]} z_t$ in Lemma 3 of Ito et al. (2024), because $z_{\max}$ is of order $\max_t p_{t,I_t}^{-\alpha}$, which can be very large. Instead, let

$$
\begin{aligned}
G &= \sum_{t=1}^{T}\frac{z_t}{\sqrt{\sum_{s=1}^{t}\frac{z_s}{h_s}}} \\
h_{\max} &= \max_{t\in[T]} h_t, \\
z_{\mathbb{E},\max} &= \max_{t\in[T]} \mathbb{E}_{I_t}[z_t].
\end{aligned}
$$

By definitions of h_t and z_t, we have $h_t \le \frac{K^{1-\alpha}-1}{\alpha}$ and

$$
\mathbb{E}_{I_t}[z_t] \le \frac{(6d)^{2-\alpha}}{2(1-\alpha)}S^{\alpha},
$$

from Lemma 11. It follows that

$$
\begin{aligned}
h_{\max} &\le \frac{K^{1-\alpha}-1}{\alpha}, \\
z_{\mathbb{E},\max} &\le \frac{(6d)^{2-\alpha}}{2(1-\alpha)}S^{\alpha}.
\end{aligned}
$$

Next, let

$$\beta'_t = \sqrt{\beta_1^2 + 2\sum_{s=1}^{t-1} \frac{z_s}{h_s}}. \tag{22}$$

Let $E = \{t \in [T] : \beta'_{t+1} \geq \sqrt{2}\beta'_t\}$ and $E^c = [T] \setminus E$. Also, let $N = |E|$ and $j = 1, 2, \ldots, N$ be the index running over the rounds in E. Similar to the proof of Lemma 2 in Ito et al. (2024), squaring both sides of $\beta_t = \beta_{t-1} + \frac{z_{t-1}}{\beta_{t-1} h_{t-1}}$ implies that

$$\beta_t^2 = \beta_{t-1}^2 + \frac{2z_{t-1}}{h_{t-1}} + \frac{z_{t-1}^2}{\beta_{t-1}^2 h_{t-1}^2} \geq \beta_{t-1}^2 + \frac{2z_{t-1}}{h_{t-1}} \geq \beta_1^2 + 2\sum_{s=1}^{t-1} \frac{z_s}{h_s},$$

which shows that $\beta'_t \leq \beta_t$. Furthermore, $\sum_{t \in E^c} \frac{z_t}{\beta_t} \leq G$ from the proof of Lemma 2 in Ito et al. (2024). Therefore,

$$\begin{aligned}
\sum_{t=1}^{T} \frac{z_t}{\beta_t} &= \sum_{t \in E^c} \frac{z_t}{\beta_t} + \sum_{t \in E} \frac{z_t}{\beta_t} \\
&\leq G + \sum_{t \in E} \frac{z_t}{\beta_t} \\
&\leq G + \frac{18d^2}{\gamma} N \\
&\leq G + \frac{18d^2}{\gamma} \log_{\sqrt{2}} \left(\frac{\beta'_{T+1}}{\beta'_1} \right) \\
&\leq G + \frac{26d^2}{\gamma} \ln \left(1 + 2\sum_{t=1}^{T} \frac{z_t}{h_t} \right),
\end{aligned}$$

where the second inequality is from the definition of z_t and the third inequality is from the fact that β'_t is multiplied by at least $\sqrt{2}$ after every round in E, thus there can be at most $N \leq \log_{\sqrt{2}} \frac{\beta'_{T+1}}{\beta'_1}$ such multiplications.

Taking the expectation over $I_{1:T}$ on both sides and using $E[\ln(X)] \leq \ln(E[X])$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\beta_t} \right] &\leq \mathbb{E}[G] + \frac{26d^2}{\gamma} \ln \left(1 + 2\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{h_t} \right] \right) \\
&= \mathbb{E}[G] + \frac{26d^2}{\gamma} \ln \left(\mathbb{E}\left[1 + 2\sum_{t=1}^{T} \frac{\mathbb{E}_{I_t}[z_t]}{h_t} \right] \right) \\
&\leq \mathbb{E}[G] + \frac{26d^2}{\gamma} \ln \left(1 + \frac{(6d)^{2-\alpha} S^\alpha}{(1-\alpha)} \mathbb{E}\left[\frac{4T}{\min_{t \in [T]} h_t} \right] \right)
\end{aligned}$$

$$\leq \mathbb{E}[G] + \frac{26d^2}{\gamma} \ln\left(1 + \frac{(6d)^{2-\alpha} S^\alpha}{(1-\alpha)} \frac{4\alpha T^{\alpha+1}}{1-\alpha}\right)$$
$$= \mathbb{E}[G] + O\left(\frac{1}{\gamma} \ln\left(\frac{\alpha S^\alpha T}{(1-\alpha)^2}\right)\right)$$

where the first equality is because h_t is $\mathbb{F}_{t-1}$-measurable and the last inequality is due to Lemma 11.

Next, Equation 45 in Ito et al. (2024) shows that $G \leq 2\sqrt{h_{\max} \sum_{t=1}^T z_t}$. Moreover, Equation 46 in Ito et al. (2024) shows that for any fixed $J \geq 1$,

$$G \leq \sqrt{8J \sum_{t=1}^T h_t z_t} + 2\sqrt{2^{-J} h_{\max} \sum_{t=1}^T z_t}.$$

As a result, we obtain the following bound:

$$\mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{\beta_t}\right] \leq \min\left\{\inf_{J\in\mathbb{N}} \mathbb{E}\left[\left\{\sqrt{8J \sum_{t=1}^T h_t z_t} + 2\sqrt{2^{-J} h_{\max} \sum_{t=1}^T z_t}\right\}\right], 2\mathbb{E}\left[\sqrt{h_{\max} \sum_{t=1}^T z_t}\right]\right\}$$
$$+ O\left(\frac{1}{\gamma} \ln\left(\frac{\alpha S^\alpha T}{(1-\alpha)^2}\right)\right). \tag{23}$$

ADVERSARIAL REGIME:

Using Jensen's inequality $E[\sqrt{X}] \leq \sqrt{E[X]}$ and Equation (23), we obtain

$$\mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{\beta_t}\right] \leq 2\sqrt{\frac{((K-1)^{1-\alpha}-1)}{\alpha} \sum_{t=1}^T \mathbb{E}[z_t]} + O\left(\frac{1}{\gamma} \ln\left(\frac{\alpha S^\alpha T}{(1-\alpha)^2}\right)\right)$$
$$\leq 2\sqrt{\frac{((K-1)^{1-\alpha}-1)}{\alpha} T\mathbb{E}[z_{\mathbb{E},\max}]} + O\left(\frac{1}{\gamma} \ln\left(\frac{\alpha S^\alpha T}{(1-\alpha)^2}\right)\right)$$
$$= O\left(\sqrt{\frac{(K^{1-\alpha}-1) S^\alpha T}{\alpha(1-\alpha)}}\right).$$

ADVERSARIAL REGIME WITH A SELF-BOUNDING CONSTRAINT:

Let $R_T = \max_{a\in[K]} R_{T,a}$. In this regime, we have $R_T + C \geq \mathbb{E}[\sum_{t=1}^T \sum_{i=1}^K q_{t,i}\Delta_i]$. Let $i^* \in [K]$ be the unique optimal arm.

Observe that given the sequence of randomly drawn arms until the beginning of round t, the quantity h_t is fixed. Therefore, we can write $\mathbb{E}[h_t z_t] = \mathbb{E}[h_t \mathbb{E}_{I_t}[z_t]]$ and obtain

$$\begin{aligned}
\mathbb{E}[h_t z_t] &= \mathbb{E}[h_t \mathbb{E}_{I_t}[z_t]] \\
&\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \mathbb{E}\left[h_t \left(\sum_{i=1}^{K} (\tilde{p}_{t,i}^{1-\alpha}) \ell_{t,i}^2 \right) \right] \\
&= \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \mathbb{E}\left[\frac{1}{\alpha} \left(\sum_{i=1}^{K} p_{t,i}^{\alpha} - 1 \right) \left(\sum_{i=1}^{K} (\tilde{p}_{t,i}^{1-\alpha}) \ell_{t,i}^2 \right) \right] \\
&\leq \frac{(6d)^{2-\alpha}}{2\alpha(1-\alpha)} \mathbb{E}\left[\left(\sum_{i=1}^{K} p_{t,i}^{\alpha} - 1 \right) \left(\sum_{\ell_{t,i} \neq 0} \tilde{p}_{t,i}^{1-\alpha} \right) \right],
\end{aligned} \tag{24}$$

where the second inequality is from $\ell_{t,i}^2 \leq 1$.

Using $p_{t,i^*}^{\alpha} - 1 \leq 0$ and $\sum_{i \neq i^*} p_{t,i}^{\alpha} \leq (K-1)^{1-\alpha} (\sum_{i \neq i^*} p_{t,i})^{\alpha}$ by Holder's inequality, we obtain

$$\begin{aligned}
\sum_{i \in [K]} p_{t,i}^{\alpha} - 1 &\leq (K-1)^{1-\alpha} (\sum_{i \neq i^*} p_{t,i})^{\alpha} \\
&\leq \frac{(K-1)^{1-\alpha}}{\Delta_{\min}^{\alpha}} (\sum_{i \in [K]} p_{t,i} \Delta_i)^{\alpha}.
\end{aligned}$$

Next, from $\tilde{p}_{t,i^*} \leq \sum_{i \neq i^*} p_{t,i}$ we have

$$\begin{aligned}
\tilde{p}_{t,i*}^{1-\alpha} &\leq \left(\sum_{i \neq i^*} \tilde{p}_{t,i} \right)^{1-\alpha} \\
&\leq \frac{1}{\Delta_{\min}^{1-\alpha}} \left(\sum_{i \neq i^*} p_{t,i} \Delta_i \right)^{1-\alpha}.
\end{aligned}$$

Therefore, by Holder's inequality,

$$\begin{aligned}
\sum_{\ell_{t,i} \neq 0} \tilde{p}_{t,i}^{1-\alpha} &\leq \left(\sum_{\ell_{t,i} \neq 0, i \neq i^*} p_{t,i}^{1-\alpha} \right) + \tilde{p}_{t,i*}^{1-\alpha} \\
&\leq \sum_{\ell_{t,i} \neq 0, i \neq i^*} \left(\Delta_i^{-\frac{1-\alpha}{\alpha}} \right)^{\alpha} (p_{t,i} \Delta_i)^{1-\alpha} + \frac{1}{\Delta_{\min}^{1-\alpha}} \left(\sum_{i \neq i^*} p_{t,i} \Delta_i \right)^{1-\alpha} \\
&\leq \left(\sum_{\ell_{t,i} \neq 0, i \neq i^*} \Delta_i^{-\frac{1-\alpha}{\alpha}} \right)^{\alpha} \left(\sum_{\ell_{t,i} \neq 0, i \neq i^*} p_{t,i} \Delta_i \right)^{1-\alpha} + \frac{1}{\Delta_{\min}^{1-\alpha}} \left(\sum_{i \neq i^*} p_{t,i} \Delta_i \right)^{1-\alpha}
\end{aligned}$$

$$\leq \frac{S^\alpha}{\Delta_{\min}^{1-\alpha}} \left(\sum_{\ell_{t,i} \neq 0, i \neq i^*} p_{t,i} \Delta_i \right)^{1-\alpha} + \frac{1}{\Delta_{\min}^{1-\alpha}} \left(\sum_{i \neq i^*} p_{t,i} \Delta_i \right)^{1-\alpha}$$
$$\leq \frac{2S^\alpha}{\Delta_{\min}^{1-\alpha}} \left(\sum_{i \in [K]} p_{t,i} \Delta_i \right)^{1-\alpha}.$$

Overall, we have

$$\mathbb{E}[h_t z_t] \leq \frac{(6d)^{2-\alpha}}{\alpha(1-\alpha)\Delta_{\min}} (K-1)^{1-\alpha} S^\alpha \mathbb{E}\left[\sum_{i=1}^K p_{t,i} \Delta_i \right]. \tag{25}$$

Furthermore, by Jensen's inequality,

$$\mathbb{E}\left[\sqrt{2^{-J} h_{\max} \sum_{t=1}^T z_t } \right] \leq \sqrt{ \mathbb{E}\left[2^{-J} h_{\max} \sum_{t=1}^T z_t \right] } \leq \sqrt{ 2^{-J} T \frac{(K-1)^{1-\alpha}}{\alpha} \frac{(6d)^{2-\alpha} S^\alpha}{2(1-\alpha)} }. \tag{26}$$

By plugging $J = \lceil \log_2(T) \rceil$, (25) and (26) into (23), we obtain

$$\mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{\beta_t} \right] \leq O\left(\sqrt{ \frac{\ln(T)(K^{1-\alpha}-1) S^\alpha \mathbb{E}[\sum_{t=1}^T \sum_{i=1}^K p_{t,i} \Delta_i]}{\alpha(1-\alpha)\Delta_{\min}} } + \sqrt{ \frac{(K^{1-\alpha}-1)S^\alpha}{\alpha(1-\alpha)} } \right)$$
$$\leq O\left(\sqrt{ \frac{\ln(T)(K^{1-\alpha}-1) S^\alpha (R_T + C)}{\alpha(1-\alpha)\Delta_{\min}} } + \sqrt{ \frac{(K^{1-\alpha}-1)S^\alpha}{\alpha(1-\alpha)} } \right).$$

In summary, keeping only the dominant $\sqrt{T}$ terms, we have the following BOBW bounds that hold simultaneously:

- In adversarial regime,

$$R_T \leq O\left(\sqrt{ \frac{(K^{1-\alpha}-1)S^\alpha T}{\alpha(1-\alpha)} } \right)$$

- In adversarial regime with a self-bounding constraint:

$$R_T \leq O\left(\frac{(K-1)^{1-\alpha} S^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}} + \sqrt{ C \frac{(K-1)^{1-\alpha} S^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}} } + \sqrt{ \frac{(K-1)^{1-\alpha} S^\alpha}{\alpha(1-\alpha)} } \right)$$

Note that we can explicitly set α sufficiently close 1 so that $\frac{K^{1-\alpha}-1}{\alpha(1-\alpha)} = O(\ln(K))$, $\frac{(K-1)^{1-\alpha}}{\alpha(1-\alpha)} = O(\ln(K))$ and while $\gamma = O(\sqrt{\frac{\alpha}{1-\alpha}})$ grows with K instead of T. For example, in Appendix G, we

show that $\alpha = 1 - \frac{1}{2\ln(K)}$ satisfies $\frac{K^{1-\alpha}-1}{\alpha(1-\alpha)} \leq 4\ln(K)$, $\frac{(K-1)^{1-\alpha}}{\alpha(1-\alpha)} \leq 4\ln(K)$ while $\gamma \lesssim \sqrt{\ln(K)}$. This ensures that

$$\begin{aligned}\gamma K \ln(T) + \frac{\beta_1(K^{1-\alpha}-1)}{\alpha} &= \gamma K \ln(T) + \frac{4K(K^{1-\alpha}-1)}{\alpha(1-\alpha)} \\ &= O\left(K\ln(K)\ln(T)\right),\end{aligned}$$

and

$$\frac{1}{1-\alpha} = O(\ln(K))$$

everywhere, so we can safely ignore the terms that do not contain $\sqrt{T}$ (in the adversarial setting) and $\frac{\ln(T)}{\Delta_{\min}}$ (in the stochastic setting).

■

B.2. Stability Proofs

In this section, we prove Lemma 12 and Lemma 13. First, we state and prove a number of supporting lemmas. In the following, we let

$$g_{\beta,\gamma}(t) = \beta t^{\alpha-1} + \frac{\gamma}{t}. \tag{27}$$

be a function defined on $(0,1) \to \mathbb{R}_+$. Note that because $\beta > 0, \alpha \in (0,1)$ and $\gamma > 0$, this function $g_{\beta,\gamma}(t)$ is monotonically decreasing in t. We will drop the subscripts β and γ whenever they are clear from the context.

The first lemma shows that $\beta_{t+1} - \beta_t$ is sufficiently small for stabilizing the FTRL update in Algorithm 1.

Lemma 14 *For any $t \geq 1$, Algorithm 1 guarantees*

$$\beta_{t+1} - \beta_t \leq (1 - \frac{1}{d})\gamma q_{t*}^{-\alpha}, \tag{28}$$

where $q_{t*} = \min(\max_{i\in[K]} q_{t,i}, 1 - \max_{i\in[K]} q_{t,i})$.

Proof Lemma 23 shows that $h_t \geq \frac{1-\alpha}{4\alpha} p_{t*}^{\alpha}$. By Lemma 24, we have $p_{t*}^{\alpha} \geq 2^{-\alpha} q_{t*}^{\alpha}$. This implies that $\frac{1}{h_t} \leq \frac{4\alpha}{1-\alpha} 2^{\alpha} q_{t*}^{-\alpha}$. By the definitions of β_{t+1}, z_t and h_t, we have

$$\begin{aligned}\beta_{t+1} - \beta_t &= \frac{z_t}{\beta_t h_t} \\ &\leq \frac{4\alpha z_t}{(1-\alpha)\beta_t} 2^{\alpha} q_{t*}^{-\alpha} \\ &\leq \frac{4\alpha}{(1-\alpha)} \frac{18d^2}{\gamma} \ell_{t,I_t}^2 2^{\alpha} q_{t*}^{-\alpha} \\ &\leq (1 - \frac{1}{d})\gamma q_{t*}^{-\alpha}\end{aligned}$$

where the last inequality uses

$$\frac{72\alpha d^2}{(1-\alpha)\gamma}\ell_{t,I_t}^2 2^\alpha \leq \frac{72\alpha d^2}{(1-\alpha)\gamma}2^\alpha \leq (1-\frac{1}{d})\gamma \tag{29}$$

for $d = 2$ and $\gamma \geq 48\sqrt{\frac{\alpha}{1-\alpha}}$. ■

Lemma 15 *For any $L \in \mathbb{R}^K, \beta > 0, \gamma > 0$ and $h \in [-1, 1]$, let*

$$x = \arg\min_{p\in\Delta_K} \langle L, p\rangle + \beta\left(\frac{1}{\alpha}(1-\sum_{i=1}^K p_i^\alpha)\right) - \gamma\sum_{i=1}^K \ln(p_i),$$

$$y = \arg\min_{p\in\Delta_K} \langle L + \frac{h}{x_1'}e_1, p\rangle + \beta\left(\frac{1}{\alpha}(1-\sum_{i=1}^K p_i^\alpha)\right) - \gamma\sum_{i=1}^K \ln(p_i).$$

Here, e_1 is the first vector in the standard basis of $\mathbb{R}^K$. If $4x_1' \geq x_1$ and $\gamma \geq 6$, then $y_1 \leq 3x_1$.

