Adaptive Learning Rate for Follow-the-Regularized-Leader: Competitive Analysis and Best-of-Both-Worlds

Shinji Ito SHINJI@MIST.I.U-TOKYO.AC.JP
The University of Tokyo and RIKEN AIP
NEC Corporation (affiliation upon submission)

Taira Tsuchiya TSUCHIYA@MIST.I.U-TOKYO.AC.JP
The University of Tokyo and RIKEN AIP

Junya Honda HONDA@I.KYOTO-U.AC.JP
Kyoto University and RIKEN AIP

Editors: Shipra Agrawal and Aaron Roth

Abstract

Follow-The-Regularized-Leader (FTRL) is known as an effective and versatile approach in online learning, where appropriate choice of the learning rate is crucial for smaller regret. To this end, we formulate the problem of adjusting FTRL's learning rate as a sequential decision-making problem and introduce the framework of competitive analysis. We establish a lower bound for the competitive ratio and propose update rules for the learning rate that achieves an upper bound within a constant factor of this lower bound. Specifically, we illustrate that the optimal competitive ratio is characterized by the (approximate) monotonicity of components of the penalty term, showing that a constant competitive ratio is achievable if the components of the penalty term form a monotone non-increasing sequence, and derive a tight competitive ratio when penalty terms are ξ-approximately monotone non-increasing. Our proposed update rule, referred to as *stability-penalty matching*, also facilitates the construction of Best-Of-Both-Worlds (BOBW) algorithms for stochastic and adversarial environments. In these environments our results contribute to achieving tighter regret bound and broaden the applicability of algorithms for various settings such as multi-armed bandits, graph bandits, linear bandits, and contextual bandits.

Keywords: follow-the-regularized-leader, adaptive learning rate, competitive analysis, best-of-both-worlds bandit algorithm

1. Introduction

In the research field of online learning and bandit algorithms, the *follow-the-regularized-leader* (FTRL) framework offers a promising approach to achieving sublinear regret. In this framework, we choose an action a_t in each round t, on the basis of $x_t \in \mathcal{X}$, a solution to the following convex optimization problem:

$$x_t \in \mathop{\arg\min}_{x \in \mathcal{X}} \left\{ \sum_{s=1}^{t-1} \hat{f}_s(x) + \frac{1}{\eta_t} \psi(x) \right\}, \tag{1}$$

where $\mathcal{X}$ is a convex set, $\{\hat{f}_s\}$ are estimators or surrogates of the loss functions, $\{\eta_t\}$ are learning rate parameters that are positive and monotone non-decreasing, and ψ is a convex regularizer function. This approach can be interpreted as a comprehensive framework that includes Online Gradient Descent (Zinkevich, 2003) and the Hedge algorithm (Littlestone and Warmuth, 1994; Arora et al., 2012; Freund and Schapire, 1997), which demonstrates its effectiveness across various online learning and bandit problems, such as multi-armed bandits (Auer et al., 2002), linear bandits (Abernethy et al., 2008; Cesa-Bianchi and Lugosi, 2012), and episodic MDPs (Lee et al., 2020).

To harness the effectiveness of FTRL, it is crucial to appropriately set the learning rate. Here, a fixed learning rate determined by time horizon T often suffices when T is predefined and the goal is the worst-case optimality. On the other hand, adaptive update of the learning rate based on feedback received at each time step has been considered when T is not predetermined and/or the goal is to achieve the optimality beyond the worst case with better practical performance. Such methods of adaptive learning rate have been shown to be beneficial in achieving data-dependent bounds (Cesa-Bianchi et al., 2007; Orabona and Pál, 2015; Erven et al., 2011) and in constructing *best-of-both-worlds* (BOBW) algorithms (Gaillard et al., 2014; Bubeck and Slivkins, 2012; Ito, 2021b; Jin et al., 2023) that attain (nearly) optimal performance in both adversarial and stochastic settings. Other literature on adaptive learning rates is also mentioned in Appendix A.

This paper aims to develop a generic methodology for sequentially adjusting learning rate in FTRL, and to investigate its limitations. A standard analysis for FTRL (e.g., in Lattimore and Szepesvári, 2020, Exercise 28.12) provides an upper bound on the regret R_T as follows:

$$R_T \lesssim \underbrace{\sum_{t=1}^{T} \eta_t z_t}_{\text{stability terms}} + \underbrace{\frac{1}{\eta_1} h_1 + \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) h_t}_{\text{penalty terms}}, \tag{2}$$

where z_t and h_t vary depending on the problem setup and the regularizer function ψ. For example, in the Hedge algorithm, i.e., when (1) is specified by $\mathcal{X} = \mathcal{P}(K) = \{x \in [0,1]^K \mid \|x\|_1 = 1\}$, $\hat{f}_t(x) = \ell_t^\top x$ with $\ell_t \in [0,1]^K$, and $\psi(x)$ is the negative Shannon entropy, z_t and h_t are bounded as $z_t \leq O\left(\sum_{i=1}^K \ell_{ti}^2 x_{ti}\right) \leq O\left(\ell_t^\top x_t\right) \leq O(1)$ and $h_t \leq -\psi(x_t) \leq \log K$. In general FTRL, a standard way of defining h_t is to set $h_t = \max_x \psi(x) - \psi(x_t)$. Some concrete examples of z_t will be discussed later, such as in Section 4. Many existing methods for sequentially updating the learning rate adjust η_t based solely on z_t (Cesa-Bianchi et al., 2007; Orabona and Pál, 2015; Erven et al., 2011). Recently, there has been consideration for adjusting the learning rate in response to h_t as well (Ito et al., 2022b; Tsuchiya et al., 2023a; Kong et al., 2023), and approaches that adjust according to both z_t and h_t have emerged (Jin et al., 2023; Tsuchiya et al., 2023b). However, these update methods using h_t are often somewhat ad-hoc, designed for specific objectives (e.g., BOBW bounds), and the optimality of these update rules themselves have not been investigated. More literature on FTRL with Tsallis entropy regularization is referenced in Appendix A.

1.1. Main contribution

We first formulate the problem of choosing the learning rate as an online decision-making problem to minimize the right-hand side of (2), which is denoted by $F(\eta_{1:T}; z_{1:T}, h_{1:T})$. For any update rule π, we denote by $F^\pi(z_{1:T}, h_{1:T})$ the value of $F(\eta_{1:T}; z_{1:T}, h_{1:T})$ for $\eta_{1:T}$ determined by π, where the update rule π is specified as a series of functions: $\pi = \{\pi_t : (z_{1:t}, h_{1:t}) \mapsto \eta_t\}_{t \in \mathbb{N}}$. We also define $F^*(z_{1:T}, h_{1:T})$ as the minimum of $F(\eta_{1:T}; z_{1:T}, h_{1:T})$ achieved by the optimal sequence $\eta_1^* \geq \eta_2^* \geq \cdots \geq \eta_T^*$ of learning rates given the entire series of $z_{1:T}$ and $h_{1:T}$ in advance. Note that each η_t^* may depend on $z_{1:T}$ and $h_{1:T}$ including the "future feedback" after the t-th round. To

Table 1: Upper bounds on F^π achieved by proposed update rules π for learning rates.