Proof Using the Lagrange multiplier method, we have the following equalities that hold for some $Z \in \mathbb{R}$,

$$\beta\left(y_1^{\alpha-1} - x_1^{\alpha-1}\right) + \gamma\left(\frac{1}{y_1} - \frac{1}{x_1}\right) = Z + \frac{h}{x_1'} \tag{30}$$

and for all $i \neq 1$,

$$\beta\left(y_i^{\alpha-1} - x_i^{\alpha-1}\right) + \gamma\left(\frac{1}{y_i} - \frac{1}{x_i}\right) = Z. \tag{31}$$

First, we show that Z and $y_1 - x_1$ has the opposite sign to h. We consider two cases:

- If $Z \geq 0$ then from (31), we have $g(y_i) - g(x_i) = Z \geq 0$. This implies $y_i \leq x_i$ and leads to $y_1 \geq x_1$. From (30), we have $Z + \frac{h}{x_1'} = g(y_1) - g(x_1) \leq 0$. Since $Z \geq 0$, this implies $h \leq 0$.
- If $Z \leq 0$ then by the same argument, we have $y_i \geq x_i$ and $y_1 \leq x_1$. Therefore, $Z + \frac{h}{x_1'} \geq 0$. Due to $Z \leq 0$, we must have $h \geq 0$.

In both cases, we have $Zh \leq 0$ and $Z(y_1 - x_1) \geq 0$. It follows that if $h \geq 0$ then we have $y_1 \leq x_1 \leq 2x_1$. If $h < 0$ then $y_1 \geq x_1$, and by rearranging (30), we obtain

$$\begin{aligned}\frac{4}{x_1} \geq -\frac{h}{x_1'} &= \underbrace{Z}_{\geq 0} + \underbrace{\gamma\left(\frac{1}{x_1} - \frac{1}{y_1}\right)}_{\geq 0} + \underbrace{\beta(x_1^{\alpha-1} - y_1^{\alpha-1})}_{\geq 0} \\ &\geq \gamma\left(\frac{1}{x_1} - \frac{1}{y_1}\right) \\ &\geq 6\left(\frac{1}{x_1} - \frac{1}{y_1}\right),\end{aligned}$$

where the last inequality is due to $\gamma \geq 6$. This implies that $\frac{3}{y_1} \geq \frac{1}{x_1}$, thus $y_1 \leq 3x_1$. ■

Lemma 16 *For any $L \in \mathbb{R}^K, \beta > 0, \beta' > 0, \gamma \geq 0$, define*

$$x = \operatorname*{arg\,min}_{p \in \Delta_K} \langle L, p \rangle + \beta \left(\frac{1}{\alpha}(1 - \sum_{i=1}^{K} p_i^\alpha) \right) - \gamma \sum_{i=1}^{K} \ln(p_i),$$
$$y = \operatorname*{arg\,min}_{p \in \Delta_K} \langle L, p \rangle + \beta' \left(\frac{1}{\alpha}(1 - \sum_{i=1}^{K} p_i^\alpha) \right) - \gamma \sum_{i=1}^{K} \ln(p_i).$$

Let $x_* = \min(\max_{i \in [K]} x_i, 1 - \max_{i \in [K]} x_i)$. For any constant $d \geq 2$, if

$$0 \leq \beta' - \beta \leq \left(1 - \frac{1}{d}\right) \gamma x_*^{-\alpha}, \tag{32}$$

then $y_i \leq d x_i$ for all $i \in [K]$.

Proof Using the Lagrange multiplier method, we have for all $i \in [K]$,

$$L_i - \beta x_i^{\alpha-1} - \frac{\gamma}{x_i} = \lambda, \tag{33}$$
$$L_i - \beta' y_i^{\alpha-1} - \frac{\gamma}{y_i} = \lambda'. \tag{34}$$

Subtracting both sides of the two equations, we obtain

$$\begin{aligned} \lambda - \lambda' + g_\beta(x_i) &= g_{\beta'}(y_i) \\ &= g_\beta(y_i) + (\beta' - \beta) y_i^{\alpha-1}. \end{aligned}$$

If $\lambda - \lambda' < 0$, then because $\beta \leq \beta'$, we have $g_\beta(x_i) > g_\beta(y_i) + (\beta' - \beta) y_i^{\alpha-1} \geq g_\beta(y_i)$. This implies $x_i < y_i$ for all $i \in [K]$, a contradiction to $\sum_{i=1}^K x_i = \sum_{i=1}^K y_i = 1$. Hence, we have $\lambda - \lambda' \geq 0$, and thus $g_\beta(x_i) \leq g_{\beta'}(y_i)$.

For any $i \in [K]$, if $x_i > x_*$ then $x_i \geq \frac{1}{2}$ and hence $y_i \leq 1 \leq 2x_i \leq d x_i$. From the condition $\beta' - \beta \leq (1 - \frac{1}{d}) \gamma x_*^{-\alpha}$, for $x_i \leq x_*$, we have

$$\begin{aligned} g_{\beta'}(y_i) &\geq g_\beta(x_i) \\ &= \beta x_i^{\alpha-1} + \frac{\gamma}{x_i} \\ &\geq (\beta' - (1 - \frac{1}{d}) \gamma x_*^{-\alpha}) x_i^{\alpha-1} + \frac{\gamma}{x_i} \\ &= \beta' x_i^{\alpha-1} - (1 - \frac{1}{d}) \gamma x_*^{-\alpha} x_i^{\alpha-1} + \frac{\gamma}{x_i} \\ &\geq \beta' x_i^{\alpha-1} - (1 - \frac{1}{d}) \gamma x_i^{-\alpha} x_i^{\alpha-1} + \frac{\gamma}{x_i} \\ &= \beta' x_i^{\alpha-1} + \frac{\gamma}{d x_i} \\ &\geq \beta' (d x_i)^{\alpha-1} + \frac{\gamma}{d x_i} \\ &= g_{\beta'}(d x_i), \end{aligned}$$

where the last inequality is due to $(d)^{\alpha-1} \leq 1$. This implies $y_i \leq dx_i$ for all $x_i \leq x_*$. ■

Let $\|x\|_A = \sqrt{x^T A x}$ be the norm of a vector $x \in \mathbb{R}^K$ induced by a positive definite matrix A. The following lemma proves Lemma 12 when the chosen arm I_t satisfies $q_{t,I_t} \leq 1 - q_{t,I_t}$.

Lemma 17 *For any $t \in [T]$, Algorithm 1 guarantees*

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min\left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)} p_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma} \ell_{t,I_t}^2 \right) \tag{35}$$

Proof Using standard local-norm analysis techniques for FTRL (for example, see Section 7.4 in Orabona (2023)), we have

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \frac{1}{2} \left\| \hat{\ell}_t \right\|^2_{(\nabla^2 \phi_t(z_t))^{-1}}, \tag{36}$$

where z_t is a point between q_t and q_{t+1}. The Hessian matrix of ϕ_t is a diagonal matrix with entries

$$\nabla^2 \phi_t(z_t) = \operatorname{diag}\left(\left(\beta_t(1-\alpha) z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2} \right)_{i=1,2,\ldots,K} \right). \tag{37}$$

Hence, its inverse is the following diagonal matrix

$$(\nabla^2 \phi(z_t))^{-1} = \operatorname{diag}\left(\left(\frac{1}{\beta_t(1-\alpha) z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2}} \right)_{i=1,2,\ldots,K} \right). \tag{38}$$

It follows that

$$\begin{aligned} \left\| \hat{\ell}_t \right\|^2_{(\nabla^2 \phi_t(z_t))^{-1}} &= \sum_{i=1}^K \hat{\ell}_{t,i}^2 \frac{1}{\beta_t(1-\alpha) z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2}} \\ &\leq \min\left(\frac{1}{\beta_t(1-\alpha)} \sum_{i=1}^K z_{t,i}^{2-\alpha} \hat{\ell}_{t,i}^2, \frac{1}{\gamma} \sum_{i=1}^K z_{t,i}^2 \hat{\ell}_{t,i}^2 \right) \\ &= \min\left(\frac{1}{\beta_t(1-\alpha)} z_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{z_{t,I_t}^2 \hat{\ell}_{t,I_t}^2}{\gamma} \right), \end{aligned} \tag{39}$$

where the last equality is due to $\hat{\ell}_{t,i} = 0$ for $i \neq I_t$. Combining (36) and (39), we obtain

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min\left(\frac{1}{2\beta_t(1-\alpha)} z_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{z_{t,I_t}^2 \hat{\ell}_{t,I_t}^2}{2\gamma} \right). \tag{40}$$

Since z_t is between q_t and q_{t+1}, we have $z_{t,I_t} \leq \max(q_{t,I_t}, q_{t+1,I_t})$. The loss estimate in Algorithm 1 uses p_{t,I_t} where $2p_{t,I_t} \geq q_{t,I_t}$ by Lemma 24, therefore we can combine the results of

Lemma 14, Lemma 16 and Lemma 15 and obtain $q_{t+1,I_t} \leq 3dq_{t,I_t}$. It follows that $z_{t,I_t} \leq 3dq_{t,I_t} \leq 6dp_{t,I_t}$, and as a result,

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min\left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)} p_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{36d^2}{2\gamma} p_{t,I_t}^2 \hat{\ell}_{t,I_t}^2 \right) \tag{41}$$

$$\leq \min\left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)} p_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma} \ell_{t,I_t}^2 \right), \tag{42}$$

where the last equality is due to $p_{t,I_t}^2 \hat{\ell}_{t,I_t}^2 = \ell_{t,I_t}^2$. ■

The next lemma proves Lemma 12 whenever the chosen arm I_t has the maximum sampling probability. The proof is largely based on Lemma 9 in Ito et al. (2024) and Equation 22 in Tsuchiya et al. (2023).

Lemma 18 *For any $t \in [T]$, if $I_t \in \arg\max_{i \in [K]} p_{t,i}$, Algorithm 1 guarantees*

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min\left(\frac{4}{\beta_t(1-\alpha)} (1 - p_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{4\ell_{t,I_t}^2}{\gamma} \right). \tag{43}$$

Proof When $I_t \in \arg\max_{i \in [K]} p_{t,i}$, we have $I_t \in \arg\max_{i \in [K]} q_{t,i}$ and thus $q_{t,I_t} \geq p_{t,I_t} \geq \frac{1}{K}$. Therefore,

$$\frac{\hat{\ell}_{t,I_t}}{\beta_t} = \frac{\ell_{t,I_t}}{p_{t,I_t}\beta_t} \leq \frac{1}{p_{t,I_t}\beta_t} \leq \frac{K}{\beta_t} \leq \frac{1-\alpha}{4} \leq \frac{1-\alpha}{4}(1 - q_{t,I_t})^{\alpha-1}, \tag{44}$$

where the third inequality is due to $\beta_t \geq \beta_1 \geq \frac{4K}{1-\alpha}$ by initialization, and the last inequality is from $(1 - q_{t,I_t})^{\alpha-1} \geq 1$ for $\alpha \in (0,1)$. Furthermore, for any $i \in [K] \setminus \{I_t\}$, we have $\frac{\hat{\ell}_{t,i}}{\beta_t} = 0 \geq -\frac{1-\alpha}{4} q_{t,i}^{\alpha-1}$. Therefore, by using Lemma 9 in Ito et al. (2024) and noting that $\hat{\ell}_{t,i} = 0$ for $i \neq I_t$, we obtain

$$\langle \frac{1}{\beta_t}\hat{\ell}_t, q_t - q_{t+1} \rangle - D_{TE}(q_{t+1}, q_t) \leq \frac{4}{\beta_t^2(1-\alpha)} (1 - q_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2. \tag{45}$$

Furthermore, Equation 22 in Tsuchiya et al. (2023) states that if $\frac{q_{t,I_t}\hat{\ell}_{t,I_t}}{\gamma} \geq -1$ and $\hat{\ell}_{t,i} = 0$ for $i \neq I_t$, then

$$\langle q_t - q_{t+1}, \hat{\ell}_t \rangle - \gamma D_{LB}(q_{t+1}, q_t) \leq \frac{q_{t,I_t}\hat{\ell}_{t,I_t}^2}{\gamma}. \tag{46}$$

Indeed, we have $\left|\frac{q_{t,I_t}\hat{\ell}_{t,I_t}}{\gamma}\right| = \left|\frac{q_{t,I_t}\ell_{t,I_t}}{p_{t,I_t}\gamma}\right| \leq \frac{1}{2\gamma} \leq \frac{1}{8}$ since $q_{t,I_t} \leq 2p_{t,I_t}$ by Lemma 24 and $\gamma \geq 4$ by definition. Therefore, (45) and (46) together implies that

$$\begin{aligned}
&\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \\
&= \langle \hat{\ell}_t, q_t - q_{t+1}\rangle - \beta_t D_{TE}(q_{t+1}, q_t) - \gamma D_{LB}(q_{t+1}, q_t) \\
&\leq \min\left(\beta_t\left(\langle \frac{1}{\beta_t}\hat{\ell}_t, q_t - q_{t+1}\rangle - D_{TE}(q_{t+1}, q_t)\right), \langle \hat{\ell}_t, q_t - q_{t+1}\rangle - \gamma D_{LB}(q_{t+1}, q_t)\right) \\
&\leq \min\left(\frac{4}{\beta_t(1-\alpha)}(1-q_{t,I_t})^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{q_{t,I_t}^2\hat{\ell}_{t,I_t}^2}{\gamma}\right) \\
&\leq \min\left(\frac{4}{\beta_t(1-\alpha)}(1-p_{t,I_t})^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{4\ell_{t,I_t}^2}{\gamma}\right),
\end{aligned} \tag{47}$$

where the first inequality is because Bregman divergences are non-negative. ■

Next, we prove Lemma 12.

Proof (Of Lemma 12) We consider two cases:

- If $I_t \notin \arg\max_{i\in[K]} p_{t,i}$ or $p_{t,I_t} \leq 1 - p_{t,I_t}$: in this case, we have $p_{t,I_t} = \min(p_{t,I_t}, 1-p_{t,I_t})$. By Lemma 17, we have

$$\begin{aligned}
\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) &\leq \min\left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)}p_{t,I_t}^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma}\ell_{t,I_t}^2\right) \\
&= \min\left(\frac{(6d)^{2-\alpha}}{2\beta_t(1-\alpha)}\tilde{p}_{t,I_t}^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma}\ell_{t,I_t}^2\right).
\end{aligned}$$

- If $p_{t,I_t} > 1 - p_{t,I_t}$: in this case, we have $I_t \in \arg\max_{i\in[K]} q_{t,i}$. By Lemma 18,

$$\begin{aligned}
\langle \hat{\ell}_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) &\leq \min\left(\frac{4}{\beta_t(1-\alpha)}(1-p_{t,I_t})^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{4\ell_{t,I_t}^2}{\gamma}\right) \\
&= \min\left(\frac{4}{\beta_t(1-\alpha)}\min(p_{t,I_t}, 1-p_{t,I_t})^{2-\alpha}\hat{\ell}_{t,i}^2, \frac{4\ell_{t,I_t}^2}{\gamma}\right).
\end{aligned}$$

Lemma 12 follows by noting that $\max\left(\frac{(6d)^{2-\alpha}}{2}, 4\right) = \frac{(6d)^{2-\alpha}}{2}$. ■

Lemma 19 *For any $L \in \mathbb{R}^K, \beta > 0, \gamma > 0$ and $h \in [-1, 1]$, let*

$$x = \arg\min_{p\in\Delta_K} \langle L, p\rangle + \beta\left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha)\right) - \gamma\sum_{i=1}^K \ln(p_i),$$

$$y = \operatorname*{arg\,min}_{p \in \Delta_K} \langle L + \frac{h}{x'_1} e_1, p \rangle + \beta \left(\frac{1}{\alpha}(1 - \sum_{i=1}^{K} p_i^{\alpha}) \right) - \gamma \sum_{i=1}^{K} \ln(p_i),$$

where $4x'_1 \geq x_1$. *Fix an arbitrary* $\omega \in (1, 2]$. *If* $\beta \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ *and* $\gamma \geq 6$, *then* $y_i \leq 3x_i$ *for all* $i \in [K]$.

Proof If $h \leq 0$, then we have $x_1 \leq y_1 \leq 3x_1$ and $y_i \leq x_i \leq 3x_i$ for all $i \neq 1$ from the proof of Lemma 15. Thus, we focus on the case $h > 0$. In this case, we have $y_1 \leq x_1 \leq 3x_1$ and $x_i \leq y_i$ for $i \neq 1$. From (30) and (31), we have

$$g(x_i) - g(y_i) = -Z \geq 0$$

for all $i \neq 1$, and

$$g(y_1) - g(x_1) = Z + \frac{h}{x'_1} \geq 0.$$

The latter implies that $-Z \leq \frac{1}{x'_1}$. Let $\epsilon = \frac{1}{\beta(1-\omega^{\alpha-1})} \leq \frac{\omega-1}{4K} \leq \frac{1}{4K}$. Similar to the proof of Lemma 13 in Ito et al. (2024), we consider two cases:

- If $x'_1 \geq \epsilon$, then $-Z \leq \frac{1}{\epsilon}$. For all $i \neq 1$,

$$\begin{aligned} g(y_i) &= g(x_i) + Z \\ &\geq g(x_i) - \frac{1}{\epsilon} \\ &= \beta x_i^{\alpha-1} - \beta(1 - \omega^{\alpha-1}) + \frac{\gamma}{x_i} \\ &\geq \beta x_i^{\alpha-1} - \beta x_i^{\alpha-1}(1 - \omega^{\alpha-1}) + \frac{\gamma}{\omega x_i} \\ &= \beta(\omega x_i)^{\alpha-1} + \frac{\gamma}{\omega x_i} = g(\omega x_i), \end{aligned}$$

  where the last inequality is due to $x_i^{\alpha-1} \geq 1$ and $\omega > 1$. This implies that for all $i \neq 1$, $y_i \leq \omega x_i \leq 3x_i$ (since $\omega \leq 2$).

- If $x'_1 < \epsilon$, then we have $x_1 \leq 4x'_1 < \frac{1}{K}$. For any $i^* \in \arg\max_{i \in [K]} x_{t,i}$, we have $i* \neq 1$. Similar to the proof of Lemma 13 in Ito et al. (2024), we have $i^* \neq 1$ and $1 \leq \frac{y_{i*}}{x_{i*}} \leq \omega$. It follows that

$$\begin{aligned} -Z &= g(x_{i*}) - g(y_{i*}) \\ &= \beta x_{i*}^{\alpha-1} + \frac{\gamma}{x_{i*}} - (\beta y_{i*}^{\alpha-1} + \frac{\gamma}{y_{i*}}) \\ &\leq \beta x_{i*}^{\alpha-1} + \frac{\gamma}{x_{i*}} - (\beta \omega^{\alpha-1} x_{i*}^{\alpha-1} + \frac{\gamma}{\omega x_{i*}}) \\ &= g(x_{i*}) - g(\omega x_{i*}). \end{aligned}$$

  As the function $g(t) - g(\omega t)$ is decreasing for $\omega > 1$, we conclude that $-Z \leq g(x_i) - g(\omega x_i)$ for all $i \in [K]$. Therefore, $g(y_i) = g(x_i) + Z \geq g(\omega x_i)$ for all $i \neq 1$, which implies $y_i \leq \omega x_i \leq 3x_i$.

In both cases, we have $y_i \leq 3x_i$ for all $i \neq 1$. Combining this with $y_1 \leq x_1$, we conclude that $y_i \leq 3x_i$ for all $i \in [K]$. ■

The following corollary is obtained by combining Lemma 16 and Lemma 19.

Corollary 20 *For any $L \in \mathbb{R}^K, \beta > 0, \gamma > 0$ and $h \in [-1, 1]$, let*

$$x = \arg\min_{p \in \Delta_K} \langle L, p \rangle + \beta \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha) \right) - \gamma \sum_{i=1}^K \ln(p_i),$$

$$y = \arg\min_{p \in \Delta_K} \langle L + \frac{h}{x_1'} e_1, p \rangle + \beta' \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha) \right) - \gamma \sum_{i=1}^K \ln(p_i).$$

Let $x_* = \min(\max_{i \in [K]} x_i, 1 - \max_{i \in [K]} x_i)$. For any $\omega \in (1, 2]$ and $d = 2$, if $2x_1' \geq x_1$, $\gamma \geq 6, \beta \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ and

$$0 \leq \beta' - \beta \leq (1 - \frac{1}{d}) \gamma x_*^{-\alpha}, \tag{48}$$

then $y_i \leq 3dx_i$ for all $i \in [K]$.

Proof Let

$$\bar{x} = \arg\min_{p \in \Delta_K} \langle L, p \rangle + \beta' \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha) \right) - \gamma \sum_{i=1}^K \ln(p_i).$$

Here, $\bar{x}$ differs from x only by the learning rates $\beta' \geq \beta$. Applying Lemma 16 with $d = 2$, we obtain $\bar{x}_i \leq dx_i$ for all $i \in [K]$. In particular, $\bar{x}_1 \leq 2x_1$. Since $x_1 \leq 2x_1'$, we have $\bar{x}_1 \leq 4x_1'$. Next, since $\bar{x}$ differs from y only by $\frac{h}{x_1'} e_1$ in the dot product, we apply Lemma 19 and obtain $y_i \leq 3\bar{x}_i$ for all $i \in [K]$. Overall, we obtain $y_i \leq 3\bar{x}_i \leq 3dx_i$ for all $i \in [K]$. ■

Finally, we are now ready to prove Lemma 13.

Proof (Of Lemma 13) Let $\omega = 2$, we have $\beta_1 = \frac{8K}{1-\alpha} \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ due to $2^\alpha \leq 1 + \alpha$ for $\alpha \in [0, 1]$. Since $z_t, h_t \geq 0$, the sequence of learning rates $(\beta_t)_t$ is increasing and hence, $\beta_t \geq \beta_1 \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ for all $t \geq 1$. Together with Lemma 14, we have $\beta_{t+1} - \beta_t \leq \left(1 - \frac{1}{d}\right) \gamma q_{t*}^{-\alpha}$ and $\beta_t \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$ for all $t \geq 1$. In addition, we have $p_{t,I_t} \geq 2q_{t,I_t}$ by Lemma 24. Applying Corollary 20 for

$$q_t = \arg\min_{x \in \Delta_K} \langle L_{t-1}, x \rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K x_i^\alpha) \right) - \gamma \sum_{i=1}^K \ln(x_i),$$

$$q_{t+1} = \arg\min_{x \in \Delta_K} \langle L_{t-1} + \frac{\ell_{t,I_t}}{p_{t,I_t}} e_{I_t}, x \rangle + \beta_{t+1} \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K x_i^\alpha) \right) - \gamma \sum_{i=1}^K \ln(x_i),$$

we obtain $q_{t+1,i} \leq 3dq_{t,i}$ for all $i \in [K]$.

For the second statement, we apply Lemma 11 in Ito et al. (2024). ■

B.3. Technical Lemmas

Lemma 21 *For any $a > 0, b > 0$ and $x \geq 0$, we have*

$$a^x + b^x \geq (a+b)^x \qquad \text{if } x \in [0,1] \tag{49}$$
$$a^x + b^x \leq (a+b)^x \qquad \text{if } x \geq 1. \tag{50}$$

.

Proof Consider the following function defined on $\mathbb{R}_+$:

$$f(x) = \ln(a^x + b^x) - x\ln(a+b). \tag{51}$$

Its derivative is

$$f'(x) = \frac{a^x \ln(a) + b^x \ln(b)}{a^x + b^x} - \ln(a+b) \tag{52}$$
$$= \frac{a^x \ln(\frac{a}{a+b}) + b^x \ln(\frac{b}{a+b})}{a^x + b^x}. \tag{53}$$

Since $\frac{a}{a+b} \leq 1$ and $\frac{b}{a+b} \leq 1$, we have $f'(x) \leq 0$. Therefore,

- If $x \in [0,1]$: we have $f(x) \geq f(1) = 0$. This implies $\ln(a^x + b^x) \geq x\ln(a+b)$ for all $x \in [0,1]$. Equivalently, $a^x + b^x \geq e^{x\ln(a+b)} = (a+b)^x$.
- If $x \geq 1$: we have $f(x) \leq f(1) = 0$. This implies $\ln(a^x + b^x) \leq x\ln(a+b)$, which leads to $a^x + b^x \leq e^{x\ln(a+b)} = (a+b)^x$ for $x \geq 1$.

■

Lemma 22 *Let $0 \leq a, b \leq 1$ and $a + b = 1$. Then, for any $x \in [0,1]$, we have*

$$(\max(a,b))^x + (\min(a,b))^x 2^{x-1} \geq 1. \tag{54}$$

Proof Without loss of generality, assume $a \geq b$. It follows that $2b \leq 1$. Consider the following function defined on $[0,1]$:

$$f(x) = \frac{b^x 2^x}{2} + a^x. \tag{55}$$

Its derivative is

$$f'(x) = \frac{1}{2}\left[b^x \ln(b) 2^x + b^x 2^x \ln(2)\right] + a^x \ln(a) \tag{56}$$
$$= b^x 2^{x-1} \ln(2b) + a^x \ln(a). \tag{57}$$

Since $2b \leq 1$ and $a \leq 1$, we have $f'(x) \leq 0$. Therefore, for all $x \in [0,1]$, we have

$$f(x) \geq f(1) = a + b = 1. \tag{58}$$

■

Lemma 23 *For any $q \in \Delta_K$, we have*

$$(-\psi_{TE}(q_t)) = \frac{1}{\alpha}\left(\sum_{i=1}^{K} q_i^\alpha - 1\right) \geq \frac{q_*^\alpha}{\alpha}(1 - 2^{\alpha-1}), \tag{59}$$

where $q_* = \min(\max_{i\in[K]} q_i, 1 - \max_{i\in[K]} q_i)$. This implies $(-\psi_{TE}(q_t)) \geq \frac{q_*^\alpha}{4\alpha}(1-\alpha)$.

Proof Let $q_{\max} = \max_{i\in[K]} q_i$. We consider two cases: $q_{\max} \leq 0.5$ and $q_{\max} > 0.5$.

- When $q_{\max} \leq 0.5$: we have $q_* = q_{\max}$. For any $i_{\max} \in \arg\max_{i\in[K]} q_i$, the inequality (59) is equivalent to

  $$q_*^\alpha 2^{\alpha-1} + \sum_{i\neq i_{\max}} q_i^\alpha \geq 1, \tag{60}$$

  Using

  $$\sum_{i\neq i_{\max}} q_i^\alpha \geq (\sum_{i\neq i_{\max}} q_i)^\alpha \tag{61}$$

  from Lemma 21 and combining with Lemma 22 leads to the desired claim.

- When $q_{\max} > 0.5$: in this case, we have $q_* = 1 - q_{\max} = \sum_{i\neq i_{\max}} q_i$. The desired inequality is equivalent to

  $$(q_{i_{\max}}^\alpha + q_*^\alpha 2^{\alpha-1}) + \sum_{i\neq i_{\max}} q_i^\alpha \geq 1 + (\sum_{i\neq i_{\max}} q_i)^\alpha. \tag{62}$$

  Again, this follows directly from

  $$\sum_{i\neq i_{\max}} q_i^\alpha \geq (\sum_{i\neq i_{\max}} q_i)^\alpha \tag{63}$$

  and Lemma 22.

The implication statement follows by $2^{\alpha-1} \leq (\alpha+3)/4$ for $\alpha \in [0,1]$. ■

Lemma 24 *Let $q \in \Delta_K$ and $p = (1 - \frac{K}{T})q + \frac{1}{T}\mathbf{1}$ where $T \geq 4K$. The following properties hold:*

- *$q_i \leq 2p_i$ for any $i \in [K]$.*
- *$q_* \leq 2p_*$ where $q_* = \min(\max_{i\in[K]} q_i, 1 - \max_{i\in[K]} q_i)$ and $p_* = \min(\max_{i\in[K]} p_i, 1 - \max_{i\in[K]} p_i)$.*
- *$p_* \geq \frac{1}{T}$.*

Proof By the definition of p, we have

$$2p_i = 2(1 - \frac{K}{T})q_i + \frac{2}{T} \geq q_i + q_i(1 - \frac{2K}{T}) \geq q_i.$$

Thus, the first statement holds.

Next, let $k \in \arg\max_{i\in[K]} q_i$. Obviously, we have $k \in \arg\max_{i\in[K]} p_i$ due to $p_i \geq p_j$ if $q_i \geq q_j$. If $q_k \leq 0.5$, then we have $q_* = q_k$. We also have $q_k \geq \frac{1}{K}$ and therefore $p_k = q_k + \frac{1-Kq_k}{T} \leq q_k \leq 0.5$. Moreover,

$$2p_* = 2p_k \geq q_k = q_*,$$

where the inequality is from the first statement. On the other hand, if $q_k > 0.5$ then we have $q_* = 1 - q_k$ and

$$\begin{aligned} p_* &= \min(p_k, 1 - p_k) \\ &= \min\left((1 - \frac{K}{T})q_k + \frac{1}{T}, 1 - \left((1 - \frac{K}{T})q_k + \frac{1}{T}\right)\right) \\ &= \min\left(q_k + \frac{1 - Kq_k}{T}, 1 - q_k + \frac{Kq_k - 1}{T}\right). \end{aligned}$$

Since $q_k \geq \frac{1}{K}$, we have $1 - q_k + \frac{Kq_k-1}{T} \geq 1 - q_k \geq \frac{1-q_k}{2}$. In addition,

$$q_k + \frac{1 - Kq_k}{T} \geq q_k + \frac{-K}{T} > 0.5 - \frac{1}{4} = \frac{1}{4} \geq \frac{1 - q_k}{2}.$$

Hence, we conclude that $p_* \geq \frac{1-q_k}{2} = \frac{q_*}{2}$. Thus, the second statement holds. The last statement follows from $p_* \geq \min_{i\in[K]} p_i \geq \frac{1}{T}$. ■

Proof (Of Lemma 11) Lemma 23 implies that for all t, we have

$$h_t \geq \frac{(p_t)_*^\alpha}{4\alpha} \geq \frac{T^{-\alpha}}{4\alpha},$$

where the last inequality is from $(p_t)_* \geq \frac{1}{T}$ by Lemma 24. Additionally,

$$\begin{aligned} \mathbb{E}_{I_t}[z_t] &\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \mathbb{E}_{I_t}\left[(\tilde{p}_{t,I_t})^{2-\alpha} \hat{\ell}_{t,I_t}^2\right] \\ &= \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \sum_{i=1}^K \frac{(\tilde{p}_{t,i})^{2-\alpha} \ell_{t,i}^2}{p_{t,i}} \\ &\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \sum_{i=1}^K (\tilde{p}_{t,i})^{1-\alpha} (\ell_{t,i}^{\frac{2}{\alpha}})^\alpha \\ &\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} (\sum_{i=1}^K \tilde{p}_{t,i})^{1-\alpha} (\sum_{i=1}^K \ell_{t,i}^{2/\alpha})^\alpha \\ &\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} S^\alpha, \end{aligned}$$

where the last inequality uses $\sum_{i=1}^K \tilde{p}_{t,i} \leq \sum_{i=1}^K p_{t,i} = 1$ and $S \geq \|\ell_t\|_0$. ■

Appendix C. Proof of the Lower Bounds in Theorem 6

C.1. Stochastic Lower Bound

Let $i^* \in [K]$ be fixed. Recall that $0 < \Delta_{\min} \leq \frac{1}{4}$ and $1 \leq U \leq \frac{K^\alpha}{4}$. We pick $b = \frac{U - \Delta_{\min}}{K^\alpha}$ so that $bK^\alpha + \Delta_{\min} = U$. We then have $\frac{U}{2K^\alpha} \leq b < \frac{1}{4}$. Our construction is as follows:

$$\ell_t = \begin{cases} -\mathbf{1} & \text{with probability } b, \\ -e_{i^*} & \text{with probability } \Delta_{\min}, \\ \mathbf{0} & \text{with probability } 1 - 2\Delta_{\min}, \end{cases}$$

where e_{i^*} is the i^*-th vector in the standard basis of $\mathbb{R}^K$. The expected loss vector is

$$\mathbb{E}\left[\ell_t\right] = -b\mathbf{1} - \Delta_{\min} e_{i^*}.$$

It follows that $\Delta_i = \Delta_{\min}$ for all $i \in [K] \setminus \{i^*\}$. In addition, the losses of arm i^* follow a Bernoulli distribution $\mathrm{Ber}(b + \Delta_{\min})$ while the losses of sub-optimal arms follow a Bernoulli distribution $\mathrm{Ber}(b)$. We verify that the constraint in (12) holds:

$$\mathbb{E}\left[\left(\sum_{i=1}^{K} \left|\ell_{t,i}\right|^{2/\alpha}\right)^{\alpha}\right] = bK^\alpha + \Delta_{\min} = U.$$

For any consistent algorithm such that $R_T = o(T^x)$ for any $x > 0$, by Lai and Robbins (1985), we have

$$\begin{aligned} \liminf_{T\to\infty} \frac{R_T}{\ln(T)} &\geq \sum_{i \neq i^*} \frac{\Delta_i}{KL(b \,\|\, b + \Delta_{\min})} \\ &\geq \frac{K\Delta_{\min}}{2KL(b \,\|\, b + \Delta_{\min})} \\ &\gtrsim \frac{Kb}{\Delta_{\min}} \\ &\gtrsim \frac{K^{1-\alpha}U}{\Delta_{\min}}, \end{aligned}$$

where the second inequality is from $K - 1 \geq \frac{K}{2}$ for all $K \geq 4$, the third inequality is by Lemma 25, and the last inequality is $b \geq \frac{U}{2K^\alpha}$.

Lemma 25 *For any $b, \Delta \in (0, \frac{1}{4}]$, we have*

$$KL(b \,\|\, b + \Delta) \leq \frac{4\Delta^2}{3b}.$$

Proof

$$KL(b \,\|\, b + \Delta) = b\ln(\frac{b}{b+\Delta}) + (1-p)\ln\left(\frac{1-b}{1-\Delta-p}\right)$$

$$
\begin{aligned}
&= -b\ln(1+\frac{\Delta}{b}) - (1-b)\ln\left(1-\frac{\Delta}{1-b}\right) \\
&\le b\left(\frac{\Delta^2}{b^2} - \frac{\Delta}{b}\right) + (1-b)\left(\frac{\Delta^2}{(1-b)^2} + \frac{\Delta}{1-b}\right) \\
&= \Delta^2\left(\frac{1}{b} + \frac{1}{1-b}\right) = \frac{\Delta^2}{b(1-b)} \le \frac{4\Delta^2}{3b},
\end{aligned}
$$

where the first inequality is due to $\ln(1+x) \ge x - x^2$ for all $x > 0$ and $\ln(1-x) \ge -x - x^2$ for all $0 \le x \le \frac{1}{3}$, and the second inequality is $1 - b \ge \frac{3}{4}$ for all $b \le \frac{1}{4}$. ■

C.2. Adversarial Lower Bound

For the adversarial lower bound, we construct a neutral environment V_0 and K competing environments $V_1, V_2, \ldots, V_K$, where:

- On V_0, the loss function is chosen by

$$
\ell_t = \begin{cases} -\mathbf{1} & \text{with probability } \eta, \\ \mathbf{0} & \text{with probability } 1-\eta. \end{cases}
$$

  It follows that the losses of all arms follow a Bernoulli distribution $\mathrm{Ber}(\eta)$ on V_0.

- On V_i for $i \in [K]$, the loss function is chosen by

$$
\ell_t = \begin{cases} -\mathbf{1} & \text{with probability } \eta, \\ -e_i & \text{with probability } \epsilon, \\ \mathbf{0} & \text{with probability } 1-\eta-\epsilon. \end{cases}
$$

  It follows that except for arm i, the losses of all other arms follow a Bernoulli distribution $\mathrm{Ber}(\eta)$ on V_i. The loss of arm i follows $\mathrm{Ber}(\eta+\epsilon)$.

Here, η and ϵ are constants chosen to be the solution of the following system of (in)equalities:

- $\eta + \epsilon \le \frac{1}{4}$.
- $\eta K^\alpha + \epsilon = U$.
- $\frac{T}{K}\frac{8\epsilon^2}{\eta} = 1$.
- $\eta K^\alpha \ge \frac{U}{2}$.

Note that for all $K \ge 4, T \ge 4K$ and $U \le \frac{K^\alpha}{4}$, the solution

$$
\begin{aligned}
\sqrt{\eta} &= \frac{-\sqrt{\frac{K}{8T}} + \sqrt{\frac{K}{8T} + 4K^\alpha U}}{2K^\alpha}, \\
\epsilon &= \sqrt{\frac{\eta K}{8T}}
\end{aligned}
\tag{64}
$$

satisfies the system of inequalities since

$$\sqrt{\eta} = \frac{-\sqrt{\frac{K}{8T}} + \sqrt{\frac{K}{8T} + 4K^\alpha U}}{2K^\alpha} \leq \frac{\sqrt{1+\frac{1}{32}}}{2K^\alpha} \leq \frac{1}{8},$$
$$\epsilon = \sqrt{\frac{\eta K}{8T}} \leq \sqrt{\frac{K}{64T}} \leq \frac{1}{8}.$$

Moreover, we have

$$\begin{aligned}
\eta &= \frac{1}{4K^{2\alpha}} \left(-\sqrt{\frac{K}{8T}} + \sqrt{\frac{K}{8T} + 4K^\alpha U} \right)^2 \\
&= \frac{1}{4K^{2\alpha}} \frac{16K^{2\alpha}U^2}{\left(\sqrt{\frac{K}{8T}} + \sqrt{\frac{K}{8T} + 4K^\alpha U} \right)^2} \\
&= \frac{4U^2}{\left(\sqrt{\frac{K}{8T}} + \sqrt{\frac{K}{8T} + 4K^\alpha U} \right)^2} \\
&\geq \frac{U^2}{4K^\alpha U} = \frac{1}{4}\frac{U}{K^\alpha} \\
&\geq \frac{K^{1-2\alpha}}{8T},
\end{aligned}$$

where the second equality is $\sqrt{a} - \sqrt{b} = \frac{a-b}{\sqrt{a}+\sqrt{b}}$, the first inequality is $\frac{K}{T} \leq \frac{K^\alpha U}{4}$ and the last inequality is $\frac{U}{K^\alpha} \geq \frac{K^{1-2\alpha}}{2T}$, both hold for all $T \geq 4K$ and $U \geq 1$. This implies that

$$\eta^2 K^{2\alpha} = \eta(\eta K^{2\alpha}) \geq \eta \frac{K^{1-2\alpha}}{8T} K^{2\alpha} = \frac{\eta K}{8T} = \epsilon^2.$$

As a result, $\eta K^\alpha \geq \epsilon = U - \eta K^\alpha$. Hence, $\eta K^\alpha \geq \frac{U}{2}$.