Input for η_t	F^*-dependent bound	$(z_{1:T}, h_{1:T})$-dependent bound
$z_{1:t}, h_{1:t}$	$4\sqrt{\xi}F^*$	$\min\left\{\sqrt{\sum_{t=1}^T z_t h_t \log T}, \sqrt{h_{\max}\sum_{t=1}^T z_t}\right\}$
$z_{1:t-1}, h_{1:t-1}, \xi$	$4\sqrt{\xi}F^* + O(z_{\max} + h_1)$	$\min\left\{\sqrt{\xi\sum_{t=1}^T z_t h_t \log T}, \sqrt{\xi h_{\max}\sum_{t=1}^T z_t}\right\}$
$z_{1:t-1}, h_{1:t-1}, \hat{h}_t$	–	$\min\left\{\sqrt{\sum_{t=1}^T z_t \hat{h}_{t+1} \log T}, \sqrt{\hat{h}_{\max}\sum_{t=1}^T z_t}\right\}$
Lower bound	$\frac{\sqrt{T-1}}{\sqrt{T}+\xi}\sqrt{\xi}F^*$	–

evaluate the performance of policies π and the complexity of this online decision-making problem, we focus on the *competitive ratio* defined as $\mathrm{CR}(\pi; z_{1:T}, h_{1:T}) = \frac{F^\pi(z_{1:T}, h_{1:T})}{F^*(z_{1:T}, h_{1:T})}$.

This study reveals that the optimal competitive ratio can be characterized by *approximate monotonicity* of $h_{1:T}$. For any fixed $\xi \geq 1$, a sequence $h_{1:T}$ is called ξ-approximately monotone non-increasing if $\xi h_{t'} \geq h_t$ for all t and $t' < t$. Letting $H_\xi^T \subseteq \mathbb{R}_{>0}^T$ denote the set of all ξ-approximately monotone non-increasing sequences, we have the following lower bound on the competitive ratio:

Theorem 1 *For any $T \in \mathbb{N}$, any $\xi \geq 1$, and for any policy $\pi = \{\pi_t : (z_{1:t}, h_{1:t}) \mapsto \eta_t\}$, there exist $z_{1:T} \in \mathbb{R}_{\geq 0}^T$ and $h_{1:T} \in H_\xi^T$ such that $\mathrm{CR}(\pi; z_{1:T}, h_{1:T}) \geq \frac{\sqrt{T-1}}{\sqrt{T}+\xi}\sqrt{\xi}$.*

This lower bound implies that conditions on $h_{1:T}$ such as approximate monotonicity are essential in order to establish non-trivial upper bounds on the competitive ratio. The proof of this theorem is given in the appendix. Note that the instance (sequences of z_t and h_t) constructed in the proof of Theorem 1 is just a hard case for the abstract subproblem of optimizing F, and it is not yet known whether such an example may appear in an actual learning process of FTRL. Thus, if z_t and h_t that appear in the actual FTRL satisfy some conditions, then there is still a possibility of improvement beyond the lower bound of Theorem 1.

This paper also provides a policy $\pi = \{\pi_t : (z_{1:t}, h_{1:t}) \mapsto \eta_t\}_{t\in\mathbb{N}}$ achieving a competitive-ratio upper bound that matches the lower bound in Theorem 1 up to a constant. This policy is expressed by the solution of the following formula:

$$\eta_1 z_1 = \frac{1}{\eta_1} h_1, \quad \eta_t z_t = \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) h_t \quad (t \geq 2), \tag{3}$$

i.e., the learning rate under which stability and penalty match in each round, which is referred to as *stability-penalty matching* (SPM) in this paper. This formula of (3) leads to the initialization of $\eta_1 = \sqrt{z_1/h_1}$ and the update rule of $\eta_t = \frac{2}{1+\sqrt{1+4\eta_{t-1}^2 z_t/h_t}}\eta_{t-1}$ for $t \geq 2$.

Theorem 2 *The policy $\pi = \{\pi_t : (z_{1:t}, h_{1:t}) \mapsto \eta_t\}_{t\in\mathbb{N}}$ given by (3) achieves $\mathrm{CR}(\pi; z_{1:T}, h_{1:T}) \leq 4\sqrt{\xi}$ for any $\xi \geq 1$, $z_{1:T} \in \mathbb{R}_{\geq 0}^T$, and $h_{1:T} \in H_\xi^T$. In addition, this policy achieves*

$$F^\pi(z_{1:T}, h_{1:T}) = O\left(\min\left\{\inf_{\varepsilon \geq \frac{1}{T}}\left\{\sqrt{\sum_{t=1}^T z_t h_t \log(\varepsilon T) + \frac{z_{\max} h_{\max}}{\varepsilon}}\right\}, \sqrt{h_{\max}\sum_{t=1}^T z_t}\right\}\right) \tag{4}$$

for any $z_{1:T} \in \mathbb{R}_{\geq 0}^T$ *and* $h_{1:T} \in \mathbb{R}_{>0}^T$, *where* $h_{\max} = \max_{t\in[T]} h_t$ *and* $z_{\max} = \max_{t\in[T]} z_t$.

The upper bound of (4) holds for any sequences of $z_{1:T}$ and $h_{1:T}$ without any requirement on the monotonicity. Upper bounds in this form is useful in developing and analyzing BOBW bandit algorithms, as can be seen in Section 1.2 and Section 4.

Theorems 1 and 2 together imply that the tight competitive ratio under the condition on the approximate monotonicity of $h_{1:T}$ is of $\Theta(\sqrt{\xi})$, and that such a tight competitive ratio is achieved by the policy given by (3).

We note that in the implementation of policy by (3), we need to know h_t and z_t at the time of determining η_t. Such a knowledge is not always available in practice as h_z and z_t may depend on η_t. To deal with such situations, we also develop learning-rate policies that do not require values of h_t and z_t when determining η_t. Bounds on F^π achieved by such policies are summarized in Table 1. The "Input" row in this table represents the knowledge required in determining η_t. For example, if the input is $z_{1:t-1}, h_{1:t-1}, \hat{h}_t$, the policy can be expressed as $\pi = \{\pi_t : (z_{1:t-1}, h_{1:t-1}, \hat{h}_t) \mapsto \eta_t\}_{t\in\mathbb{N}}$. The value $\hat{h}_t$ in this table is an arbitrary upper bound on h_t that is available when determining η_t. A typical example of $\hat{h}_t$ is to set $\hat{h}_t = h_{t-1}$, which is justified when $h_t = O(h_{t-1})$ holds and this condition can be ensured, e.g., via Lemmas 24 and 25 in this paper and via lemmas in Jin et al. (2023, Appendix C.3). Another example of $\hat{h}_t$ is to define $\hat{h}_t = \xi\tilde{h}_{t-1} := \xi \min_{s\in[t-1]} h_s$, which is an upper bound of h_t if $h_{1:T} \in H_\xi^T$. Bounds shown in Table 1 are achieved by variants of the policy given by (3), which are provided in Section 3.

1.2. Application: best-of-both-worlds regret bounds

Bounds on F dependent on $(z_{1:T}, h_{1:T})$ such as (4) are useful in developing BOBW bandit algorithms. Examples dealt with in this paper are summarized in Table 2, where we use the notation of $\log_+(x) = \max\{1, \log(x)\}$. The regret bounds presented in Table 2 are achieved through an algorithmic framework detailed in Algorithm 1. Notably, Algorithm 1 in Section 4 adopts a methodology similar to those found in prior studies, such as Auer et al. (2002); Eldowa et al. (2023); Cesa-Bianchi and Lugosi (2012); Zimmert and Seldin (2021), with the distinct exceptions of its learning rate and regularization definitions. Specifically, the employed regularization function utilizes a hybrid regularizer based on Tsallis entropy, a concept previously explored in Zimmert et al. (2019); Tsuchiya et al. (2023a); Masoudian and Seldin (2021); Jin et al. (2023) and thus, is not a novel contribution of this work. The seminal contribution of this paper lies in the innovative update rules for the learning rate, demonstrating their effectiveness through BOBW results. These findings underscore the proposed SPM learning rates capability to significantly enhance performance.