With the choice of η and ϵ in (64), we verify that (12) holds:

$$\mathbb{E}\left[\left(\sum_{i=1}^K \left|\ell_{t,i}\right|^{2/\alpha} \right)^\alpha \right] = \eta K^\alpha + \epsilon = U.$$

Let $\mathcal{A}$ denote the algorithm of a learner. Let $N_i = \sum_{t=1}^T \mathbb{1}\{I_t = i\}$ denote the number of times arm i is pulled by $\mathcal{A}$. Let $\mathbb{P}_0$ and $\mathbb{P}_i$ denote the distribution of the observed losses on V_0 and V_i, respectively. Similarly, let $\mathbb{E}_0$ and $\mathbb{E}_i$ denote the expectation taken on V_0 and V_i, respectively.

We first run $\mathcal{A}$ on V_0. Let $a = \arg\min_{i\in[K]} \mathbb{E}_0[N_i]$ be the arm that is pulled the least in expectation. Since $\sum_{i=1}^K N_i = T$, we have $\mathbb{E}_0[N_a] \leq \frac{T}{K}$.

By the standard arguments in establishing lower bounds for adversarial bandits (e.g. Auer et al., 2002a, Equation 28-30), we have

$$\mathbb{E}_a[N_a] \leq \mathbb{E}_0[N_a] + \frac{T}{2}\|\mathbb{P}_a - \mathbb{P}_0\|_1$$

$$
\begin{aligned}
&\leq \mathbb{E}_0[N_a] + \frac{T}{2}\sqrt{2\ln(2) KL(\mathbb{P}_0 \parallel \mathbb{P}_a)} \\
&\leq \mathbb{E}_0[N_a] + \frac{T}{2}\sqrt{2\ln(2)\mathbb{E}_0[N_a] KL(\eta \parallel \eta+\epsilon)} \\
&\leq \frac{T}{K} + \frac{T}{2}\sqrt{2\ln(2)\frac{T}{K} KL(\eta \parallel \eta+\epsilon)} \\
&\leq \frac{T}{K} + \frac{T}{2}\sqrt{2\ln(2)\frac{T}{K}\frac{4\log_2(e)\epsilon^2}{\eta}} \\
&= \frac{T}{K} + \frac{T}{2}\sqrt{\frac{T}{K}\frac{8\epsilon^2}{\eta}},
\end{aligned}
$$

where

- the second inequality is Pinsker's inequality,
- the third inequality is due to the chain rule for KL-divergence (Cover and Thomas, 2006) and the fact that V_0 and V_a differ only by the loss distribution of arm a,
- the fourth inequality is because $\mathbb{E}_0[N_a] \leq \frac{T}{K}$,
- the last inequality is the reverse Pinsker's inequality (Sason, 2015).

This further implies that $\mathbb{E}_a[N_a] \leq \frac{T}{K} + \frac{T}{2} \leq \frac{3T}{4}$ for $K \geq 4$. Hence,

$$
\begin{aligned}
\mathbb{E}_a[R_{T,a}] &= \epsilon(T - \mathbb{E}_a[N_a]) \\
&\geq \epsilon(T - \frac{3T}{4}) = \frac{T\epsilon}{4} \\
&= \sqrt{\frac{TK\eta}{32}} \\
&= \Omega(\sqrt{K^{1-\alpha}UT}),
\end{aligned}
$$

where the last equality is due to $\eta K^\alpha \geq \frac{U}{2}$.

Appendix D. Proofs for Section 4

Before proving Theorem 7, in Appendix D.1, we first establish the BOBW regret bound for an algorithm that combines real-time SPM and Optimistic FTRL without any reservoir samplings. The full procedure is given in Algorithm 3. Later on, in Appendix D.3, we will use this regret bound in the analysis of Algorithm 4.

D.1. A General SPM-based Regret Bound for Optimistic FTRL

We consider the adversarial multi-armed bandits with losses in $[0, 1]$. Note that the analysis can be trivially extended to the $[-1, 1]$ case by increasing β_t and γ by a multiplicative factor of 2.

Input: $K \geq 1, T \geq 4K, \alpha \in (0,1), \beta_1 = \frac{4K}{1-\alpha}, \gamma = \max(3, 48\sqrt{\frac{\alpha}{1-\alpha}}), d = 2.$
Initialize $L_{0,i} = 0$ for $i \in [K]$
for *each round* $t = 1, \ldots, T$ **do**
  Compute $m_t \in [0,1]^K$
  Compute $q_t = \arg\min_{x \in \Delta_K} \langle m_t + L_{t-1}, x \rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K x_i^\alpha) \right) - \gamma \sum_{i=1}^K \ln(x_i)$
  Compute $p_t = \left(1 - \frac{K}{T}\right) q_t + \frac{1}{T}\mathbf{1}$
  Draw $I_t \sim p_t$ and observe ℓ_{t,I_t}
  Compute loss estimate $\hat{\ell}_{t,i} = m_{t,i} + \frac{(\ell_{t,i} - m_{t,i})\mathbb{1}\{I_t = i\}}{p_{t,i}}$ $L_{t,i} = L_{t-1,i} + \hat{\ell}_{t,i}$
  Compute $z_t = \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \tilde{p}_{t,I_t}^{2-\alpha} (\hat{\ell}_{t,I_t} - m_{t,I_t})^2, \frac{\beta_t 18 d^2}{\gamma} (\ell_{t,I_t} - m_{t,I_t})^2 \right)$
  Compute $h_t = \left(\frac{1}{\alpha}(\sum_{i=1}^K p_{t,i}^\alpha - 1) \right)$
  Compute $\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}$
end

Algorithm 3: Optimistic FTRL using Tsallis entropy plus log-barrier regularization for losses in $[0,1]$

In round t, the learner computes $m_t \in [0,1]^K$ before drawing arm I_t and uses Optimistic FTRL with the hybrid regularizer

$$\begin{aligned} q_t &= \arg\min_{x \in \Delta_K} \langle m_t + \sum_{s=1}^{t-1} \hat{\ell}_s, x \rangle + \phi_t(x) \\ &= \arg\min_{x \in \Delta_K} \langle m_t + L_{t-1}, x \rangle + \beta_t \psi_{TE}(x) + \gamma \psi_{LB}(x) \\ &= \arg\min_{x \in \Delta_K} \langle m_t + L_{t-1}, x \rangle + \beta_t \left(1 - \sum_{i=1}^K x_i^\alpha \right) - \gamma \sum_{i=1}^K \ln(p_i). \end{aligned}$$

Let $p_t = \left(1 - \frac{K}{T}\right) q_t + \frac{1}{T}\mathbf{1}$. The learner draws $I_t \sim p_t$ and use the unbiased loss estimator

$$\hat{\ell}_{t,i} = m_{t,i} + \frac{\ell_{t,i} - m_{t,i}}{p_{t,i}} \mathbb{1}\{I_t = i\}.$$

SPM learning rates: the learning rates are set according to SPM rule (Ito et al., 2024), where

$$\begin{aligned} h_t &= (-\psi_{TE})(p_t) = \frac{1}{\alpha} \left(\sum_{i=1}^K p_{t,i}^\alpha - 1 \right), \\ z_t &= \min \left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \tilde{p}_{t,I_t}^{2-\alpha} (\hat{\ell}_{t,I_t} - m_{t,I_t})^2, \frac{\beta_t 18 d^2}{\gamma} (\ell_{t,I_t} - m_{t,I_t})^2 \right). \end{aligned}$$

Details are given in Algorithm 3.

D.2. Analysis for Algorithm 3

Similar to Ito et al. (2022), the analysis uses

$$\begin{aligned} r_{t+1} &= \arg\min_{x\in\Delta_K} \langle L_{t-1} + \hat{\ell}_t, x\rangle + \frac{\beta_{t+1}}{\alpha}(1 - \sum_{i=1}^{K} x_i^\alpha) - \gamma \sum_{i=1}^{K} \ln(x_i) \\ &= \arg\min_{x\in\Delta_K} \langle L_{t-1} + m_t + \frac{\ell_{tI_t} - m_{t,I_t}}{p_{t,I_t}} e_{I_t}, x\rangle + \frac{\beta_{t+1}}{\alpha}(1 - \sum_{i=1}^{K} x_i^\alpha) - \gamma \sum_{i=1}^{K} \ln(x_i), \end{aligned}$$

where $2p_{t,I_t} \geq q_{t,I_t}$. Observe that $(\ell_{t,I_t} - m_{t,I_t}) \in [-1, 1]$ and

$$\begin{aligned} \beta_{t+1} - \beta_t &= \frac{z_t}{\beta_t h_t} \\ &\leq \frac{18d^2}{h_t \gamma} \\ &\leq \left(1 - \frac{1}{d}\right) \gamma q_{t*}^{-\alpha}, \end{aligned}$$

similar to the proof of Lemma 14. Therefore, we can invoke Corollary 20 with $\omega = 2$ and obtain $r_{t+1,i} \leq 3q_{t,i} \leq 6p_{t,i}$ for all $i \in [K]$. Combining this with Lemma 1 in Ito et al. (2022) and our Lemma 12, we obtain

$$\begin{aligned} \sum_{t=1}^{T} \langle \hat{\ell}_t, q_t - u\rangle &\leq \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T} (\phi_t(r_{t+1}) - \phi_{t+1}(r_{t+1})) \\ &\quad + \sum_{t=1}^{T} \langle \hat{\ell}_t - m_t, q_t - r_{t+1}\rangle - D_t(r_{t+1}, q_t) \\ &\leq \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T} (\beta_{t+1} - \beta_t)(-\psi_{TE}(r_{t+1})) + \sum_{t=1}^{T} \frac{z_t}{\beta_t} \\ &\leq \phi_{T+1}(u) - \phi_1(r_1) + 6\left(\sum_{t=1}^{T} (\beta_{t+1} - \beta_t)(-\psi_{TE}(p_{t+1})) + \sum_{t=1}^{T} \frac{z_t}{\beta_t}\right) \\ &= \phi_{T+1}(u) - \phi_1(r_1) + 6\left(\sum_{t=1}^{T} (\beta_{t+1} - \beta_t) h_t + \sum_{t=1}^{T} \frac{z_t}{\beta_t}\right) \\ &= \phi_{T+1}(u) - \phi_1(r_1) + 12 \sum_{t=1}^{T} \frac{z_t}{\beta_t} \\ &\leq \gamma K \ln(T) + \frac{\beta_1(K^{1-\alpha} - 1)}{\alpha} + 12 \sum_{t=1}^{T} \frac{z_t}{\beta_t}. \end{aligned}$$

It follows that

$$R_T \leq \mathbb{E}\left[\sum_{t=1}^{T} \langle \hat{\ell}_t, q_t - u\rangle\right] + 3K$$

$$\lesssim \gamma K \ln(T) + \frac{\beta_1(K^{1-\alpha}-1)}{\alpha} + \mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\beta_t}\right].$$

Applying (23) with $S = K$, we obtain the following bounds on $\sum_{t=1}^{T} \frac{z_t}{\beta_t}$ in each environment.

IN ADVERSARIAL REGIME WITH A SELF-BOUNDING CONSTRAINT:

With $J = \log_2(T)$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\beta_t}\right] &\lesssim \sqrt{\ln(T) \sum_{t=1}^{T} \mathbb{E}[h_t z_t]} \\
&\leq \sqrt{\ln(T) \sum_{t=1}^{T} \mathbb{E}[h_t \mathbb{E}_{I_t}[z_t]]} \\
&\leq \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \mathbb{E}\left[h_t \left(\sum_{i=1}^{K} (\tilde{p}_{t,i}^{1-\alpha})(\ell_{t,i} - m_{t,i})^2\right)\right] \\
&= \frac{(6d)^{2-\alpha}}{2(1-\alpha)} \mathbb{E}\left[\left(\frac{1}{\alpha}\sum_{i=1}^{K} p_{t,i}^{\alpha} - 1\right)\left(\sum_{i=1}^{K} (\tilde{p}_{t,i}^{1-\alpha})(\ell_{t,i} - m_{t,i})^2\right)\right] \\
&\leq \frac{(6d)^{2-\alpha}}{2\alpha(1-\alpha)} \mathbb{E}\left[\left(\sum_{i=1}^{K} p_{t,i}^{\alpha} - 1\right)\left(\sum_{\ell_{t,i}\neq 0} \tilde{p}_{t,i}^{1-\alpha}\right)\right],
\end{aligned}$$

where the last inequality is from $(\ell_{t,i} - m_{t,i})^2 \leq 1$. Observe that the last bound is exactly the bound in (24), hence

$$\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\beta_t}\right] \lesssim O\left(\frac{(K-1)^{1-\alpha}K^{\alpha}\ln(T)}{\alpha(1-\alpha)\Delta_{\min}} + \sqrt{C\frac{(K-1)^{1-\alpha}K^{\alpha}\ln(T)}{\alpha(1-\alpha)\Delta_{\min}}} + \sqrt{\frac{(K-1)^{1-\alpha}K^{\alpha}}{\alpha(1-\alpha)}}\right) \tag{65}$$

holds for Optimistic FTRL as well.

IN ADVERSARIAL BANDITS:

$$\begin{aligned}\sqrt{\mathbb{E}\left[h_{\max}\sum_{t=1}^{T} z_t\right]} &\lesssim \sqrt{\frac{K^{1-\alpha}-1}{\alpha(1-\alpha)}\mathbb{E}\left[\sum_{t=1}^{T} p_{t,I_t}^{-\alpha}(\ell_{t,I_t}-m_{t,I_t})^2\right]} \\ &= \sqrt{\frac{K^{1-\alpha}-1}{\alpha(1-\alpha)}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} p_{t,i}^{1-\alpha}(\ell_{t,i}-m_{t,i})^2\right]} \\ &\leq \sqrt{\frac{(K^{1-\alpha}-1)}{\alpha(1-\alpha)}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\ell_{t,i}-m_{t,i})^2\right]},\end{aligned} \tag{66}$$

where the inequality is due to $p_{t,i}^{1-\alpha} \leq 1$.

D.3. Proof for Theorem 7

Let $\mu_t = \frac{1}{s}\sum_{s=1}^{t} \ell_s$ and

$$Q = \sum_{t=1}^{T} \|\ell_t - \mu_T\|_2^2 .$$

Using the reservoir sampling technique in Hazan and Kale (2011), we can use a prediction vector m_t satisfying $\mathbb{E}[m_t] = \mu_t$ and $\mathrm{Var}[m_t] \leq \frac{Q}{t\ln(T)}$.

REGRET ANALYSIS

Without loss of generality, assume $\ln(T) \in \mathbb{N}$ (otherwise, this increases at most a constant factor in the regret bound). For any fixed $u \in \Delta_K$, we have,

$$\sum_{t=1}^{T}\langle \ell_t, p_t - u\rangle = \sum_{t=K\ln(T)+1} \langle \ell_t, p_t - u\rangle + \sum_{t=1}^{K\ln(T)} \langle \ell_t, p_t - u\rangle \tag{67}$$

$$\leq \sum_{t=K\ln(T)+1} \langle \ell_t, p_t - u\rangle + K\ln(T) \tag{68}$$

$$= \underbrace{\sum_{t=K\ln(T)+1} \mathbb{1}\{b_t = 1\}\langle \ell_t, p_t - u\rangle}_{(A)} + \underbrace{\sum_{t=1}^{T} \mathbb{1}\{b_t = 0\}\langle \ell_t, p_t - u\rangle}_{(B)} + K\ln(T). \tag{69}$$

Recall that $b_t = 1$ indicates a reservoir sampling round where, for $t > K\ln(T)$, the sampling probability is the uniform distribution $p_t = \frac{1}{K}\mathbf{1}$, and $b_t = 0$ indicates an FTRL round. Next, we bound the expectation of A and B in the equation above. First, we have

$$\mathbb{E}[A] \leq \mathbb{E}\left[\sum_{t=K\ln(T)+1}^{T} \mathbb{1}\{b_t = 1\}\right] = \sum_{t=K\ln(T)+1}^{T} \mathbb{P}[b_t = 1] \leq \sum_{t=1}^{T} \frac{K\ln(T)}{t} \leq O(K(\ln(T))^2). \tag{70}$$

Input: $K \geq 1, T \geq 4K, \alpha \in (0,1), \beta_1 = \frac{8K}{1-\alpha}, \gamma = \max(6, 48\sqrt{\frac{\alpha}{1-\alpha}}), d = 2.$
Initialize $\mathbb{S}_i = \emptyset, \tilde{\mu}_{0,i} = 0, L_{0,i} = 0$ for $i \in [K]$
for *each round* $t = 1, \ldots, T$ **do**
  Sample $b_t \sim \mathrm{Ber}(\min(\frac{K \ln(T)}{t}, 1))$
  if $b_t = 1$ **then**
    if $t \leq K \ln(T)$ **then**
      Draw $I_t = t \mod K + 1$ and observe ℓ_{t,I_t}
      Add ℓ_{t,I_t} to the reservoir $\mathbb{S}_{I_t}$ of arm I_t
    end
    else
      Draw $I_t \sim \mathrm{Unif}([K])$ and observe ℓ_{t,I_t}
      Draw a random element by $\mathrm{Unif}(\mathbb{S}_{I_t})$ and replace it by ℓ_{t,I_t}
    end
    Update the mean estimate $\tilde{\mu}_{t,I_t}$ in the reservoir $\mathbb{S}_{I_t}$ (Hazan and Kale, 2011)
    Compute $m_t = \tilde{\mu}_t$
  end
  else
    Compute $m_t = m_{t-1}$
    Compute $q_t = \arg\min_{x \in \Delta_K} \langle m_t + L_{t-1}, x \rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K x_i^\alpha) \right) - \gamma \sum_{i=1}^K \ln(x_i)$
    Compute $p_t = \left(1 - \frac{K}{T}\right) q_t + \frac{1}{T}\mathbf{1}$
    Draw $I_t \sim p_t$ and observe ℓ_{t,I_t}
    Compute loss estimate $\hat{\ell}_{t,i} = m_{t,i} + \frac{(\ell_{t,i} - m_{t,i}) \mathbb{1}\{I_t = i\}}{p_{t,i}}$
    Update $L_{t,i} = L_{t-1,i} + \hat{\ell}_{t,i}$
    Compute $z_t = \min \left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \tilde{p}_{t,I_t}^{2-\alpha} (\hat{\ell}_{t,I_t} - m_{t,I_t})^2, \frac{\beta_t 18 d^2}{\gamma} (\ell_{t,I_t} - m_{t,I_t})^2 \right)$
    Compute $h_t = \left(\frac{1}{\alpha} (\sum_{i=1}^K p_{t,i}^\alpha - 1) \right)$
    Compute $\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}$
  end
end

Algorithm 4: SPM with Optimistic FTRL and Reservoir Sampling for losses in $[0,1]$

Next, the set of rounds with $b_t = 0$ are the Optimistic FTRL rounds; hence, we can apply (65) and (66). In the adversarial regime with a self-bounding constraint, we have

$$\begin{aligned}
\mathbb{E}[B] &\lesssim \sqrt{\ln(T) \mathbb{E}[\mathbb{1}\{b_t = 0\} h_t z_t]} \\
&\lesssim \sqrt{\frac{\ln(T)}{\alpha(1-\alpha)} \mathbb{E}\left[\mathbb{1}\{b_t = 0\} \left(\sum_{i=1}^K p_{t,i}^\alpha - 1 \right) \left(\sum_{i=1}^K \tilde{p}_{t,i}^{1-\alpha} \right) \right]} \\
&\leq \sqrt{\frac{\ln(T)}{\alpha(1-\alpha)} \mathbb{E}\left[\left(\sum_{i=1}^K p_{t,i}^\alpha - 1 \right) \left(\sum_{i=1}^K \tilde{p}_{t,i}^{1-\alpha} \right) \right]},
\end{aligned}$$

where the last inequality is from $\left(\sum_{i=1}^K p_{t,i}^\alpha - 1\right)\left(\sum_{i=1}^K \tilde{p}_{t,i}^{1-\alpha}\right) \geq 0$ in the rounds where $b_t = 1$ (in such rounds, p_t is either a one-hot vector if $t \leq K \ln T$ or $\frac{1}{K}\mathbf{1}$ if $t > K \ln T$). By (65), we obtain

$$\mathbb{E}[B] \lesssim O\left(\frac{(K-1)^{1-\alpha}K^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}} + \sqrt{C\frac{(K-1)^{1-\alpha}K^\alpha \ln(T)}{\alpha(1-\alpha)\Delta_{\min}}} + \sqrt{\frac{(K-1)^{1-\alpha}K^\alpha}{\alpha(1-\alpha)}}\right).$$

In the adversarial regime, we have

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^T \mathbb{1}\{b_t = 0\} z_t] &\lesssim \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{b_t = 0\} \sum_{i=1}^K (\ell_{t,i} - m_{t,i})^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{b_t = 0\} \sum_{i=1}^K (\ell_{t,i} - \tilde{\mu}_{t,i})^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{b_t = 0\} \|\ell_t - \tilde{\mu}_t\|_2^2\right] \\
&\leq \left(\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{b_t = 0\} \|\ell_t - \mu_t\|_2^2\right] + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{b_t = 0\} \|\tilde{\mu}_t - \mu_t\|_2^2\right]\right) \\
&\leq \left(\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{b_t = 0\} \|\ell_t - \mu_T\|_2^2\right] + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{b_t = 0\} \|\tilde{\mu}_t - \mu_t\|_2^2\right]\right) \\
&\leq \left(Q + \sum_{t=1}^T \frac{Q}{t\ln(T)}\right) \leq 3Q,
\end{aligned}$$

where the first inequality is triangle inequality, the second inequality is $\mathbb{E}\left[\sum_{t=1}^T \|\ell_t - \mu_t\|_2^2\right] \leq \mathbb{E}\left[\sum_{t=1}^T \|\ell_t - \mu_T\|_2^2\right]$ by Lemma 10 in Hazan and Kale (2011), the third inequality is by Lemma 11 in Hazan and Kale (2011), and the last inequality is due to $\sum_{t=1}^T \frac{1}{t} \leq \ln(T) + 1$. Overall, the regret for adversarial bandits is

$$R_T \lesssim \sqrt{\frac{(K^{1-\alpha} - 1)Q}{\alpha(1-\alpha)}}.$$

Appendix E. Proofs for Section 5

We have

$$\mathbb{E}[\sum_{t=1}^T \langle \ell_t, p_t\rangle - u] = \mathbb{E}[\sum_{t=1}^T \langle \ell_t, q_t - u + \frac{\mathbf{1} - Kq_t}{T}\rangle]$$

$$\leq \mathbb{E}[\sum_{t=1}^{T}\langle \ell_t, q_t - u\rangle] + K$$
$$= \mathbb{E}[\sum_{t=1}^{T}\langle \hat{\ell}_t, q_t - u\rangle] + K.$$

Furthermore, let

$$r_{t+1} = \arg\min_{x\in\Delta_K}\langle L_{t-1} + \hat{\ell}_t, x\rangle + \sum_{i=1}^{K}\beta_{t,i}\left(\frac{1}{\alpha}(-x_i^\alpha) + (1-x_i)\ln(1-x_i) + x_i\right) - \gamma\sum_{i=1}^{K}\ln(x_i).$$

Then,

$$\begin{aligned}
&\sum_{t=1}^{T}\langle \hat{\ell}_t, q_t - u\rangle \\
&\leq \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T}\sum_{i=1}^{K}(\beta_{t+1,i} - \beta_{t,i})\left(\frac{p_{t+1,i}^\alpha}{\alpha} + (p_{t+1,i}-1)\ln(1-p_{t+1,i}) - p_{t+1,i}\right) + \sum_{t=1}^{T}\frac{z_{t,I_t}}{\beta_{t,I_t}} \\
&\leq \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T}\frac{2}{\alpha}(\beta_{t+1,I_t} - \beta_{t,I_t})p_{t+1,I_t}^\alpha + \sum_{t=1}^{T}\frac{z_{t,I_t}}{\beta_{t,I_t}} \\
&\leq \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T}\frac{12}{\alpha}(\beta_{t+1,I_t} - \beta_{t,I_t})p_{t,I_t}^\alpha + \sum_{t=1}^{T}\frac{z_{t,I_t}}{\beta_{t,I_t}} \\
&= \phi_{T+1}(u) - \phi_1(r_1) + \sum_{t=1}^{T}12(\beta_{t+1,I_t} - \beta_{t,I_t})h_{t,I_t} + \sum_{t=1}^{T}\frac{z_{t,I_t}}{\beta_{t,I_t}} \\
&= \phi_{T+1}(u) - \phi_1(r_1) + 13\sum_{t=1}^{T}\frac{z_{t,I_t}}{\beta_{t,I_t}} \\
&= \phi_{T+1}(u) - \phi_1(r_1) + 13\sum_{i=1}^{K}\sum_{t=1}^{T}\frac{z_{t,i}}{\beta_{t,i}},
\end{aligned}$$

where the first inequality is from $p_{t+1,i} \geq 0$ and Lemma 30, the second inequality is from $p_{t+1,I_t}^\alpha \leq (6p_{t,I_t})^\alpha \leq 6p_{t,I_t}^\alpha$ and the last equality is $z_{t,i} = 0$ for all $i \neq I_t$.

From the previous section, we have for all $i \in [K]$,

$$\sum_{t=1}^{T}\frac{z_{t,i}}{\beta_{t,i}} \lesssim \min\left\{\sqrt{\mathbb{E}\left[\ln(T)\sum_{t=1}^{T}h_{t,i}z_{t,i}\right]} + \sqrt{\frac{1}{T}\mathbb{E}\left[h_{i,\max}\sum_{t=1}^{T}z_{t,i}\right]}, \sqrt{\mathbb{E}\left[h_{i,\max}\sum_{t=1}^{T}z_{t,i}\right]}\right\}. \tag{71}$$

First, we have $h_{i,\max} = \max_t h_{t,i} = \frac{1}{\alpha}\max_t p_{t,i}^\alpha \leq \frac{1}{\alpha}$. In addition, $\mathbb{E}_{I_t}[z_{t,i}] \leq \mathbb{E}_{I_t}[\mathbb{1}\{I_t = i\}p_{t,i}^{-\alpha}(\ell_{t,i} - m_{t,i})^2] = p_{t,i}^{1-\alpha}(\ell_{t,i} - m_{t,i})^2 \leq 1$. Therefore, the sum $\frac{1}{T}\mathbb{E}\left[h_{i,\max}\sum_{t=1}^{T}z_{t,i}\right]$ is

bounded by $\frac{1}{\alpha}$. We can simplify (71) by

$$\sum_{t=1}^{T} \frac{z_{t,i}}{\beta_{t,i}} \lesssim \min\left\{ \sqrt{\ln(T)\mathbb{E}\left[\sum_{t=1}^{T} h_{t,i} z_{t,i}\right]} + \sqrt{\frac{1}{\alpha}}, \sqrt{\mathbb{E}\left[\sum_{t=1}^{T} z_{t,i}\right]} \right\}. \tag{72}$$

A Bound for Stochastic Bandits from $\sqrt{\ln(T)\mathbb{E}\left[\sum_{t=1}^{T} h_{t,i} z_{t,i}\right]}$

We have

$$h_{t,i} z_{t,i} \lesssim \mathbb{1}\{I_t = i\}(\ell_{t,i} - m_{t,i})^2 p_{t,i}^{\alpha} \min\left(\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min\left\{ p_{t,I_t}^{-\alpha}, \frac{1 - p_{t,I_t}}{p_{t,I_t}^2} \right\} \right).$$

Therefore, by Lemma 34,

$$\mathbb{E}_{I_t}[h_{t,i} z_{t,i}] \lesssim \frac{1}{1-\alpha} \tilde{p}_{t,i} (\ell_{t,i} - m_{t,i})^2.$$

Denote $P_i = \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}\{I_t = i\}]$. Bounding $(\ell_{t,i} - m_{t,i})^2 \le 1$ for any $m_t \in [0,1]^K$, we obtain

$$\begin{aligned}
\mathrm{Reg}_T &\lesssim \sum_{i=1}^{K} \sqrt{\mathbb{E}\left[\ln(T) \sum_{t=1}^{T} h_{t,i} z_{t,i}\right]} \\
&\lesssim \sqrt{\frac{\ln(T)}{\alpha(1-\alpha)}} \left(\sum_{i \neq i^*} \sqrt{P_i} + \sqrt{\left(\sum_{i \neq i^*} P_i\right)} \right) + K \ln(T) \\
&\le \sqrt{\frac{\ln(T)}{\alpha(1-\alpha)}} \left(\sum_{i \neq i^*} \sqrt{P_i} + \frac{1}{\sqrt{K-1}} \sum_{i \neq i^*} \sqrt{P_i} \right) + K \ln(T) \\
&\lesssim \sqrt{\frac{\ln(T)}{\alpha(1-\alpha)}} \left(\sum_{i \neq i^*} \sqrt{P_i} \right) + K \ln(T).
\end{aligned}$$

Similar to Ito et al. (2022), by using $\mathrm{Reg}_T = 2\mathrm{Reg}_T - \mathrm{Reg}_T$ and $2\sqrt{ax} - bx \le \frac{a}{b}$ for $a = \frac{\ln(T)}{\alpha(1-\alpha)}$, $b = \Delta_i$ and $x = P_i$, we obtain (note that we set $\alpha = \frac{1}{2}$)

$$\mathrm{Reg}_T \lesssim \frac{1}{\alpha(1-\alpha)} \sum_{i \neq i^*} \frac{\ln(T)}{\Delta_i}.$$

Bounds for Adversarial Bandits

Fix $i \in [K]$. Since $\tilde{p}_{t,i} \le p_{t,i}$, we have $h_{t,i} z_{t,i} \lesssim \mathbb{1}\{I_t = i\}(\ell_{t,i} - m_{t,i})^2$. Therefore, by setting $m_{t,i}$ to be the output of an online learning algorithm with fully-observable squared loss as in Ito et al.

(2022), i.e.,

$$m_{t,i} = \frac{1}{1 + \sum_{s=1}^{t-1} \mathbb{1}\{I_s = i\}} \left(\frac{1}{2} + \sum_{s=1}^{t-1} \mathbb{1}\{I_s = i\} \ell_{t,i} \right)$$

and then applying their Lemma 3, we obtain for any fixed $m^* \in [0,1]^K$,

$$\sum_{t=1}^{T} \mathbb{1}\{I_t = i\}(\ell_{t,i} - m_{t,i})^2 \lesssim \sum_{t=1}^{T} \mathbb{1}\{I_t = i\}(\ell_{t,i} - m_i^*)^2 + \ln(1 + \sum_{t=1}^{T} \mathbb{1}\{I_t = i\}).$$

As already shown in Ito et al. (2022), for each appropriately chosen m^*, we would recover the data-dependent bounds of order $\sqrt{KQ_\infty \ln(T)}$ (with $m^* \in \arg\min_{\bar{\ell} \in \mathbb{R}^K} \sum_{t=1}^{T} \left\| \ell_t - \bar{\ell} \right\|_2^2$), $\sqrt{KL^* \ln(T)}$ (with $m^* = \mathbf{0}$) and $\sqrt{K(T - L^*) \ln(T)}$ (with $m^* = \mathbf{1}$).

On the other hand, from the quantity $\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} z_{t,i}\right]}$ and Jensen's inequality, we obtain

$$\begin{aligned}
\mathrm{Reg}_T &\lesssim \sqrt{K \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{K} z_{t,i} \right]} \\
&\lesssim \sqrt{K \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{K} p_{t,i}^{1-\alpha} (\ell_{t,i} - m_{t,i})^2 \right]} \\
&\lesssim \sqrt{K \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{K} p_{t,i}^{1-\alpha} \right]} \\
&\leq \sqrt{K \mathbb{E}\left[\sum_{t=1}^{T} K^{\alpha} \right]} \\
&= \sqrt{K^{1+\alpha} T} = K^{\frac{1}{4}} \sqrt{KT} \qquad (\alpha = 1/2),
\end{aligned}$$

which grows with $\sqrt{T}$ in the worst-case.

E.1. Stability Proofs

In this section, we define the following function

$$g(x) = x_i^{\alpha - 1} + \ln(1 - x). \tag{73}$$

Note that g is decreasing. In addition, let $d_f(y, x) = f(y) - f(x) - f'(x)(y - x)$ denote the Bregman divergence associated with a one-dimensional strictly convex function $f : \mathbb{R} \to \mathbb{R}$. Note that $d_f(y, x) \geq 0$ for all $x, y \in \mathbb{R}$.

Lemma 26 *For any $L \in \mathbb{R}^K, \beta \in R_+^K, \gamma > 0$ and $d \geq 2$, let*

$$x = \arg\min_{p \in \Delta_K} \langle L, p \rangle + \sum_{i=1}^{K} \beta_i \left(\frac{-p_i^\alpha}{\alpha} + (1-p_i)\ln(1-p_i) + p_i \right) - \gamma \sum_{i=1}^{K} \ln(p_i)$$
$$y = \arg\min_{p \in \Delta_K} \langle L, p \rangle + \sum_{i=1}^{K} \beta'_i \left(\frac{-p_i^\alpha}{\alpha} + (1-p_i)\ln(1-p_i) + p_i \right) - \gamma \sum_{i=1}^{K} \ln(p_i).$$

If $0 \leq \beta'_1 - \beta_1 \leq \left(1 - \frac{1}{d}\right) \gamma x_1^{-\alpha}$ and $\beta'_i = \beta_i$ for $i > 1$, then $y_1 \leq d x_1$.

Proof If $dx_1 \geq 1$ then $y_1 \leq 1 \leq dx_1$ trivially. Hence, we assume $dx_1 \leq 1$. By the Lagrange multiplier method, we have for $i = 2, \ldots, K$ and some $\lambda, \lambda' \in \mathbb{R}$,

$$L_i - \beta_i(x_i^{\alpha-1} + \ln(1-x_i)) - \frac{\gamma}{x_i} = \lambda,$$
$$L_i - \beta_i(x_i^{\alpha-1} + \ln(1-y_i)) - \frac{\gamma}{y_i} = \lambda'.$$

Similary, for $i = 1$, we have

$$L_1 - \beta_1(x_1^{\alpha-1} + \ln(1-x_1)) - \frac{\gamma}{x_1} = \lambda,$$
$$L_1 - \beta'_1(x_1^{\alpha-1} + \ln(1-y_1)) - \frac{\gamma}{y_1} = \lambda'.$$

Taking $Z = \lambda' - \lambda$ over all K pairs of equations, we obtain

$$\beta_i(g(x_i) - g(y_i)) + \gamma \left(\frac{1}{x_i} - \frac{1}{y_i} \right) = Z \tag{74}$$
$$\beta_1 g(x_1) - \beta'_1 g(y_1) + \gamma \left(\frac{1}{x_1} - \frac{1}{y_1} \right) = Z. \tag{75}$$

If $Z \geq 0$, then since $\beta_i > 0$ and both $g(x)$ and $\frac{\gamma}{x}$ are decreasing, we have $y_i \geq x_i$ for all $i \neq 1$. This straightforwardly implies that $y_1 \leq x_1$. Thus, we focus on the case $Z < 0$. In this case, we have $y_i < x_i$ for all $i \neq 1$ and $y_1 > x_1$. We consider two cases:

- If $g(x_1) \leq 0$: from $y_1 \geq x_1$, we have $g(y_1) \leq g(x_1) \leq 0$. Hence, from $0 < \beta_1 \leq \beta'_1$, we obtain

  $$\beta'_1 g(y_1) \leq \beta_1 g(y_1) \leq \beta_1 g(x_1).$$

  This implies that $0 > Z = \beta_1 g(x_1) - \beta'_1 g(y_1) + \gamma \left(\frac{1}{x_1} - \frac{1}{y_1} \right) \geq 0$, a contradiction.

- If $g(x_1) > 0$: in this case, (75) and $Z < 0$ implies $\beta'_1 g(y_1) = \beta_1 g(x_1) + \gamma \left(\frac{1}{x_1} - \frac{1}{y_1} \right) - Z > 0$. Furthermore, by re-arranging, we obtain

  $$\beta'_1 g(y_1) + \frac{\gamma}{y_1} \geq \beta_1 g(x_1) + \frac{\gamma}{x_1}$$

$$
\begin{aligned}
&\geq \left(\beta_1' - \left(1 - \frac{1}{d}\right)\gamma x_1^{-\alpha}\right)(x_1^{\alpha-1} + \ln(1-x_1)) + \frac{\gamma}{x_1} \\
&= \beta_1'(x_1^{\alpha-1} + \ln(1-x_1)) - \left(1 - \frac{1}{d}\right)\gamma x^{-1} - \left(1 - \frac{1}{d}\right)\gamma x_1^{-\alpha}\ln(1-x_1) + \frac{\gamma}{x_1} \\
&= \beta_1'(x_1^{\alpha-1} + \ln(1-x_1)) + \frac{\gamma}{dx_1} - \left(1 - \frac{1}{d}\right)\gamma x_1^{-\alpha}\ln(1-x_1) \\
&\geq \beta_1'(x_1^{\alpha-1} + \ln(1-x_1)) + \frac{\gamma}{dx_1} \\
&\geq \beta_1'((dx_1)^{\alpha-1} + \ln(1-dx_1)) + \frac{\gamma}{dx_1} \\
&= \beta_1' g(dx_1) + \frac{\gamma}{dx_1},
\end{aligned}
$$

where the second inequality is from $\beta_1 \geq \beta_1' - \left(1 - \frac{1}{d}\right)\gamma x_1^{-\alpha}$, the third inequality is due to $\ln(1-x_1) < 0$ and the last inequality is from $d \geq 2 > 1$. Since $\beta g(x) + \frac{\gamma}{x}$ is decreasing for all $\beta > 0, \gamma > 0$, we conclude that $y_1 \leq dx_1$.