As demonstrated in Table 2, the SPM learning rates introduced in this paper achieve BOBW regret bounds with tight dependencies on T, for any value of $\alpha \in (0, 1)$ in the α-Tsallis entropy. Specifically, we attain an $O(\log T)$ bound in stochastic environments and an $O(\sqrt{T})$ bound in adversarial environments. When designing FTRL-based BOBW algorithms, various regularizers have been investigated, including α-Tsallis entropy (Zimmert and Seldin, 2021; Jin et al., 2023), log-barrier (Wei and Luo, 2018; Ito et al., 2022a), and Shannon entropy (Ito et al., 2022b). However, achieving bounds in a tight order for both stochastic and adversarial scenarios has been confirmed only when we use the $1/2$-Tsallis entropy regularizer (for instance, see Jin et al., 2023). This research marks the first instance of demonstrating optimality in terms of T for $\alpha \neq 1/2$, thereby presenting a method that allows the parameter α to be adjusted. This adaptability ensures the achievement of

Table 2: Bounds on z_t and regret for FTRL with α-Tsallis entropy and SPM learning rates. Based on the values of $B(\alpha)$ in the upper table, we establish the BOBW regret bounds in the lower table.

Setting	Parameters	Bound on z_t	$B(\alpha)$	$\min_\alpha B(\alpha)$
Multi-armed bandit	K: # arms	$\frac{1}{1-\alpha}\sum_{i \neq i^*} q_{ti}^{1-\alpha}$	$\frac{K-1}{\alpha(1-\alpha)}$	$K-1$
Graph bandit	K: # arms, ζ: independence number	$\frac{\zeta^\alpha(1-q_{ti^*})^{1-\alpha}}{1-\alpha}$	$\frac{\zeta(K/\zeta)^{1-\alpha}}{\alpha(1-\alpha)}$	$\zeta \log_+\left(\frac{K}{\zeta}\right)$
Linear bandit	K: # arms, d: dimensionality	$\frac{d(1-q_{ti^*})^{1-\alpha}}{1-\alpha}$	$\frac{dK^{1-\alpha}}{\alpha(1-\alpha)}$	$d \log K$
Contextual bandit	M: # arms, K: # experts	$\frac{K(1-q_{ti^*})^{1-\alpha}}{1-\alpha}$	$\frac{MK^{1-\alpha}}{\alpha(1-\alpha)}$	$M \log K$

Environment	Regret upper bound
Adversarial	$O\left(\sqrt{B(\alpha)T}\right)$
Stochastic	$O\left(\frac{B(\alpha)}{\Delta_{\min}}\log_+\left(\frac{\Delta_{\min}^2 T}{B(\alpha)}\right)\right)$
Corrupted Stochastic	$O\left(\frac{B(\alpha)}{\Delta_{\min}}\log_+\left(\frac{\Delta_{\min}^2 T}{B(\alpha)}\right) + \sqrt{\frac{CB(\alpha)}{\Delta_{\min}}\log_+\left(\frac{\Delta_{\min} T}{C}\right)}\right)$

optimal bounds relative to problem-specific parameters, such as the independence number in graph bandits or the number of experts in contextual bandits.

The bounds presented in Table 2 of this study can be compared with existing results as follows: For multi-armed bandits, the current state-of-the-art for comparison would be the work of Jin et al. (2023). When $\alpha \neq 1/2$, their bounds in terms of T are $O(\sqrt{T \log T})$ in adversarial environments and $O(\log T)$ in stochastic environments. Our study improves upon these by achieving $O(\sqrt{T})$ and $O(\log T)$, respectively, thus presenting a tight dependency on T. However, their bounds have advantages in achieving the tight dependency on the suboptimality gaps of individual arms and allowing for multiple optimal arms. In the case of $\alpha = 1/2$, our results essentially replicate the bounds of Tsallis-INF (Zimmert and Seldin, 2021; Masoudian and Seldin, 2021), ignoring constant factors. In graph bandits, compared to the bounds by Dann et al. (2023), our results show an improved dependency on $\log K$, achieving the same bounds as the adversarial graph-bandit algorithm by Eldowa et al. (2023) for adversarial environments, which are tight within a constant factor. This can be seen as an extension of the adversarial-only results by Eldowa et al. (2023) to the BOBW results. For contextual and linear bandits, our bounds are nearly identical to those reported by Dann et al. (2023), but notably better when considering corrupted settings; Our method achieves the refined bound in which $\log T$ is replaced by $\log(\Delta_{\min} T/C)$, indicating a superior guarantee under certain conditions suggested by Masoudian and Seldin (2021).

The proposed approach, similarly to other FTRL-based algorithms, achieves bounds of $o(\sqrt{T})$-regret in stochastic regimes with adversarial corruption, and more generally, in adversarial regimes with self-bounding constraints (Zimmert and Seldin, 2021). The specific form of these bounds is presented in Table 2, where $C \geq 0$ represents the corruption level, indicating the magnitude of adversarial corruption. Compared to the $O\left(\frac{1}{\Delta_{\min}} \log T + \sqrt{\frac{C}{\Delta_{\min}} \log T}\right)$-bounds commonly found in existing studies (Dann et al., 2023; Zimmert and Seldin, 2021), our work refines these to a form of $O\left(\frac{1}{\Delta_{\min}} \log_+ \left(\Delta_{\min}^2 T\right) + \sqrt{\frac{C}{\Delta_{\min}} \log_+ \frac{\Delta_{\min} T}{C}}\right)$. Similar bounds for the multi-armed bandit problem have been demonstrated by Masoudian and Seldin (2021), and for an understanding of the significance of these refined bounds, we refer to this paper. This study is the first to achieve such refined bounds for α-Tsallis entropy with $\alpha \neq 1/2$ and to extend their applicability beyond multi-armed bandit problems.

2. Problem Setup

We consider the problem of updating the learning rate η_t so that the RHS of (2) is minimized. To this end, we define $F(\beta_{1:T}; z_{1:T}, h_{1:T})$ by

$$F(\beta_{1:T}; z_{1:T}, h_{1:T}) := \sum_{t=1}^{T} \left(\frac{z_t}{\beta_t} + (\beta_t - \beta_{t-1}) h_t \right), \tag{5}$$

for $\beta_{1:T} = (\beta_t)_{t=1}^T \in \mathbb{R}_{>0}^T$, $z_{1:T} = (z_t)_{t=1}^T \in \mathbb{R}_{\geq 0}^T$, and $h_{1:T} = (h_t)_{t=1}^T \in \mathbb{R}_{>0}^T$, where we let $\beta_0 = 0$ for notational simplicity. The value of F is equal to the main components of the RHS of (2), under the variable transformation of $\beta_t = 1/\eta_t$. We address a sequential decision-making problem where the objective is to choose β_t based on the information up to the t-th round, given by $(z_{1:t}, h_{1:t})$, or up to the $(t-1)$-th round, given by $(z_{1:t-1}, h_{1:t-1})$, with the goal of minimizing the value of F.