■

Lemma 27 *For any $L \in \mathbb{R}^K, \beta \in R_+^K, \gamma > 0$ and $h \in [-1, 1]$, let*

$$
\begin{aligned}
x &= \arg\min_{p \in \Delta_K} \langle L, p\rangle + \sum_{i=1}^K \beta_i \left(\frac{-p_i^\alpha}{\alpha} + (1-p_i)\ln(1-p_i) + p_i\right) - \gamma \sum_{i=1}^K \ln(p_i) \\
y &= \arg\min_{p \in \Delta_K} \langle L + \frac{h}{x_1'}, p\rangle + \sum_{i=1}^K \beta_i \left(\frac{-p_i^\alpha}{\alpha} + (1-p_i)\ln(1-p_i) + p_i\right) - \gamma \sum_{i=1}^K \ln(p_i),
\end{aligned}
$$

where $4x_1' \geq x_1$. If $\gamma \geq 6$ then $y_1 \leq 3x_1$.

Proof Using the Lagrange multiplier method, we have the following equalities that hold for some $Z \in \mathbb{R}$,

$$
\beta_1\left(g(y_1) - g(x_1)\right) + \gamma\left(\frac{1}{y_1} - \frac{1}{x_1}\right) = Z + \frac{h}{x_1'} \tag{76}
$$

and for all $i \neq 1$,

$$
\beta_i\left(g(y_i) - g(x_i)\right) + \gamma\left(\frac{1}{y_i} - \frac{1}{x_i}\right) = Z. \tag{77}
$$

First, we show that Z and $y_1 - x_1$ has the opposite sign to h. We consider two cases:

- If $Z \geq 0$ then from (77) and the monotonic decreasing property of $\beta g(x) + \frac{\gamma}{x}$, we have $y_i \leq x_i$ and this leads to $y_1 \geq x_1$. Combining $y_1 \geq x_1$ and (76), we have $Z + \frac{h}{x_1'} \leq 0$. Since $Z \geq 0$, this implies $h \leq 0$.

- If $Z \le 0$ then by the same argument, we have $y_i \ge x_i$ and $y_1 \le x_1$. Therefore, $Z + \frac{h}{x_1'} \ge 0$. Due to $Z \le 0$, we must have $h \ge 0$.

In both cases, we have $Zh \le 0$ and $Z(y_1 - x_1) \ge 0$. It follows that if $h \ge 0$ then we have $y_1 \le x_1 \le 3x_1$. If $h < 0$ then $y_1 \ge x_1$, and by rearranging (76), we obtain

$$\begin{aligned}\frac{4}{x_1} \ge -\frac{h}{x_1'} &= \underbrace{Z}_{\ge 0} + \underbrace{\gamma\left(\frac{1}{x_1} - \frac{1}{y_1}\right)}_{\ge 0} + \underbrace{\beta(g(x_1) - g(y_1))}_{\ge 0} \\ &\ge \gamma\left(\frac{1}{x_1} - \frac{1}{y_1}\right) \\ &\ge 6\left(\frac{1}{x_1} - \frac{1}{y_1}\right),\end{aligned}$$

where the last inequality is due to $\gamma \ge 6$. This implies that $\frac{3}{y_1} \ge \frac{1}{x_1}$, thus $y_1 \le 3x_1$. ■

By combining Lemma 26 and Lemma 27, we obtain the following corollary. The proof of this corollary is nearly identical to that of Corollary 20.

Corollary 28 *For any $t \in [T]$, Algorithm 2 guarantees that*

$$r_{t+1,I_t} \le 3dq_{t,I_t}.$$

Lemma 29 *For all $t \in [T]$, Algorithm 2 guarantees that*

$$\langle \hat{\ell}_t - m_t, q_t - r_{t+1} \rangle - D_t(r_{t+1}, q_t) \le \frac{z_{t,I_t}}{\beta_{t,I_t}},$$

where

$$z_{t,I_t} = (\ell_{t,I_t} - m_{t,I_t})^2 \min\left\{\frac{(6d)^{2-\alpha}}{2(1-\alpha)} \min\left\{p_{t,I_t}^{-\alpha}, \frac{1 - p_{t,I_t}}{p_{t,I_t}^2}\right\}, \frac{\beta_{t,I_t} 18d^2}{\gamma}\right\}.$$

Proof Let

$$\begin{aligned}f_1(x) &= \frac{-x^\alpha}{\alpha}, \\ f_2(x) &= (1-x)\ln(1-x) + x, \\ f_3(x) &= -\ln(x).\end{aligned}$$

Since $\phi_t(x) = \sum_{i=1}^K \left(\beta_{t,i}(f_1(x_i) + f_2(x_i)) + \gamma f_3(x_i)\right)$, we have

$$\begin{aligned}D_t(r_{t+1}, q_t) &= \sum_{i=1}^K \left(\beta_{t,i}(d_{f_1}(r_{t+1,i}, q_{t,i}) + d_{f_2}(r_{t+1,i}, q_{t,i})) + \gamma d_{f_3}(r_{t+1,i}, q_{t,i})\right) \\ &\ge \beta_{t,I_t}(d_{f_1}(r_{t+1,I_t}, q_{t,I_t}) + d_{f_2}(r_{t+1,I_t}, q_{t,I_t})) + \gamma d_{f_3}(r_{t+1,I_t}, q_{t,I_t}).\end{aligned}$$

Furthermore, as $\hat{\ell}_{t,i} - m_{t,i} = 0$ for all $i \neq I_t$, we have

$$\begin{aligned}\langle \hat{\ell}_t - m_t, q_t - r_{t+1} \rangle - D_t(r_{t+1}, q_t) &= \frac{\ell_{t,I_t} - m_{t,I_t}}{p_{t,I_t}}(r_{t+1,I_t} - q_{t,I_t}) - D_t(r_{t+1}, q_t) \\ &\leq \min(A, B, C),\end{aligned}$$

where

$$\begin{aligned}A &= \frac{\ell_{t,I_t} - m_{t,I_t}}{p_{t,I_t}}(r_{t+1,I_t} - q_{t,I_t}) - \beta_{t,I_t} d_{f_1}(r_{t+1,I_t}, q_{t,I_t}), \\ B &= \frac{\ell_{t,I_t} - m_{t,I_t}}{p_{t,I_t}}(r_{t+1,I_t} - q_{t,I_t}) - \beta_{t,I_t} d_{f_2}(r_{t+1,I_t}, q_{t,I_t}), \\ C &= \frac{\ell_{t,I_t} - m_{t,I_t}}{p_{t,I_t}}(r_{t+1,I_t} - q_{t,I_t}) - \gamma d_{f_3}(r_{t+1,I_t}, q_{t,I_t}).\end{aligned}$$

Here, we used $x - (a + b + c) \leq \min(x - a, x - b, x - c)$ for $a, b, c \geq 0$.

Note that $\ell_{t,I_t} - m_{t,I_t} \in [-1, 1]$ for $0 \leq \ell_{t,I_t}, m_{t,I_t} \leq 1$. By Corollary 28 and the fact that $r_{t+1} \in \Delta_K$, we have $0 \leq r_{t+1,I_t} \leq 3dq_{t,I_t}$. Combining this with Lemma 32, we have

$$\begin{aligned}\min(A, B) &\leq \frac{(3d)^{2-\alpha}(\ell_{t,I_t} - m_{t,I_t})^2}{\beta_{t,I_t} p_{t,I_t}^2} \min\left\{ \frac{q_{t,I_t}^{2-\alpha}}{2(1-\alpha)}, 1 - q_{t,I_t} \right\} \\ &\leq \frac{(6d)^{2-\alpha}(\ell_{t,I_t} - m_{t,I_t})^2}{\beta_{t,I_t} p_{t,I_t}^2} \min\left\{ \frac{p_{t,I_t}^{2-\alpha}}{(1-\alpha)}, 2(1 - p_{t,I_t}) \right\} \\ &\leq \frac{(6d)^{2-\alpha}(\ell_{t,I_t} - m_{t,I_t})^2}{2(1-\alpha)\beta_{t,I_t}} \min\left\{ p_{t,I_t}^{-\alpha}, \frac{(1 - p_{t,I_t})}{p_{t,I_t}^2} \right\},\end{aligned} \tag{78}$$

where the second inequality is due to $q_{t,I_t} \leq 2p_{t,I_t}$ and $1 - q_{t,I_t} \leq 2(1 - p_{t,I_t})$ by Lemma 33 and the last inequality is $1 - \alpha \leq 1$.

The second-order derivative of $\gamma f_3(x)$ is $\frac{\gamma}{x^2}$. Therefore, by Lemma 31, we have

$$C \leq \frac{(\ell_{t,I_t} - m_{t,I_t})^2 v^2}{2\gamma p_{t,I_t}^2} \leq \frac{(\ell_{t,I_t} - m_{t,I_t})^2 18d^2}{\gamma}, \tag{79}$$

where $v \leq 3dq_{t,I_t} \leq 6dp_{t,I_t}$ is a point between r_{t+1,I_t} and q_{t,I_t}.

Combining (78) and (79) implies the desired bound. ■

E.2. Technical Lemmas

Lemma 30 *For any $\alpha \in [0, 1]$ and $x \in [0, 1]$,*

$$x^\alpha \geq (x - 1)\ln(1 - x). \tag{80}$$

Proof For $\alpha \in [0,1]$, we have $x^\alpha \geq x$. Let

$$h(x) = x + (1-x)\ln(1-x). \tag{81}$$

Its derivative is $h'(x) = 1 - \ln(1-x) - 1 = -\ln(1-x) \geq 0$ for all $x \in [0,1]$. Therefore, $h(x) \geq h(0) = 0$ for all $x \in [0,1]$. We conclude that

$$x^\alpha - (x-1)\ln(1-x) \geq x - (x-1)\ln(1-x) = h(x) \geq 0. \tag{82}$$

■

Recall that $d_f(y,x) = f(y) - f(x) - f'(x)(y-x)$ denote the Bregman divergence associated with a one-dimensional strictly convex function $f : \mathbb{R} \to \mathbb{R}$. The following lemma is essentially a one-dimensional local-norm analysis of FTRL, whose proof can be found in standard literature. We provide a proof here for completeness.

Lemma 31 *For any $a \in \mathbb{R}, x, y \in (0,1)$ and strictly convex function $f : (0,1) \to \mathbb{R}$, we have*

$$a(x-y) - d_f(y,x) \leq \frac{1}{2}\frac{a^2}{f''(z)}$$

for some z between x and y.

Proof The inequality trivally holds when $x = y$, hence we assume $x \neq y$. By Taylor's theorem, we have $d_f(y,x) = \frac{f''(z)}{2}(x-y)^2$ for some z between x and y. Note that the strict convexity of f implies $f''(z) > 0$. We have

$$\begin{aligned} a(x-y) - d_f(x,y) &= a(x-y) - \frac{f''(z)}{2}(x-y)^2 \\ &= \frac{1}{2}\left(- \left((x-y)\sqrt{f''(z)} - \frac{a}{\sqrt{f''(z)}} \right)^2 + \frac{a^2}{f''(z)} \right) \leq \frac{a^2}{2f''(z)}. \end{aligned}$$

■

Next, the following two lemma establish the foundation for choosing $z_{t,i}$ in Algorithm 2.

Lemma 32 *Let $\alpha \in (0,1), \beta > \frac{4}{1-\alpha}$ be fixed and*

$$\begin{aligned} f_1(x) &= \left(\frac{-x^\alpha}{\alpha} \right), \\ f_2(x) &= ((1-x)\ln(1-x) + x). \end{aligned}$$

For any $d \geq 1, h \in [-1,1], q \in (0,1]$ and $p \geq \frac{q}{2}$, we have

$$\begin{aligned} &\min\left\{ \max_{0 \leq u \leq dq} \left(\frac{h}{p}(q-u) - \beta d_{f_1}(u,q) \right), \max_{u \in \mathbb{R}} \left(\frac{h}{p}(q-u) - \beta d_{f_2}(u,q) \right) \right\} \\ &\leq \frac{d^{2-\alpha}h^2}{\beta p^2} \min\left\{ \frac{q^{2-\alpha}}{2(1-\alpha)}, 1-q \right\}. \end{aligned}$$

Proof First, we bound $\max_{0\le u\le dq}\left(\frac{h}{p}(q-u)-\beta d_{f_1}(u,q)\right)$. For any $u\ge 0$, by Lemma 31, we have for some v between q and u:

$$\left(\frac{h}{p}(q-u)-\beta d_{f_1}(u,q)\right) \le \frac{1}{2}\frac{h^2}{p^2}\frac{v^{2-\alpha}}{\beta(1-\alpha)} \tag{83}$$

$$\le \frac{h^2}{\beta p^2}\frac{d^{2-\alpha}q^{2-\alpha}}{2(1-\alpha)}, \tag{84}$$

where we used the fact that the second-order derivative of $\beta f_1(v)$ is $\beta(1-\alpha)v^{\alpha-2}$ and $v\le \max(q,u)\le dq$. It follows that

$$\max_{0\le u\le dq}\left(\frac{h}{p}(q-u)-\beta d_{f_1}(u,q)\right) \le \frac{h^2}{\beta p^2}\frac{d^{2-\alpha}q^{2-\alpha}}{2(1-\alpha)}. \tag{85}$$

Next, using Lemma 5 in Ito et al. (2022), we have

$$\max_{u\in\mathbb{R}}\left(\frac{h}{p}(q-u)-\beta d_{f_2}(u,q)\right) = \beta\max_{u\in\mathbb{R}}\left(\frac{h}{\beta p}(q-u)-d_{f_2}(u,q)\right)$$

$$= \beta(1-q)\left(\exp\left(\frac{h^2}{\beta^2p^2}\right)-\frac{h}{\beta p}-1\right).$$

We consider two cases:

- If $\beta p\ge 1$: in this case, we have $\frac{h}{\beta p}\le 1$ for any $h\in[-1,1]$. From the inequality $\exp(a)-a-1\le a^2$ for $a\le 1$, we have $\exp\left(\frac{h^2}{\beta^2p^2}\right)-\frac{h}{\beta p}-1\le\frac{h^2}{\beta^2p^2}$. Therefore,

  $$\max_{u\in\mathbb{R}}\left(\frac{h}{p}(q-u)-\beta d_{f_2}(u,q)\right)\le(1-q)\frac{h^2}{\beta p^2}.$$

  This implies that

  $$\min\left\{\max_{0\le u\le dq}\left(\frac{h}{p}(q-u)-\beta d_{f_1}(u,q)\right),\max_{u\in\mathbb{R}}\left(\frac{h}{p}(q-u)-\beta d_{f_2}(u,q)\right)\right\}$$
  $$\le\frac{(d)^{2-\alpha}h^2}{\beta p^2}\min\left\{\frac{q^{2-\alpha}}{2(1-\alpha)},1-q\right\}.$$

- If $\beta p<1$: in this case, we have $q\le 2p\le\frac{2}{\beta}\le\frac{1-\alpha}{2}$. This implies that $\frac{q}{1-\alpha}\le\frac{1}{2}$ and also $q\le\frac{1}{2}$. Combining this with $q^{1-\alpha}\le 1$, we obtain

  $$\frac{q^{2-\alpha}}{2(1-\alpha)}\le\frac{q}{2(1-\alpha)}\le\frac{1}{4}\le 1-q.$$

  It follows that by (85),

  $$\max_{0\le u\le dq}\left(\frac{h}{p}(q-u)-\beta d_{f_1}(u,q)\right)\le\frac{h^2}{\beta p^2}\frac{(d)^{2-\alpha}q^{2-\alpha}}{2(1-\alpha)}$$
  $$=\frac{(d)^{2-\alpha}h^2}{\beta p^2}\min\left\{\frac{q^{2-\alpha}}{2(1-\alpha)},1-q\right\}.$$

In both cases, we have

$$\min\left\{\max_{0\le u\le dq}\left(\frac{h}{p}(q-u)-\beta d_{f_1}(u,q)\right),\max_{u\in\mathbb{R}}\left(\frac{h}{p}(q-u)-\beta d_{f_2}(u,q)\right)\right\}$$
$$\le\frac{(d)^{2-\alpha}h^2}{\beta p^2}\min\left\{\frac{q^{2-\alpha}}{2(1-\alpha)},1-q\right\}.$$

■

Lemma 33 *For any $K\ge 3, T\ge 4K$ and $q\in[0,1]$, let*

$$p=\left(1-\frac{K}{T}\right)q+\frac{1}{T}.$$

Then, we have $1-q\le 2(1-p)$.

Proof The desired inequality is equivalent to $2p-q\le 1$. By the definition of p, we have

$$\begin{aligned}2p-q&=2\left(1-\frac{K}{T}\right)q+\frac{2}{T}-q\\&=\left(1-\frac{2K}{T}\right)q+\frac{2}{T}\\&\le\left(1-\frac{2K}{T}\right)+\frac{2}{T}\\&\le 1.\end{aligned}$$

■

Lemma 34 *Fix an index $i\in[K]$ and let $p\in\Delta_K$ be an arbitrary vector in Δ_K. Let $\alpha\in(0,1)$ be a constant and $I\sim p$ be a random variable distributed according to p. We have*

$$\mathbb{E}_{I\sim p}\left[\mathbb{1}\{I=i\}p_i^{\alpha}\min\left(p_i^{-\alpha},\frac{1-p_i}{p_i^2}\right)\right]\le 2\min(p_i,1-p_i).$$

Proof By the definition of I, the left-hand side is equal to

$$\begin{aligned}\mathbb{E}_{I\sim p}\left[\mathbb{1}\{I=i\}p_i^{\alpha}\min\left(p_i^{-\alpha},\frac{1-p_i}{p_i^2}\right)\right]&=p_i^{1+\alpha}\min\left(p_i^{-\alpha},\frac{1-p_i}{p_i^2}\right)\\&=\min\left(p_i,\frac{(1-p_i)}{p_i^{1-\alpha}}\right).\end{aligned}$$

We consider two cases: $p_i\le\frac{1}{2}$ and $p_i>\frac{1}{2}$.

- If $p_i \leq \frac{1}{2}$: since $p_i^{1-\alpha} \leq 1$, we have $\frac{1-p_i}{p_i^{1-\alpha}} \geq 1 - p_i \geq p_i$. Hence, $\min\left(p_i, \frac{(1-p_i)}{p_i^{1-\alpha}}\right) = p_i \leq 2p_i = 2\min(p_i, 1-p_i)$.
- If $p_i > \frac{1}{2}$: we then have $\frac{(1-p_i)}{p_i^{1-\alpha}} \leq \frac{1-p_i}{p_i} \leq 2(1-p_i)$. Therefore, $\min\left(p_i, \frac{(1-p_i)}{p_i^{1-\alpha}}\right) \leq \min(p_i, 2(1-p_i)) \leq 2\min(p_i, 1-p_i)$.