For any policy π of choosing β_t, let $F^\pi(z_{1:T}, h_{1:T})$ be the value of $F(\beta_{1:T}, z_{1:T}, h_{1:T})$ for $\beta_{1:T}$ determined by π. We measure the performance of policies π based on the competitive ratio given by

$$\mathrm{CR}(\pi; z_{1:T}, h_{1:T}) = \frac{F^\pi(z_{1:T}, h_{1:T})}{F^*(z_{1:T}, h_{1:T})}, \tag{6}$$

where F^* represents the minimum value of F achieved by the offline optimization procedure depending on the entire series of $z_{1:t}$ and $h_{1:t}$ in hindsight, i.e.,

$$F^*(z_{1:T}, h_{1:T}) = \inf\{F(\beta_{1:T}; z_{1:T}, h_{1:T}) \mid 0 \leq \beta_1 \leq \beta_2 \leq \cdots \leq \beta_T\}. \tag{7}$$

Remark 3 *The constraint of $\beta_t \leq \beta_{t+1}$ is equivalent to the constraint that the learning rate η_t is monotone non-increasing, i.e., $\eta_t \geq \eta_{t+1}$. Although this constraint is not absolutely necessary in the algorithm design, it is often needed when obtaining regret upper bounds of the form of* (2).

In interpreting the competitive ratio as defined in (6) of this paper, it is essential to be aware of its practical implications and limitations. A smaller competitive ratio implies that, upon fixing any sequences of $z_{1:T}$ and $h_{1:T}$, the performance closely approximates that for the optimal sequence $\beta_{1:T}$ of learning rates. However, in the context of actual applications to FTRL, the scenario is more complex because the values of z_t and h_t are influenced by the learning rate $\beta_{1:t}$ itself. This leads to a critical insight: Our competitive analysis does not incorporate how changes in the learning rate might affect $z_{1:T}$ and $h_{1:T}$ directly. In other words, the "optimality" of the learning rate update rules, in the sense of the competitive ratio, merely signifies optimality from the perspective of dependency on $z_{1:T}$ and $h_{1:T}$, without considering the effects that learning rates have on $z_{1:T}$ and $h_{1:T}$. Despite this limitation, bounds dependent on $z_{1:T}$ and $h_{1:T}$ provide various benefits, such as data-dependent bounds (Cesa-Bianchi et al., 2007; Erven et al., 2011; De Rooij et al., 2014; Orabona and Pál, 2015) and BOBW bounds (Zimmert and Seldin, 2021) that are also discussed in Section 4, and are thus of practical utility.

This paper shows that the optimal competitive ratio for some reasonable classes of policies can be characterized by *approximate monotonicity* of $h_{1:T}$:

Definition 4 *Let $\xi \geq 1$. We call a sequence $h_{1:T}$ is ξ-approximately non-increasing if $\xi h_{t'} \geq h_t$ holds for any t and t' such that $t' < t$.*

Note that 1-approximately non-increasing sequences are monotone non-increasing. For any $\xi \geq 1$, let H_ξ^T denote the set of ξ-approximately non-increasing sequences, i.e.,

$$H_\xi^T = \left\{ h_{1:T} \in \mathbb{R}_{>0}^T \mid t' < t \Longrightarrow \xi h_{t'} \geq h_t \right\}. \tag{8}$$

In our analysis, we use the following property of ξ-approximately non-increasing sequences:

Lemma 5 *Suppose $h_{1:T} \in H_\xi^T$. Then, $\tilde{h}_{1:T} \in H_1^T$ defined by $\tilde{h}_t = \min_{s \in \{1,2,\ldots,t\}} h_s$ satisfies $\tilde{h}_t \leq h_t \leq \xi \tilde{h}_t$ for all t.*

This lemma implies that the parameter $\xi \geq 1$ represents the ratio of how well the sequence $h_{1:T}$ can be approximated by a monotone non-increasing sequence. All omitted proofs are given in the appendix. Sequence $\tilde{h}_{1:T} \in H_1^T$ given in Lemma 5 will be utilized in Section 3. For any nonnegative integer $n \in \mathbb{Z}_{\geq 0}$, we denote $[n] = \{1, 2, \ldots, n\}$. We also use the natation of $z_{\max} = \sup_t z_t$ and $h_{\max} = \sup_t h_t$.

3. Stability-Penalty Matching

Assume that at the time of choosing β_t, we are given an access to $\hat{h}_t$, an upper bound or an approximated value of h_t. Consider the following two update rules:

$$\text{Rule 1} \quad \pi = \left\{ \pi_t : (z_{1:t}, h_{1:t-1}, \hat{h}_t) \mapsto \beta_t \right\} \qquad \beta_0 = 0, \quad \beta_t = \beta_{t-1} + \frac{z_t}{\beta_t \hat{h}_t} \quad (t \geq 1), \tag{9}$$

$$\text{Rule 2} \quad \pi = \left\{ \pi_t : (z_{1:t-1}, h_{1:t-1}, \hat{h}_t) \mapsto \beta_t \right\} \quad \beta_1 > 0, \quad \beta_t = \beta_{t-1} + \frac{z_{t-1}}{\beta_{t-1} \hat{h}_t} \quad (t \geq 2). \tag{10}$$

We set learning rates by $\eta_t = 1/\beta_t$ with β_t given by these rules. We refer to these update rule as *stability-penalty-matching* (SPM) learning rate, as they are designed so that the t-th stability term z_t/β_t (or the $(t-1)$-th stability term z_{t-1}/β_{t-1}) matches the t-th penalty term $(\beta_t - \beta_{t-1})h_t$.

The update rule of (9) can be viewed as a quadratic equation in β_t, whose positive solution is $\beta_t = \frac{\beta_{t-1}}{2}\left(1 + \sqrt{1 + z_t/(\beta_{t-1}^2 \hat{h}_t)}\right)$. Specifically in our analysis, we consider two typical settings of $\hat{h}_t$: One is to set $\hat{h}_t = h_t$, and the other is to set

$$\hat{h}_1 = \xi h_1, \quad \hat{h}_t = \xi \tilde{h}_{t-1} = \xi \min_{s \in [t-1]} h_s \quad (t \geq 2), \tag{11}$$

where the latter ensures $h_t \le \hat{h}_{t+1} \le \hat{h}_t$ and $\hat{h}_{t+1} \le \xi h_t$, which are used in our analysis. These inequalities follow from Lemma 5.

Remark 6 *The SPM learning rate can replicate several existing learning rate update rules under certain parameter settings. For example, if we set $h_t = \bar{h}$ for all t, (9) and (10) lead to $\beta_t = \Theta\left(\sqrt{\bar{h}\sum_{s=1}^{t} z_s}\right)$ and $\beta_t = \Theta\left(\beta_1 + \sqrt{\bar{h}\sum_{s=1}^{t-1} z_s}\right)$, respectively, which correspond to AdaFTRL-type learning rates (Cesa-Bianchi et al., 2007; De Rooij et al., 2014; Erven et al., 2011; Orabona and Pál, 2015; Ito, 2021b). This approach is known to achieve regret bounds of $O\left(\sqrt{\bar{h}\sum_{t=1}^{T} z_t}\right)$. By considering another example, when $z_t = \Theta(\hat{h}_t)$, (9) leads to $\beta_t = \Theta(\sqrt{t})$. As a corresponding case, in Tsallis-INF using the $1/2$-Tsallis entropy (Zimmert and Seldin, 2021), we can see that $z_t \approx h_t$, and it is known to be advantageous to use a learning rate of $\beta_t = \Theta(\sqrt{t})$. Further, when we set $z_t = p_t^{1-\alpha}$ and $\hat{h}_t = p_{t-1}^{\alpha}$ for some $p_t \in (0,1)$ and $\alpha \in (0,1)$, (10) leads to $\beta_t = \Theta\left(\beta_1 + \sqrt{\sum_{s=1}^{t-1} p_s^{1-2\alpha}}\right)$, which replicates the learning rate designed by Jin et al. (2023) for FTRL-based MAB algorithms with $(1-\alpha)$-Tsallis entropy regularizers.*

We show that SPM update rules achieve the following:

Theorem 7 *Suppose $h_t \le \hat{h}_t$ holds for all t. If β_t is given by (9), it holds that*

$$F(\beta_{1:T}; z_{1:T}, h_{1:T}) = O\left(\min\left\{\inf_{\varepsilon \ge \frac{1}{T}}\left\{\sqrt{\sum_{t=1}^{T} z_t \hat{h}_t \log(\varepsilon T) + \frac{z_{\max}\hat{h}_{\max}}{\varepsilon}}\right\}, \sqrt{\hat{h}_{\max}\sum_{t=1}^{T} z_t}\right\}\right). \quad (12)$$

If β_t is given by (10), it holds that

$$F(\beta_{1:T}; z_{1:T}, h_{1:T})$$
$$= O\left(\min\left\{\inf_{\varepsilon \ge \frac{1}{T}}\left\{\sqrt{\sum_{t=1}^{T} z_t \hat{h}_{t+1} \log(\varepsilon T) + \frac{z_{\max}\hat{h}_{\max}}{\varepsilon}}\right\}, \sqrt{\hat{h}_{\max}\sum_{t=1}^{T} z_t}\right\} + \frac{z_{\max}}{\beta_1} + \beta_1 \hat{h}_1\right). \quad (13)$$

The following bounds dependent on F^* also hold:

(a) If $\pi = \{\pi_t : (z_{1:t}, h_{1:t}) \mapsto \beta_t\}$ is given by (9) with $\hat{h}_t = h_t$, it holds for any T, $\xi \ge 1$, $z_{1:T} \in \mathbb{R}_{\ge 0}^T$, and $h_{1:T} \in H_\xi^T$ that $F^\pi(z_{1:T}, h_{1:T}) \le 4\sqrt{\xi} F^*(z_{1:T}, h_{1:T})$.

(b) If $\pi = \{\pi_t : (z_{1:t-1}, h_{1:t}) \mapsto \beta_t\}$ is given by (10) with $\hat{h}_t = h_t$, it holds for any T, $\xi \ge 1$, $z_{1:T} \in \mathbb{R}_{\ge 0}^T$, and $h_{1:T} \in H_\xi^T$ that $F^\pi(z_{1:T}, h_{1:T}) \le 4\sqrt{\xi} F^*(z_{1:T}, h_{1:T}) + O\left(z_{\max}\beta_1 + \beta_1 \hat{h}_1\right)$.

(c) For any fixed $\xi \ge 1$, if $\pi = \{\pi_t : (z_{1:t}, h_{1:t-1}) \mapsto \beta_t\}$ is given by (9) with $\hat{h}_t$ defined as (11), it holds for any T, $z_{1:T} \in \mathbb{R}_{\ge 0}^T$, and $h_{1:T} \in H_\xi^T$ that $F^\pi(z_{1:T}, h_{1:T}) \le 4\sqrt{\xi} F^*(z_{1:T}, h_{1:T}) + O\left(\sqrt{\xi h_{\max} z_{\max}}\right)$.

(d) For any fixed $\xi \ge 1$, if $\pi = \{\pi_t : (z_{1:t-1}, h_{1:t-1}) \mapsto \beta_t\}$ is given by (10) with $\hat{h}_t$ defined as (11), it holds for any T, $z_{1:T} \in \mathbb{R}_{\ge 0}^T$, and $h_{1:T} \in H_\xi^T$ that $F^\pi(z_{1:T}, h_{1:T}) \le 4\sqrt{\xi} F^*(z_{1:T}, h_{1:T}) + O\left(z_{\max}\beta_1 + \beta_1 \hat{h}_1\right)$.

Note that, in the application to the design of best-of-both-worlds algorithm provided in Section 4, the bound in (13) is mainly used, and the results on the competitive ratio (a)-(d) in Theorem 7 are not used in particular. The upper bound on the competitive given in (a) of Theorem 7, combined with the lower bound in Theorem 1, leads to the following corollary:

Corollary 8 *For the class of policy $\{\pi = \{\pi_t : (z_{1:t}, h_{1:t}) \mapsto \beta_t\}\}$ and for any $\xi \geq 1$, the competitive ratio is bounded as follows:*

$$\inf_{\pi=\{\pi_t:(z_{1:t},h_{1:t})\mapsto\beta_t\}} \sup_{T\in\mathbb{N}, z_{1:T}\in\mathbb{R}^T_{\geq 0}, h_{1:T}\in H^T_\xi} \mathrm{CR}(\pi; z_{1:T}, h_{1:T}) \in \left[\sqrt{\xi}, 4\sqrt{\xi}\right]. \tag{14}$$

For any $z_{1:T} \in \mathbb{R}^T_{\geq 0}$ and $h_{1:T} \in \mathbb{R}^T_{>0}$, define $G(z_{1:T}, h_{1:T})$ by

$$G(z_{1:T}, h_{1:T}) = \sum_{t=1}^{T} \left(\sum_{s=1}^{t} \frac{z_s}{h_s} \right)^{-1/2} z_t. \tag{15}$$

Using this function G, we can provide upper bounds on F as follows:

Lemma 9 *Suppose $h_t \leq \hat{h}_t$ holds for all t. If β_t is given by (9), $F(\beta_{1:T}; z_{1:T}, h_{1:T}) \leq 2G(z_{1:T}, \hat{h}_{1:T})$. If β_t is given by (10), $F(\beta_{1:T}; z_{1:T}, h_{1:T}) \leq 2G(z_{1:T}, \hat{h}_{2:T+1}) + 7\frac{z_{\max}}{\beta_1} + \beta_1 h_1$.*

The value of G can be bounded as follows:

Lemma 10 *Let $\theta_0 > \theta_1 > \theta_2 > \cdots > \theta_J > 0$ be an arbitrary positive and monotone decreasing sequence such that $\theta_0 \geq h_{\max}$. Denote $\mathcal{T}_j = \{t \in [T] \mid \theta_{j-1} \geq h_t > \theta_j\}$ for $j \in [J]$ and $\mathcal{T}_{J+1} = \{t \in [T] \mid \theta_J \geq h_t\}$. We then have $G(z_{1:T}, h_{1:T}) \leq 2\sum_{j=1}^{J+1} \sqrt{\theta_{j-1} \sum_{t\in\mathcal{T}_j} z_t}$. Consequently, by choosing $\theta_j = h_{\max}2^{-j}$, we obtain*

$$G(z_{1:T}, h_{1:T}) \leq \min\left\{ \inf_{J\in\mathbb{N}} \left\{ \sqrt{8J\sum_{t=1}^{T} h_t z_t} + 2\sqrt{2^{-J} T h_{\max} z_{\max}} \right\}, 2\sqrt{h_{\max}\sum_{t=1}^{T} z_t} \right\}. \tag{16}$$

On the other hand, $F^*(z_{1:T}, h_{1:T})$ can be bounded from below as follows:

Lemma 11 *Let $\theta_0 > \theta_1 > \theta_2 > \cdots > \theta_J > \theta_{J+1} = 0$ be an arbitrary positive and monotone decreasing sequence. Denote $\mathcal{T}_j = \{t \in [T] \mid \theta_{j-1} \geq h_t > \theta_j\}$ for $j \in [J]$. Suppose that $h_{1:T}$ is a monotone non-increasing sequence. We then have $F^*(z_{1:T}, h_{1:T}) \geq 2\sum_{j=1}^{J} \sqrt{(\theta_j - \theta_{j+1}) \sum_{t\in\mathcal{T}_j} z_t}$.*

To see the relation between F^* and G, define $H(z_{1:T}, h_{1:T})$ by

$$H(z_{1:T}, h_{1:T}) = \sum_{j=1}^{\infty} \sqrt{\theta_{j-1} \sum_{t\in\mathcal{T}_j} z_t}, \quad \text{where} \quad \theta_j = h_{\max}2^{-j}, \quad \mathcal{T}_j = \{t \in [T] | \theta_{j-1} \geq h_t > \theta_j\}$$

with $\theta_j = h_{\max}2^{-j}$ and $\mathcal{T}_j = \{t \in [T] | \theta_{j-1} \geq h_t > \theta_j\}$. Lemma 10 implies $G(z_{1:T}, h_{1:T}) \leq 2H(z_{1:T}, h_{1:T})$ holds for any $z_{1:T}$ and $h_{1:T}$. Further, Lemma 11 means that $F^*(z_{1:T}, h_{1:T}) \geq H(z_{1:T}, h_{1:T})$ holds if $h_{1:T}$ is monotone non-increasing.