■

Appendix F. SPM for Adversarial Sleeping Bandits

Intuitively, the sparsity constraint $\|\ell_t\|_0 \leq S$ indicates that there are at most S arms containing non-trivial information in each round, however the learner does not know the arms with non-trivial information. In this sense, sparse bandits is conceptually more difficult than adversarial sleeping bandits (Kleinberg et al., 2010), where in each round t the learner is given, by an adversary, a set $\mathbb{A}_t \subseteq [K]$ of active arms to choose from. Note that the learner is not allowed to choose an arm in $[K] \setminus \mathbb{A}_t$. The performance of the learner is measured by its per-action regret

$$R_{T,a} = \sum_{t=1}^{T} \mathbb{1}\{a \in \mathbb{A}_t\}(\ell_{t,I_t} - \ell_{t,a}).$$

A natural question is whether Algorithm 1 can be extended to this adversarial sleeping bandits setting. The following theorem answers this question in the positive.

Theorem 35 *For any $K \geq 4, T \geq 4k$, Algorithm 5 (in Appendix F) guarantees that for all $a \in [K]$,*

$$\mathbb{E}[R_{T,a}] \leq O\left(\sqrt{\frac{(K^{1-\alpha}-1)(\max_{t\in[T]}|\mathbb{A}_t|)^\alpha}{\alpha(1-\alpha)}T}\right),$$

Our Algorithm 5 is a combination of Algorithm 1 and the SB-EXP3 algorithm in Nguyen and Mehta (2024). More specifically, Algorithm 5 uses the estimated cumulative *regret* (instead of losses) to compute q_t in the FTRL update. Then, the sampling probability vector p_t is obtained by a filtering step $p_{t,i} = \frac{q_{t,i}\mathbb{1}\{i\in\mathbb{A}_t\}}{\sum_{j=1}^K q_{t,j}\mathbb{1}\{j\in\mathbb{A}_t\}}$. While the bound in Theorem 35 is of the same order as in Nguyen and Mehta (2024, Theorem 2), it has the advantage of not requiring the knowledge of $\max_t|\mathbb{A}_t|$ in advance nor any complicated two-level doubling trick.

Algorithm: We use the same regularization function in Algorithm 1,

$$\begin{aligned}\Phi_t(p) &= \beta_t\psi_{TE}(p) - \gamma\psi_{LB}(p)\\ &= \frac{\beta_t}{\alpha}\left(1 - \sum_{i=1}^K p_i^\alpha\right) - \gamma\sum_{i=1}^K \ln(p_i).\end{aligned}$$

Input: $K \geq 1, T \geq 4K, \alpha \in (0,1), \beta_1 = \frac{8K}{1-\alpha}, \gamma = \max\left(6, 48\sqrt{\frac{\alpha}{1-\alpha}}\right), d = 2.$

Initialize $R_{0,i} = 0$ for $i \in [K]$

for *each round* $t = 1, \ldots, T$ **do**

  The adversary reveals the set of active arms $\mathbb{A}_t$

  Compute $q_t = \arg\min_{x \in \Delta_K} \langle -R_{t-1}, x \rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K x_i^\alpha)\right) - \gamma \sum_{i=1}^K \ln(x_i)$

  for *arm* $i \in [K]$ **do**

    Compute $p_{t,i} = \frac{q_{t,i} \mathbb{1}\{i \in \mathbb{A}_t\}}{\sum_{j=1}^K q_{t,j} \mathbb{1}\{j \in \mathbb{A}_t\}}$

  end

  Draw $I_t \sim p_t$ and observe ℓ_{t,I_t}

  for *active arm* $i \in [K]$ **do**

    if $i \in \mathbb{A}_t$ **then**

      Compute loss estimate $\hat{\ell}_{t,i} = \frac{\ell_{t,i} \mathbb{1}\{I_t = i\}}{p_{t,i}}$

    end

    else

      Set $\hat{\ell}_{t,i} = \ell_{t,I_t}$

    end

    Compute $r_{t,i} = \ell_{t,I_t} - \hat{\ell}_{t,i}$

    Update $R_{t,i} = R_{t-1,i} + r_t$

  end

  Compute $z_t = \min\left(\frac{(4d)^{2-\alpha}}{(1-\alpha)} \sum_{i \in \mathbb{A}_t} \min(p_{t,i}, 1 - p_{t,i})^{1-\alpha}, \beta_t \frac{18d^2}{\gamma} \sum_{i \in \mathbb{A}_t} \tilde{p}_{t,i}\right)$

  Compute $h_t = \frac{1}{\alpha}(\sum_{i=1}^K q_{t,i}^\alpha - 1)$

  Compute $\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}$

end

Algorithm 5: SPM Approach for Fully-Adversarial Sleeping Bandits

Instead of running FTRL on the sequence of losses, we run FTRL on the sequence of *estimated regret* $R_{t,i} = \sum_{s=1}^t \mathbb{1}\{i \in \mathbb{A}_s\}(\ell_{s,I_s} - \hat{\ell}_{s,i})$, i.e.,

$$\begin{aligned} q_t = \arg\min_{x \in \Delta_K} F_t(x) &:= \arg\min_{x \in \Delta_K} \langle -R_{t-1}, x \rangle + \Phi_t(x) \\ &= \arg\min_{x \in \Delta_K} \langle -R_{t-1}, x \rangle + \beta_t \left(\frac{1}{\alpha}(1 - \sum_{i=1}^K x_i^\alpha)\right) - \gamma \sum_{i=1}^K \ln(x_i). \end{aligned}$$

Given the set of active arms $\mathbb{A}_t$, the sampling probability p_t is $p_{t,i} = \frac{\mathbb{1}\{i \in \mathbb{A}_t\}}{\sum_{j=1}^K \mathbb{1}\{j \in \mathbb{A}_t\} q_{t,j}}$. An arm $I_t \sim p_t$ is drawn. The learning rates are set by SPM rules (Ito et al., 2024): $\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}$, where

$$z_t = \min\left(\frac{(4d)^{2-\alpha}}{(1-\alpha)} \sum_{i \in \mathbb{A}_t} \min(p_{t,i}, 1 - p_{t,i})^{1-\alpha}, \beta_t \frac{18d^2}{\gamma} \sum_{i \in \mathbb{A}_t} \tilde{p}_{t,i}\right),$$

$$h_t = (-\psi_{TE}(q_t)) = \frac{1}{\alpha}\left(\sum_{i=1}^K q_{t,i}^\alpha - 1\right).$$

F.1. Regret Analysis

In this section, we prove Theorem 35.

Proof Let $I_{a,t} = \mathbb{1}\{a \in \mathbb{A}_t\}$. By the definition of $\hat{\ell}_t$ and the fact that $p_{t,i} = 0$ for $i \notin \mathbb{A}_t$, for any $a \in \mathbb{A}_t$, we have

$$\mathbb{E}_{I_t}[\hat{\ell}_{t,a}] = \sum_{i=1}^{K} p_{t,i} \frac{\ell_{t,a}\mathbb{1}\{a=i\}}{p_{t,a}} = \sum_{i \in \mathbb{A}_t} p_{t,i} \frac{\ell_{t,a}\mathbb{1}\{a=i\}}{p_{t,a}} = \ell_{t,a}. \tag{86}$$

Therefore, the per-action regret with respect to $a \in [K]$ is

$$\begin{aligned}
R_{T,a} &= \mathbb{E}\left[\sum_{t=1}^{T} I_{a,t}(\ell_{t,I_t} - \ell_{t,a})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} I_{a,t}(\langle \hat{\ell}_t, q_t\rangle - \langle \ell_t, e_a\rangle)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} I_{a,t}(\langle \hat{\ell}_t, q_t - e_a\rangle)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} I_{a,t}(\langle \hat{\ell}_t - \ell_{t,I_t}\mathbf{1}, q_t - e_a\rangle)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \langle -r_t, q_t - e_a\rangle\right],
\end{aligned}$$

where

- the second equality uses

$$\begin{aligned}
\langle \hat{\ell}_t, q_t\rangle &= \sum_{i=1}^{K} \hat{\ell}_{t,i} q_{t,i} \\
&= \sum_{i \in \mathbb{A}_t} \hat{\ell}_{t,i} q_{t,i} + \sum_{i \notin \mathbb{A}_t} \hat{\ell}_{t,i} q_{t,i} \\
&= \hat{\ell}_{t,I_t} q_{t,I_t} + \ell_{t,I_t} \sum_{i \notin \mathbb{A}_t} q_{t,i} \\
&= \frac{\ell_{t,I_t}}{p_{t,I_t}} q_{t,I_t} + \ell_{t,I_t} \sum_{i \notin \mathbb{A}_t} q_{t,i} \\
&= \ell_{t,I_t} \sum_{j \in \mathbb{A}_t} q_{t,j} + \ell_{t,I_t} \sum_{i \notin \mathbb{A}_t} q_{t,i} \\
&= \ell_{t,I_t}.
\end{aligned}$$

- the fourth equality uses $\langle \ell_{t,I_t}\mathbf{1}, q_t - e_a\rangle = \ell_{t,I_t}(\sum_{i=1}^{K} q_{t,i} - e_{a,i}) = 0$.

- the last equality uses

$$r_{t,i} = \begin{cases} \ell_{t,I_t} - \hat{\ell}_{t,i} & i \in \mathbb{A}_t \\ 0 & i \notin \mathbb{A}_t. \end{cases}$$

Let $u_a = (1 - \frac{K}{T})e_a + \frac{1}{T}\mathbf{1}$. We have

$$\sum_{t=1}^{T} \langle -r_t, q_t - e_a \rangle = \sum_{t=1}^{T} \langle -r_t, q_t - u_a \rangle + \sum_{t=1}^{T} \langle -r_t, u_a - e_a \rangle.$$

The expectation of the second term is bounded by

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{T} \langle -r_t, u_a - e_a \rangle] &= \sum_{t=1}^{T} \langle \mathbb{E}[-r_t], u_a - e_a \rangle \\
&= \sum_{t=1}^{T} \sum_{j=1}^{K} \mathbb{E}[-r_{t,j}](u_{a,i} - e_{a,i}) \\
&= \sum_{t=1}^{T} \sum_{j=1}^{K} \mathbb{E}[\hat{\ell}_{t,j} - \ell_{t,I_t}](u_{a,i} - e_{a,i}) \\
&= \sum_{t=1}^{T} \sum_{j=1}^{K} (\ell_{t,j} - \mathbb{E}_{i \sim p_t}[\ell_{t,i}])(u_{a,i} - e_{a,i}) \\
&\leq 2K.
\end{aligned}$$

Therefore, we only need to focus on the first term $\sum_{t=1}^{T} \langle -r_t, q_t - u_a \rangle$. Next, by the definition of $F_t(x) = \langle \sum_{s=1}^{t-1} -r_s, x \rangle + \phi_t(x)$ and $q_t = \arg\min_{x \in \Delta_K} F_t(x)$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \langle -r_t, q_t - u_a \rangle \\
&= \left(\sum_{t=1}^{T} \langle -r_t, q_t \rangle \right) - (F_{T+1}(u_a) - \phi_{T+1}(u_a)) \\
&= \left(\sum_{t=1}^{T} \langle -r_t, q_t \rangle \right) - F_1(q_1) + \left(\sum_{t=1}^{T} F_t(q_t) - F_{t+1}(q_{t+1}) \right) + \phi_{T+1}(u_a) + F_{T+1}(q_{T+1}) - F_{T+1}(u_a) \\
&\leq \left(\sum_{t=1}^{T} \langle -r_t, q_t \rangle \right) - F_1(q_1) + \left(\sum_{t=1}^{T} F_t(q_t) - F_{t+1}(q_{t+1}) \right) + \phi_{T+1}(u_a) \\
&= \phi_{T+1}(u_a) - F_1(q_1) + \left(\sum_{t=1}^{T} F_t(q_t) - F_{t+1}(q_{t+1}) + \langle -r_t, q_t \rangle \right)
\end{aligned}$$

$$\leq \gamma K \ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha} - 1) + \underbrace{\left(\sum_{t=1}^{T} F_t(q_t) - F_{t+1}(q_{t+1}) + \langle -r_t, q_t \rangle \right)}_{\heartsuit}.$$

We bound each term in $\heartsuit$ as follows:

$$\begin{aligned}
&F_t(q_t) - F_{t+1}(q_{t+1}) + \langle -r_t, q_t \rangle \\
&= \langle -R_{t-1}, q_t \rangle + \langle R_t, q_{t+1} \rangle + \phi_t(q_t) - \phi_{t+1}(q_{t+1}) + \langle -r_t, q_t \rangle \\
&= \langle -R_{t-1}, q_t - q_{t+1} \rangle + \phi_t(q_t) - \phi_t(q_{t+1}) + \phi_t(q_{t+1}) - \phi_{t+1}(q_{t+1}) + \langle -r_t, q_t - q_{t+1} \rangle \\
&= (\beta_{t+1} - \beta_t) h_{t+1} + (\langle -R_{t-1}, q_t - q_{t+1} \rangle + \phi_t(q_t) - \phi_t(q_{t+1})) \\
&\quad + \langle -r_t, q_t - q_{t+1} \rangle \\
&\leq (\beta_{t+1} - \beta_t) h_{t+1} + \langle -r_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t),
\end{aligned}$$

where the inequality is from $\langle -R_{t-1} + \nabla \phi_t(q_t), q_{t+1} - q_t \rangle \geq 0$ by the optimality of q_t and hence,

$$\begin{aligned}
-D_t(q_{t+1}, q_t) &= \phi_t(q_t) - \phi_t(q_{t+1}) + \langle \nabla \phi_t(q_t), q_{t+1} - q_t \rangle \\
&\geq \phi_t(q_t) - \phi_t(q_{t+1}) + \langle -R_{t-1}, q_t - q_{t+1} \rangle.
\end{aligned}$$

We have $q_{t+1,i} \leq 4d q_{t,i}$ for all $i \in [K]$ from the combination of the results of Lemma 36, Lemma 16 and Lemma 37. It follows that

$$\begin{aligned}
&\sum_{t=1}^{T} \langle -r_t, q_t - u_a \rangle \\
&\leq \gamma K \ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha} - 1) + \sum_{t=1}^{T} (\beta_{t+1} - \beta_t) h_{t+1} + \sum_{t=1}^{T} \langle -r_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \\
&\leq \gamma K \ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha} - 1) + 4d \sum_{t=1}^{T} (\beta_{t+1} - \beta_t) h_t + \sum_{t=1}^{T} \langle -r_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \\
&= \gamma K \ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha} - 1) + 4d \sum_{t=1}^{T} \frac{z_t}{\beta_t} + \sum_{t=1}^{T} \langle -r_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t),
\end{aligned}$$

where the second inequality is from Lemma 13.

Using Lemma 39 and noting that β_t is fixed before round t, we have

$$\begin{aligned}
\mathbb{E}_{I_t}\left[\langle -r_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \right] &\leq \mathbb{E}_{I_t}\left[\min\left(\frac{(4d)^{2-\alpha}}{\beta_t(1-\alpha)} \tilde{p}_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma} \tilde{p}_{t,I_t}^2 \hat{\ell}_{t,I_t}^2 \right) \right] \\
&\leq \min\left(\mathbb{E}_{I_t}\left[\frac{(4d)^{2-\alpha}}{\beta_t(1-\alpha)} \tilde{p}_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2 \right], \mathbb{E}_{I_t}\left[\frac{18d^2}{\gamma} \tilde{p}_{t,I_t}^2 \hat{\ell}_{t,I_t}^2 \right] \right) \\
&\leq \min\left(\frac{(4d)^{2-\alpha}}{\beta_t(1-\alpha)} \sum_{i \in \mathbb{A}_t} \tilde{p}_{t,i}^{1-\alpha}, \frac{18d^2}{\gamma} \sum_{i \in \mathbb{A}_t} \tilde{p}_{t,i} \right)
\end{aligned}$$

$$= z_t.$$

It follows that

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle -r_t, q_t - u_a\rangle\right] \leq \gamma K \ln(T) + \frac{\beta_1}{\alpha}(K^{1-\alpha} - 1) + 6\mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t}{\beta_t}\right].$$

Let

$$\begin{aligned} z_{\max} &= \max_{t\in[T]} z_t, \\ h_{\max} &= \max_{t\in[T]} h_t \\ A &= \max_{t\in[T]} |\mathbb{A}_t| \end{aligned}$$

The quantity $z_{\max}$ is bounded by

$$\begin{aligned} z_{\max} &\leq \frac{(4d)^{2-\alpha}}{1-\alpha}\max_{t\in[T]}\sum_{i=1}^{K}\tilde{p}_{t,i}^{1-\alpha} \\ &\leq \frac{(4d)^{2-\alpha}}{1-\alpha}\max_{t\in[T]}\sum_{i\in\mathbb{A}_t}\tilde{p}_{t,i}^{1-\alpha} \\ &\leq \frac{(4d)^{2-\alpha}A^{\alpha}}{1-\alpha}. \end{aligned}$$

Hence, we can bound $\mathbb{E}\left[\sum_{t=1}^{T}\frac{z_t}{\beta_t}\right]$ using the same analysis for SPM learning rates in Ito et al. (2024) and obtain

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T}\frac{z'_t}{\beta_t}\right] \\ &\leq O\left(\min\left\{\inf_{J\in\mathbb{N}}\mathbb{E}\left[\left\{\sqrt{8J\sum_{t=1}^{T}h_t z_t} + 2\sqrt{2^{-J}Th_{\max}z_{\max}}\right\}\right], \mathbb{E}\left[\sqrt{Th_{\max}z_{\max}}\right]\right\} + \mathbb{E}\left[\frac{z_{\max}}{\beta_1}\right]\right) \\ &\leq O\left(\min\left\{\inf_{J\in\mathbb{N}}\mathbb{E}\left[\left\{\sqrt{8J\sum_{t=1}^{T}h_t z_t} + \sqrt{2^{-J}Th_{\max}z_{\max}}\right\}\right], \mathbb{E}\left[\sqrt{Th_{\max}z_{\max}}\right]\right\}\right), \end{aligned} \tag{87}$$

where the second inequality is due to $\frac{z_{\max}}{\beta_1} \leq O\left(\frac{A^{\alpha}}{(1-\alpha)\frac{4K}{1-\alpha}}\right) = O(1)$. Here, we used

$$\begin{aligned} h_{\max} &= \max_{t\in[T]} h_t \\ &= \frac{1}{\alpha}(\sum_{i=1}^{K} q_{t,i}^{\alpha} - 1) \end{aligned}$$

$$\leq \frac{K^{1-\alpha}-1}{\alpha}.$$

In the adversarial regime, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{z'_t}{\beta_t}\right] \leq O(\left(\mathbb{E}[\sqrt{Th_{\max}z_{\max}}]\right))$$
$$\leq O\left(\sqrt{T\frac{(K^{1-\alpha}-1)A^{\alpha}}{\alpha(1-\alpha)}}\right).$$

Hence, the regret bound is of order

$$\mathbb{E}[R_{T,a}] \leq O\left(\sqrt{T\frac{(K^{1-\alpha}-1)A^{\alpha}}{\alpha(1-\alpha)}}\right).$$

■

F.2. Stability Proofs

Recall from Section B.2 that the function $g : [0,1] \to \mathbb{R}_+$ defined by

$$g(x) = \beta x^{\alpha-1} + \frac{\gamma}{x}$$

is decreasing in $x \in [0,1]$ for $\beta, \gamma > 0$.