Remark 12 *For the policy π given by (9) with $\hat{h}_t = h_t$, if $h_{1:T}$ is monotone non-increasing, we can see that each of $F^\pi(z_{1:T}, h_{1:T})$, $F^*(z_{1:T}, h_{1:T})$, $G(z_{1:T}, h_{1:T})$ and $H(z_{1:T}, h_{1:T})$ is in the constant factor of the others. In fact, we have $F^\pi(z_{1:T}, h_{1:T}) \leq 2G(z_{1:T}, h_{1:T}) \leq 4H(z_{1:T}, h_{1:T}) \leq 4F^*(z_{1:T}, h_{1:T}) \leq 4F^\pi(z_{1:T}, h_{1:T})$.*

Lemma 13 *If $h_{1:T}$ is ξ-approximately non-increasing for some $\xi \geq 1$ we have*

$$G(z_{1:T}, h_{1:T}) \leq 2\sqrt{\xi} F^*(z_{1:T}, h_{1:T}). \tag{17}$$

Lemma 14 *If $h_{1:T+1} \in H_1^{T+1}$, we then have $H(z_{1:T}, h_{1:T}) \leq H(z_{1:T}, h_{2:T+1}) + 4\sqrt{h_{\max} z_{\max}}$.*

By using the lemmas presented so far, we can prove Theorem 7:

Proof sketch of Theorem 7 Bounds on F of (12) and (13) immediately follow from Lemmas 9 and 10. In the following, we show bounds that depend on F^*. Suppose $h_{1:T} \in H_\xi^T$. Then, $\tilde{h}_t := \min_{s \in [t]} h_s$ satisfies $\tilde{h}_t \leq h_t \leq \xi \tilde{h}_t \leq \xi \tilde{h}_{t-1}$ and $\tilde{h}_{1:T} \in H_1^T$, i.e., $\tilde{h}_t \geq \tilde{h}_{t+1}$. Hence, if $\beta_{1:T}$ is given by (9) with $\hat{h}_t = h_t$, we have

$$F(\beta_{1:T}; z_{1:T}, h_{1:T}) \leq 2G(z_{1:T}, h_{1:T}) \leq 4\sqrt{\xi} F^*(z_{1:T}, h_{1:T}), \tag{18}$$

where the first and second inequalities follow from Lemmas 9 and 13, respectively. This means that (a) in Theorem 7 holds. We next see that (c) holds. If $\beta_{1:T}$ is given by (9) with (11), we then have

$$\begin{aligned}
F(\beta_{1:T}; z_{1:T}, h_{1:T}) &\leq 2G(z_{1:T}, \hat{h}_{1:T}) && \text{(Lemma 9)} \\
&= 2G(z_{1:T}, \xi \tilde{h}_{0:T-1}) = 2\sqrt{\xi} G(z_{1:T}, \tilde{h}_{0:T-1}) && \text{(Definitions of } \tilde{h}_t \text{ and } G \text{ in (15))} \\
&\leq 4\sqrt{\xi} H(z_{1:T}, \tilde{h}_{0:T-1}) && \text{(Lemma 10)} \\
&\leq 4\sqrt{\xi} \left(H(z_{1:T}, \tilde{h}_{1:T}) + 4\sqrt{h_{\max} z_{\max}} \right) && \text{(Lemma 14)} \\
&\leq 4\sqrt{\xi} \left(F^*(z_{1:T}, \tilde{h}_{1:T}) + 4\sqrt{h_{\max} z_{\max}} \right) && \text{(Lemma 11)} \\
&\leq 4\sqrt{\xi} \left(F^*(z_{1:T}, h_{1:T}) + 4\sqrt{h_{\max} z_{\max}} \right), && \text{(Definition of } F^* \text{ and } \tilde{h}_t \leq h_t)
\end{aligned}$$

which completes the proof of (c) in Theorem 7. Other bounds (b) and (d) can be shown in a similar manner. For a complete proof, please refer to Appendix C.7.

4. Application: best-of-both-worlds bandit algorithm

This section provides examples of best-of-both-worlds bandit algorithms based on the stability-penalty-matching learning rate. In problem examples in this paper, we consider the following procedure of online learning: A player is given the number of actions K, and some information of the setup before the game starts. In each round of $t \in \{1, 2, \ldots\}$, the environment chooses a loss vector $\ell_t \in [-1, 1]^K$ while the player chooses an action $I(t) \in [K]$, and then incurs the loss of $\ell_{t,I(t)} \in [-1, 1]$. The available feedback and the structure behind ℓ_t are different depending on the problem setup. The performance of the player is measured by the regret defined as follows:

$$R_T(i^*) = \mathbf{E}\left[\sum_{t=1}^T \ell_{t,I(t)} - \sum_{t=1}^T \ell_{t,i^*} \right], \quad R_T = \max_{i^* \in [K]} R_T(i^*). \tag{19}$$

Let $p_t \in \mathcal{P}(K) = \{p \in [0,1]^K \mid \|p\|_1 = 1\}$ denote the distribution from which an action $I(t)$ is chosen, i.e., $\Pr[I(t) = i | \mathcal{H}_{t-1}] = p_{ti}$, where $\mathcal{H}_{t-1} = \{(\ell_s, I(s))\}_{s=1}^{t-1}$. In an *adversarial regime*, the loss ℓ_t can be chosen in an adversarial manner depending on $\mathcal{H}_{t-1}$. Special cases such as stochastic environments, in which ℓ_t independently follows an identical unknown distribution, can be captured in the following regime:

Definition 15 (Adversarial regime with a self-bounding constraint (Zimmert and Seldin, 2021)) *For $\Delta \in \mathbb{R}_{\geq 0}^K$, $C \geq 0$, and $T \in \mathbb{N}$, the environment is in an adversarial regime with a (Δ, C, T) self-boundig constraint if the regret is bounded from below as $R_T \geq R'_T - C$, where we define*

$$R'_T = \mathbf{E}\left[\sum_{t=1}^{T} \Delta_{I(t)}\right] = \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} \Delta_i p_{ti}\right]. \tag{20}$$

As discussed in Zimmert and Seldin (2021, Section 5), this regime includes stochastic environments with adversarial corruption, where each $\Delta_i \geq 0$ represents the suboptimality gap for action i, and C corresponds to the magnitude of corruption. Following prior studies such as (Zimmert and Seldin, 2021) and (Jin et al., 2023), we assume that there is a unique optimal action $i^* \in [K]$, and that $\Delta_i > 0$ holds for all $i \in [K] \setminus \{i^*\}$. Denote $\Delta_{\min} = \min_{i \in [K]\setminus\{i^*\}} \Delta_i$.

4.1. Algorithmic framework for best-of-both-worlds

This subsection provide an algorithmic framework for online learning problems based on FTRL, which has been considered in a variety of problems including multi-armed bandits (Auer et al., 2002; Zimmert and Seldin, 2021), combinatorial semi-bandits (Zimmert et al., 2019), graph bandits (Alon et al., 2017; Eldowa et al., 2023), linear bandits (Cesa-Bianchi and Lugosi, 2012), and contextual bandits (Auer et al., 2002).

Our algorithmic framework computes the probability distribution $q_t \in \mathcal{P}(K)$ given by

$$q_t \in \operatorname*{arg\,min}_{p \in \mathcal{P}(K)} \left\{ \sum_{s=1}^{t-1} \left\langle \hat{\ell}_s, p \right\rangle + \beta_t \psi(p) + \bar{\beta}\bar{\psi}(p) \right\}, \tag{21}$$

where $\hat{\ell}_t$ is an unbiased estimator of ℓ_t. Regularizers ψ and $\bar{\psi}$ are defined as follows:

$$\psi(p) = -\frac{1}{\alpha}\sum_{i=1}^{K}(p_i^{\alpha} - p_i), \quad \bar{\psi}(p) = -\frac{1}{\bar{\alpha}}\sum_{i=1}^{K}(p_i^{\bar{\alpha}} - p_i), \quad \text{where} \quad \alpha \in (0,1), \ \bar{\alpha} = 1 - \alpha. \tag{22}$$

We refer ψ (and $\bar{\psi}$) as the *α-Tsallis entropy* (and the $\bar{\alpha}$-Tsallis entropy) in this paper. The additional regularizer $\bar{\beta}\bar{\psi}$ is introduced to ensure the condition of $h_t = O(h_{t-1})$ is satisfied. Similar techniques, referred to as *hybrid regularizers*, have also been used in existing studies such as Masoudian et al. (2022), Tsuchiya et al. (2023b), and Jin et al. (2023). We then choose an action $I(t) \in [K]$ from the distribution $p_t \in \mathcal{P}(K)$ defined by

$$p_t = (1 - \gamma_t)\, q_t + \gamma_t p_0, \tag{23}$$

where $\gamma_t \in [0, 1/2]$ and $p_0 \in \mathcal{P}(K)$ is a distribution which we refer to as the *exploration basis*. By a standard analysis of FTRL (e.g., Exercise 28.12 in Lattimore and Szepesvári, 2020), we have

$$R_T \leq \mathbf{E}\left[\sum_{t=1}^{T}\left(2\gamma_t + \left\langle \hat{\ell}_t, q_t - q_{t+1}\right\rangle - \beta_t D(q_{t+1}, q_t) + (\beta_t - \beta_{t-1})h_t + \bar{\beta}h'\right)\right], \tag{24}$$

Algorithm 1 FTRL with Tsallis-entropy regularizers and SPM learning rates

```
Require: $K \in \mathbb{N}, 0 \leq \alpha < 1, \beta_1 > 0, \bar{\beta} \geq 0, p_0 \in \mathcal{P}(K)$.
for $t = 1, 2, \ldots$ do
    Compute $q_t \in \mathcal{P}(K)$ given by (21) with $\psi(p)$ and $\bar{\psi}(p)$ defined in (22).
    Set $h_t = -\psi(q_t)$ and $z_t \geq 0$ based on $q_t$. Compute $\gamma_t$ based on $z_t$ and $\beta_t$. Set $p_t$ by (23).
    Choose $I(t)$ so that $\Pr[I(t) = i] = p_{ti}$ and get feedback from the environment.
    Compute $\hat{\ell}_t$ based on the feedback.
    Set $\beta_{t+1}$ by he update rule of (10) with $\hat{h}_{t+1} = h_t$.
end for
```

where $D(p, q) = \psi(p) - \psi(q) - \langle \nabla \psi(q), p - q \rangle$ is the Bregman divergence associated with ψ, and we define $h_t = -\psi(q_t)$ and $h' = -\bar{\psi}(q_1) \leq \frac{1}{\bar{\alpha}} K^{1-\bar{\alpha}}$. We note that $h_t \leq h_1 = h_{\max}$ holds for all t.

To obtain BOBW regret bounds, we design p_0, $\hat{\ell}_t$, α, β_t, $\bar{\beta}$, and $\gamma_t \in [0, 1/2]$, so that

$$h_t = O(h_{t-1}), \quad \mathbf{E}\left[2\gamma_t + \left\langle \hat{\ell}_t, q_t - q_{t+1} \right\rangle - \beta_t D(q_{t+1}, q_t) | \mathcal{H}_{t-1}\right] = O\left(\frac{z_t}{\beta_t}\right) \tag{25}$$

hold for some $z_t \in [0, z_{\max}]$. We then have $R_T \leq \mathbf{E}\left[F(\beta_{1:T}; z_{1:T}, h_{1:T})\right] + \bar{\beta} h'$. By applying the SPM update rule (10) with $\hat{h}_t = h_{t-1}$, we obtain

$$R_T = O\left(\mathbf{E}\left[\min\left\{\sqrt{h_1 \sum_{t=1}^{T} z_t}, \inf_{\varepsilon \geq \frac{1}{T}}\left\{\sqrt{\sum_{t=1}^{T} h_t z_t \log(\varepsilon T) + \frac{h_1 z_{\max}}{\varepsilon}}\right\}\right\}\right] + \kappa\right) \tag{26}$$

as a direct consequence of Theorem 7, where we denote $\kappa = \frac{z_{\max}}{\beta_1} + \beta_1 h_1 + \bar{\beta} h'$. In an adversarial regime with a (Δ, C, T) self-bounding constraint, if

$$h_t z_t \leq \omega(\Delta) \cdot \langle \Delta, q_t \rangle \tag{27}$$

holds for some $\omega(\Delta) > 0$, we have $R_T = O\left(\sqrt{(R_T + C)\omega(\Delta) \log T} + \kappa\right)$, which implies $R_T = O\left(\omega(\Delta) \log T + \sqrt{C\omega(\Delta) \log T} + \kappa\right)$.

The proposed algorithm is summarized in Algorithm 1. We note that the input of p_0 is not required if $\gamma_t = 0$ for all t. Feedback information from the environment and the construction of $\hat{\ell}_t$ vary with each problem setting. From the discussion in this section, we can show that Algorithm 1 achieves BOBW regret bounds as follows:

Proposition 16 *Suppose that (25) holds and that some $z_{\max} > 0$ satisfies $z_t \leq z_{\max}$ for all t with probablity 1. Then Algorithm 1 achieves $R_T = O\left(\mathbf{E}\left[\sqrt{h_1 \sum_{t=1}^{T} z_t + \kappa}\right]\right) \leq O\left(\sqrt{h_1 z_{\max} T} + \kappa\right)$ in adversarial regimes, where $\kappa = \frac{z_{\max}}{\beta_1} + \beta_1 h_1 + \bar{\beta}\bar{h}$. Further, in adversarial regimes with (Δ, C, T) self-bounding constraints, if (27) holds for some $\omega(\Delta)$, Algorithm 1 achieves*

$$R_T = O\left(\omega(\Delta) \log_+\left(\frac{h_1 z_{\max} T}{\omega(\Delta)^2 + C\omega(\Delta)}\right) + \sqrt{C\omega(\Delta) \log_+\left(\frac{h_1 z_{\max} T}{\omega(\Delta)^2 + C\omega(\Delta)}\right)} + \kappa\right). \tag{28}$$

In the subsections below, we use the following notation:

$$\kappa = \frac{z_{\max}}{\beta_1} + \beta_1 h_1 + \bar{\beta}\bar{h}, \quad q_{t*} = \min\left\{ \|q_t\|_\infty, 1 - \|q_t\|_\infty \right\}, \quad \tilde{q}_{ti} = \min\left\{ q_{ti}, q_{t*} \right\}. \tag{29}$$

In the following, we demonstrate that using Algorithm 1, we can achieve the BOBW regret bounds for multi-armed bandit and linear bandit problems as shown in Table 2. The results for graph bandits and for contextual bandits are described in Appendices D.6 and D.7, respectively.