Lemma 36 *For any $t \geq 1$, Algorithm 5 guarantees*

$$\beta_{t+1} - \beta_t \leq (1-\frac{1}{d})\gamma q_{t*}^{-\alpha}, \tag{88}$$

where $q_{t*} = \min(\max_{i\in\mathbb{A}_t} q_{t,i}, 1 - \max_{i\in\mathbb{A}_t} q_{t,i})$.

Proof Equation (59) shows that $h_t \geq \frac{1-\alpha}{4\alpha}q_{t*}^{\alpha}$. This implies that $\frac{1}{h_t} \leq \frac{4\alpha}{1-\alpha}q_{t*}^{-\alpha}$. By the definitions of β_{t+1}, z_t and h_t, we have

$$\beta_{t+1} - \beta_t = \frac{z_t}{\beta_t h_t}$$
$$\leq \frac{4\alpha z_t}{(1-\alpha)\beta_t}q_{t*}^{-\alpha}$$
$$\leq \frac{4\alpha}{(1-\alpha)}\frac{18d^2}{\gamma}q_{t*}^{-\alpha}$$
$$\leq (1-\frac{1}{d})\gamma q_{t*}^{-\alpha}$$

where the last inequality uses

$$\frac{72\alpha d^2}{(1-\alpha)\gamma} \leq (1-\frac{1}{d})\gamma \tag{89}$$

for $d = 2$ and $\gamma \geq 48\sqrt{\frac{\alpha}{1-\alpha}}$. ■

Lemma 37 *For any $K \geq 3, \alpha \in (0,1), \beta > 0, \gamma \geq 0, R \in \mathbb{R}^K$ and $h \in [-1,1]$, let $S \subseteq [K]$ be a subset of $[K]$ where $1 \in S$. Let $e_S \in \{0,1\}^K$ be a vector such that $e_{S,i} = \mathbb{1}\{i \in S\}$. Define*

$$x = \arg\min_{p \in \Delta_K} \langle -R, p\rangle + \frac{\beta}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha) - \gamma \sum_{i=1}^K \ln(p_i)$$
$$y = \arg\min_{p \in \Delta_K} \langle -R + \frac{h}{x_1'} e_1 - h e_S, p\rangle + \frac{\beta}{\alpha}(1 - \sum_{i=1}^K p_i^\alpha) - \gamma \sum_{i=1}^K \ln(p_i),$$

where $1 \geq x_1' \geq x_1$. Fix an $\omega \in (1,2]$. If $\gamma \geq 6$ and $\beta \geq \frac{4K}{(\omega-1)(1-\omega^{\alpha-1})}$, then $y_i \leq 4x_i$ for all $i \in [K]$.

Proof Using the Lagrange multiplier methods, we have the following three equalities that hold for some $Z \in \mathbb{R}$:

$$g(x_1) - g(y_1) = Z + h - \frac{h}{x_1'}, \tag{90}$$
$$g(x_i) - g(y_i) = Z + h \qquad \text{for } i \in S \setminus \{1\}, \tag{91}$$
$$g(x_i) - g(y_i) = Z \qquad \text{for } i \notin S. \tag{92}$$

When $h \leq 0$:

First, we prove that $Z + h \leq 0$. Assume the contrary that $Z + h \geq 0$. Since $h \in [-1, 0]$, this implies that $Z > 0$ and $Z + h + \frac{(-h)}{x_1'} > 0$. Hence, $g(x_i) > g(y_i)$ for all $i \in [K]$, which is a contradiction since both x and y are in Δ_K. Thus, we must have $Z + h \leq 0$.

For any $i \in S \setminus \{1\}$, we have $g(x_i) - g(y_i) = Z + h \leq 0$ and therefore $y_i \leq x_i$. Next, we consider two cases of Z: $Z \geq 0$ and $Z < 0$.

- If $Z \geq 0$: we have $0 \leq Z \leq -h \leq 1$. For all $i \notin S$, we have $g(y_i) = g(x_i) - Z \geq g(x_i) - 1 \geq g(2x_i)$ by Lemma 40, which implies $y_i \leq 2x_i$. Thus, we only need to show that $y_1 \leq 2x_1$. If $y_1 \leq x_1$ then this is trivially true. If $y_1 \geq x_1$,

$$\begin{aligned}\frac{2}{x_1} \geq \frac{-h}{x_1'} &= (-(Z+h)) + \beta(x^{\alpha-1} - y^{\alpha-1}) + \gamma\left(\frac{1}{x_1} - \frac{1}{y_1}\right) \\ &\geq \gamma\left(\frac{1}{x_1} - \frac{1}{y_1}\right) \\ &\geq 4\left(\frac{1}{x_1} - \frac{1}{y_1}\right),\end{aligned}$$

  which leads to $y_1 \leq 2x_1$.

- If $Z < 0$: in this case, for all $i \notin S$, we have $g(x_i) \leq g(y_i)$ which implies $x_i \geq y_i$. As $x_i \geq y_i$ for all $i \neq 1$, we must have $x_1 \leq y_1$. Therefore, we have $y_1 \leq 2x_1$ by the same argument in the previous case.

When $h \geq 0$:

First, we prove that $Z + h \geq 0$. Assume the contrary that $Z + h < 0$. Since $h \in [0, 1]$, this implies $Z < 0$ and $Z + h - \frac{h}{x_1'} < 0$. Hence, $g(x_i) < g(y_i)$ for all $i \in [K]$, which is a contradiction since both x and y are in Δ_K. Thus, we must have $Z + h \geq 0$.

For any $i \in S \setminus \{1\}$, we have $g(x_i) - g(y_i) = Z + h \geq 0$ and therefore $x_i \leq y_i$. Next, we consider two cases of Z: $Z \geq 0$ and $Z < 0$.

- If $Z \geq 0$: in this case, due to the monotonicity of the function g, we have $x_i \leq y_i$ for $i \neq 1$ and therefore, $x_1 \geq y_1$. This implies $Z + h - \frac{h}{x_1'} \leq 0$. Let

$$\epsilon = \frac{1}{\beta(1 - \omega^{\alpha-1})} \leq \frac{\omega - 1}{4K} \leq \frac{1}{K}$$

  as in the proof of Lemma 19. We further consider two cases of x_1'.

  - If $x_1' \geq \epsilon$: we have $Z \leq Z + h \leq \frac{h}{x_1'} \leq \frac{1}{\epsilon}$. Therefore, for all $i \neq 1$,

$$\begin{aligned} g(y_i) &= g(x_i) - Z - h\mathbb{1}\{i \in S\} \\ &\geq g(x_i) - Z - h \\ &\geq g(x_i) - \frac{1}{\epsilon} \\ &= \beta x_i^{\alpha-1} - \beta(1 - \omega^{\alpha-1}) + \frac{\gamma}{x_i} \\ &\geq \beta x_i^{\alpha-1} - \beta x_i^{\alpha-1}(1 - \omega^{\alpha-1}) + \frac{\gamma}{\omega x_i} \\ &= \beta(\omega x_i)^{\alpha-1} + \frac{\gamma}{\omega x_i} = g(\omega x_i), \end{aligned}$$

    where the last inequality is due to $x_i^{\alpha-1} \geq 1$ and $\omega > 1$. This implies that for all $i \neq 1$, $y_i \leq \omega x_i \leq 3x_i$ since $\omega \leq 2$.
  - If $x_1' < \epsilon$: we have $x_1 \leq \epsilon\frac{1}{2K}$ and $\sum_{i=1}^{K}(y_i - x_i) = x_1 - y_1 \leq \epsilon\frac{\omega-1}{K}$. Let $i* = \arg\max_{i\in[K]} x_i$. We have $x_{i*} \geq \frac{1}{K} > \frac{1}{2K}$, hence $i* \neq 1$. Furthermore,

$$\begin{aligned} \frac{1}{K}\left(\frac{y_{i*}}{x_{i*}} - 1\right) &\leq x_{i*}\left(\frac{y_{i*}}{x_{i*}} - 1\right) \\ &= y_{i*} - x_{i*} \\ &\leq \sum_{i\neq 1}(y_i - x_i) \\ &\leq \frac{\omega - 1}{K}, \end{aligned}$$

    which implies that $y_{i*} \leq \omega x_{i*}$. Therefore, using the fact that $g(x) - g(\omega x)$ is also decreasing in x, for all $i \neq 1$, we have

$$\begin{aligned} g(y_i) &= g(x_i) - (Z + h\mathbb{1}\{i \in S\}) \\ &\geq g(x_i) - Z - 1 \qquad \text{since } h \in [0, 1] \end{aligned}$$

$$\begin{aligned}
&\geq g(x_i) - (g(x_{i*}) - g(y_{i*})) - 1 \\
&\geq g(x_i) - (g(x_{i*}) - g(\omega x_{i*})) - 1 \\
&\geq g(x_i) - (g(x_i) - g(\omega x_i)) - 1 \\
&= g(\omega x_i) - 1 \\
&\geq g(2\omega x_i),
\end{aligned}$$

where the second inequality is from $Z \leq Z + h\mathbb{1}\{i^* \in S\} = g(x_{i*}) - g(y_{i*})$, the third inequality is from $g(y_{i*}) \geq g(\omega x_{i*})$, and the last inequality is $g(x) - 1 \geq g(2x)$ by Lemma 40. From $g(y_i) \geq g(2\omega x_i)$, we conclude that $y_i \leq 2\omega x_i \leq 4x_i$ for all $i \neq 1$.

- If $Z < 0$: since $x'_1 \leq 1$ and $h \in [0,1]$, we have $Z + h - \frac{h}{x'_1} < 0$. It follows that $g(x_1) - g(y_1) < 0$, hence $x_1 \geq y_1$. Moreover, for $i \notin S$, we also have $x_i \geq y_i$ due to $0 > Z = g(x_i) - g(y_i)$. Thus, we only need to show $y_i \leq 3x_i$ for $i \in S \setminus \{1\}$. For such i, we have

$$g(y_i) = g(x_i) - (h + Z) \geq g(x_i) - h \geq g(x_i) - 1 \geq g(2x_i),$$

where the last inequality is from Lemma 40. This implies $y_i \leq 2x_i$.

■

Lemma 38 *For any $t \in [T]$ and constant $c \in [0,1]$, Algorithm 5 guarantees*

$$\sum_{i=1}^{K} (q_{t,i})^{2-c} r_{t,i}^2 \leq 2(\tilde{p}_{t,I_t})^{2-c} \hat{\ell}_{t,I_t}^2.$$

Proof Since $r_{t,i} = 0$ for $i \notin \mathbb{A}_t$, $r_{t,i} = \ell_{t,I_t}$ for $i \in \mathbb{A}_t \setminus \{I_t\}$ and $\hat{\ell}_{t,I_t} = \frac{\ell_{t,I_t}}{p_{t,I_t}}$, we have

$$\begin{aligned}
\sum_{i=1}^{K} q_{t,i}^{2-c} r_{t,i}^2 &= \sum_{i \in \mathbb{A}_t} q_{t,i}^{2-c} r_{t,i}^2 \\
&= \ell_{t,I_t}^2 \sum_{i \in \mathbb{A}_t, i \neq I_t} q_{t,i}^{2-c} + q_{t,I_t}^{2-c} \ell_{t,I_t}^2 \left(1 - \frac{1}{p_{t,I_t}}\right)^2 \\
&= \frac{\ell_{t,I_t}^2}{p_{t,I_t}^2} \left(p_{t,I_t}^2 \sum_{i \in \mathbb{A}_t, i \neq I_t} q_{t,i}^{2-c} + q_{t,I_t}^{2-c} \left(p_{t,I_t} - 1\right)^2 \right) \\
&\leq \frac{\ell_{t,I_t}^2}{p_{t,I_t}^2} \left(p_{t,I_t}^2 \sum_{i \in \mathbb{A}_t, i \neq I_t} p_{t,i}^{2-c} + p_{t,I_t}^{2-c} \left(p_{t,I_t} - 1\right)^2 \right) \\
&\leq \frac{\ell_{t,I_t}^2}{p_{t,I_t}^2} \left(p_{t,I_t}^2 \left(\sum_{i \in \mathbb{A}_t, i \neq I_t} p_{t,i} \right)^{2-c} + p_{t,I_t}^{2-c} \left(p_{t,I_t} - 1\right)^2 \right) \\
&= \hat{\ell}_{t,I_t}^2 (p_{t,I_t}(1 - p_{t,I_t}))^{2-c} \left(p_{t,I_t}^c + (1 - p_{t,I_t})^c \right)
\end{aligned}$$

$$\leq 2\hat{\ell}_{t,I_t}^2 (p_{t,I_t}(1-p_{t,I_t}))^{2-c}$$
$$\leq 2\tilde{p}_{t,I_t}^{2-c}\hat{\ell}_{t,I_t}^2,$$

where the first inequality is due to $q_{t,i} \leq p_{t,i}$ for $i \in \mathbb{A}_t$, the second inequality is from repeatedly applying $a^x + b^x \leq (a+b)^x$ for $x = 2 - c \geq 1$ by Lemma 21, the third inequality is $p_{t,I_t}^c \leq 1$ and $(1-p_{t,I_t})^c \leq 1$, and the last inequality is $x(1-x) \leq \min(x, 1-x)$ for $x \in [0,1]$. ■

Lemma 39 *For any $t \in [T]$, Algorithm 5 guarantees*

$$\langle -r_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \leq \min\left(\frac{(4d)^{2-\alpha}}{\beta_t(1-\alpha)}\tilde{p}_{t,I_t}^{2-\alpha}\hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma}\tilde{p}_{t,I_t}^2\hat{\ell}_{t,I_t}^2\right) \tag{93}$$

Proof Using standard local-norm analysis techniques for FTRL (for example, see Section 7.4 in Orabona (2023)), we have

$$\langle -r_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \leq \frac{1}{2}\|r_t\|^2_{(\nabla^2\phi_t(z_t))^{-1}}, \tag{94}$$

where z_t is a point between q_t and q_{t+1}. The Hessian matrix of ϕ_t is a diagonal matrix with entries

$$\nabla^2\phi_t(z_t) = \operatorname{diag}\left(\left(\beta_t(1-\alpha)z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2}\right)_{i=1,2,\ldots,K}\right). \tag{95}$$

Hence, its inverse is the following diagonal matrix

$$(\nabla^2\phi(z_t))^{-1} = \operatorname{diag}\left(\left(\frac{1}{\beta_t(1-\alpha)z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2}}\right)_{i=1,2,\ldots,K}\right). \tag{96}$$

It follows that

$$\begin{aligned}\|r_t\|^2_{(\nabla^2\phi_t(z_t))^{-1}} &= \sum_{i=1}^K r_{t,i}^2 \frac{1}{\beta_t(1-\alpha)z_{t,i}^{\alpha-2} + \frac{\gamma}{z_{t,i}^2}} \\ &\leq \min\left(\frac{1}{\beta_t(1-\alpha)}\sum_{i=1}^K z_{t,i}^{2-\alpha}r_{t,i}^2, \frac{1}{\gamma}\sum_{i=1}^K z_{t,i}^2 r_{t,i}^2\right)\end{aligned} \tag{97}$$

where the last equality is due to $\hat{\ell}_{t,i} = 0$ for $i \neq I_t$. Combining (94) and (97), we obtain

$$\langle -r_t, q_t - q_{t+1}\rangle - D_t(q_{t+1}, q_t) \leq \min\left(\frac{1}{\beta_t(1-\alpha)}\sum_{i=1}^K z_{t,i}^{2-\alpha}r_{t,i}^2, \frac{1}{\gamma}\sum_{i=1}^K z_{t,i}^2 r_{t,i}^2\right). \tag{98}$$

Since z_t is between q_t and q_{t+1}, we have $z_{t,I_t} \leq \max(q_{t,I_t}, q_{t+1,I_t})$. The loss estimate in Algorithm 5 uses p_{t,I_t} where $p_{t,I_t} \geq q_{t,I_t}$, therefore we can combine the results of Lemma 36, Lemma 16 and Lemma 37 and obtain $q_{t+1,i} \leq 4dq_{t,i}$ for all $i \in [K]$. It follows that $z_{t,i} \leq 4dq_{t,i}$, and as a result,

$$\langle \hat{\ell}_t, q_t - q_{t+1} \rangle - D_t(q_{t+1}, q_t) \leq \min\left(\frac{(4d)^{2-\alpha}}{2\beta_t(1-\alpha)} \sum_{i=1}^{K} q_{t,i}^{2-\alpha} r_{t,i}^2, \frac{9d^2}{2\gamma} \sum_{i=1}^{K} q_{t,i}^2 r_{t,i}^2 \right) \tag{99}$$

$$\leq \min\left(\frac{(4d)^{2-\alpha}}{\beta_t(1-\alpha)} \tilde{p}_{t,I_t}^{2-\alpha} \hat{\ell}_{t,I_t}^2, \frac{18d^2}{\gamma} \tilde{p}_{t,I_t}^2 \hat{\ell}_{t,I_t}^2 \right), \tag{100}$$

where the last inequality is from applying Lemma 38 twice: once with $c = \alpha$ and once with $c = 0$. ■

F.3. Technical Lemmas

Lemma 40 *For any $x \in [0,1]$, if $\gamma \geq 2$ then $g(x) - 1 \geq g(2x)$.*

Proof We have

$$\begin{aligned} g(x) - 1 &= \beta x^{\alpha-1} + \frac{\gamma}{x} - 1 \\ &\geq \beta 2^{\alpha-1} x^{\alpha-1} + \frac{\gamma}{2x} + \frac{\gamma}{2x} - 1 \\ &= g(2x) + \frac{\gamma - 2x}{2x} \\ &\geq g(2x). \end{aligned}$$

■

Appendix G. Setting α appropriately close to 1

Recall that we assume $K \geq 3$ in our algorithms. Let $b = 1 - \alpha$. We will set $\alpha = 1 - \frac{0.5}{\ln(K)}$, which is equivalent to setting $b = \frac{0.5}{\ln(K)}$. Note that $\alpha \geq 1 - \frac{0.5}{\ln(3)} > 0.5$. Taking exponent on both sides of

$$\ln(1 + 2b\ln(K)) = \ln(2) \geq 0.5 = b\ln(K), \tag{101}$$

we obtain $1 + 2b\ln(K) \geq K^b$. This implies

$$\frac{K^{1-\alpha} - 1}{\alpha(1-\alpha)} \leq \frac{2(K^b - 1)}{b} \leq 4\ln(K). \tag{102}$$

Furthermore,

$$\frac{(K-1)^{1-\alpha}}{\alpha(1-\alpha)} \leq \frac{2(K-1)^{1-\alpha}}{1-\alpha} = \ln(K)(K-1)^{\frac{0.5}{\ln K}} \leq \ln(K) K^{\frac{0.5}{\ln K}} \leq 2\ln K, \tag{103}$$

where the last inequality is due to $K^{\frac{0.5}{\ln(K)}} = (e^{\ln K})^{\frac{0.5}{\ln(K)}} = e^{0.5} < 2$. In addition,

$$\frac{\alpha}{1-\alpha} \leq \frac{1}{1-\alpha} = \frac{1}{b} = 2\ln(K). \tag{104}$$

This implies that $\gamma = \max\left(6, 48\sqrt{\frac{\alpha}{1-\alpha}}\right) \lesssim \sqrt{\ln(K)}$.