4.2. Multi-armed bandit

In the multi-armed bandit problem, we assume that $\ell_t \in [0,1]^K$ and that the player gets only feedback of the incurred loss of $\ell_{t,I(t)}$. We set arbitrary $\alpha \in (0,1)$ and set

$$\beta_1 \geq \frac{4K}{1-\alpha}, \quad z_t = \frac{1}{1-\alpha} \sum_{i=1}^{K} \tilde{q}_{ti}^{1-\alpha}, \quad \gamma_t = 0, \quad \hat{\ell}_{ti} = \frac{\mathbf{1}[I(t) = i]}{p_{ti}} \ell_{ti}. \tag{30}$$

In addition, we set $\bar{\beta} \geq 0$ as follows:

$$\alpha \leq \frac{1}{2} \implies \bar{\beta} = 0, \qquad \alpha > \frac{1}{2} \implies \bar{\beta} \geq \frac{32K}{(1-\alpha)^2 \beta_1}. \tag{31}$$

As shown in Appendix D.4, conditions (30) and (31) are sufficient conditions for (25). Further, we can show that $h_t = -\psi(q_t)$ and z_t in (30) satisfy $h_1 z_t \leq \frac{K-1}{\alpha(1-\alpha)}$ and (27) with

$$\omega(\Delta) = \frac{2}{\alpha(1-\alpha)} \left(\sum_{i \neq i^*} \Delta_i^{-\frac{\alpha}{1-\alpha}} \right)^{1-\alpha} \left(\sum_{i \neq i^*} \Delta_i^{-\frac{1-\alpha}{\alpha}} \right)^{\alpha} \leq 2 \frac{K-1}{\alpha(1-\alpha)\Delta_{\min}}. \tag{32}$$

Hence, from Proposition 16, we have the following:

Theorem 17 *For the K-armed bandit problem, Algorithm 1 with* (30) *and* (31) *achieves BOBW regret bounds in Proposition 16 with $h_1 z_{\max} = O\left(\frac{K-1}{\alpha(1-\alpha)}\right)$ and $\omega(\Delta)$ given by* (32).

Note that if $\alpha = 1/2$ then $\omega(\Delta) = O\left(\sum_{i \neq i^*} 1/\Delta_i\right)$, which recovers the regret bounds shown by Zimmert and Seldin (2021); Masoudian and Seldin (2021).

4.3. Linear bandit

In the *linear bandit* problems, each arm $i \in [K]$ is associated with a d-dimensional *feature vector* $\phi_i \in \mathbb{R}^d$. The environment in each round determines a *loss vector* $\theta_t \in \mathbb{R}^d$, for which the loss $\ell_{ti} \in [-1,1]$ satisfies $\mathbf{E}[\ell_{ti} | \theta_t] = \langle \theta_t, \phi_i \rangle$. After choosing an arm $I(t)$, the player observes only the incurred loss of $\ell_{t,I(t)}$. Without loss of generality, we assume that $d \leq K$ and that $\{\phi_i\}_{i=1}^K$ spans $\mathbb{R}^d$. For any distribution $p \in \mathcal{P}(K)$, denote

$$S(p) = \sum_{i=1}^{K} p_i \phi_i \phi_i^\top = \mathop{\mathbf{E}}_{I \sim p}\left[\phi_I \phi_I^\top \right]. \tag{33}$$

Then, there exists a distribution $p \in \mathcal{P}(K)$ such that $\phi_i^\top S(p)^{-1}\phi_i \le d$ (see, e.g., Lattimore and Szepesvári, 2020, Theorem 21.1). We choose $p_0 \in \mathcal{P}(K)$ so that

$$\phi_i^\top S(p_0)^{-1}\phi_i \le cd \quad (i \in [K]) \tag{34}$$

holds for some $c = O(1)$. Let $\alpha \ge 1/2$ and set

$$\beta_1 \ge \frac{8cd}{1-\alpha},\ \bar{\beta} \ge \frac{32d}{(1-\alpha)^2\beta_1},\ z_t = \frac{dq_{t*}^{1-\alpha}}{1-\alpha},\ \gamma_t = \frac{4cz_t}{\beta_t},\ \hat{\ell}_{ti} = \ell_{t,I(t)}\phi_{I(t)}^\top S(p_t)^{-1}\phi_i. \tag{35}$$

If p_0 satisfies (34) and if parameters are given by (35), then (25) holds. Further, $h_t = -\psi(q_t)$ and z_t in (35) satisfy $h_1 z_t \le \frac{d}{\alpha(1-\alpha)}K^{1-\alpha}$ and (27) with $\omega(\Delta)$ defined as

$$\omega(\Delta) = \frac{d}{\alpha(1-\alpha)}\Delta_{\min}^{\alpha-1}\left(\sum_{i\neq i^*}\Delta_i^{-\frac{\alpha}{1-\alpha}}\right)^{1-\alpha} \le \frac{dK^{1-\alpha}}{\alpha(1-\alpha)\Delta_{\min}}. \tag{36}$$

Hence, Proposition 16 leads to the following regret bounds:

Theorem 18 *For linear bandit problems of K arms associated with d-dimensional vectors, Algorithm 1 with p_0 satisfying (34) and parameters given by (35) achieves BOBW regret bounds in Proposition 16 with $h_1 z_{\max} = O\left(\frac{dK^{1-\alpha}}{\alpha(1-\alpha)}\right)$ and $\omega(\Delta)$ given by (36).*

Note that we obtain $\frac{dK^{1-\alpha}}{\alpha(1-\alpha)} = O(d\log K)$ by setting $\alpha = 1 - \frac{1}{4\log K}$, which recovers the regret upper bound by Dann et al. (2023, Corollary 12).

5. Conclusion

In this paper, we formulated a sequential decision-making problem of adjusting the learning rate in FTRL. For this problem, we showed that simple strategies, referred to as stability-penalty matching, achieve optimal performance in terms of the worst-case competitive ratio. We also demonstrated that these strategies for adaptive learning rates, combined with Tsallis-entropy regularizers, are useful in designing best-of-both-worlds algorithms for various bandit problems, including multi-armed bandits, graph bandits, linear bandits, and contextual bandits. The proposed approach is expected to have a wider range of applications, including use in further problem setups and the design of algorithms to achieve data-dependent regret bounds.

Acknowledgments

JH was supported by JSPS, KAKENHI Grant Number JP21K11747, Japan. TT was supported by JST, ACT-X Grant Number JPMJAX210E, Japan.

References

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Conference on Learning Theory*, 2008.

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012.

Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems*, volume 28, pages 2197–2205, 2015.

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Algorithmic Learning Theory*, pages 229–243, 2006.

Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

Idan Amir, Guy Azov, Tomer Koren, and Roi Livni. Better best of both worlds bounds for bandits with switching costs. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *International Conference on Machine Learning*, pages 217–226, 2009.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

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

Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *Conference on Learning Theory*. PMLR, 2023.

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

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

Khaled Eldowa, Emmanuel Esposito, Tommaso Cesari, and Nicolò Cesa-Bianchi. On the minimax regret for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tim Erven, Wouter M Koolen, Steven Rooij, and Peter Grünwald. Adaptive hedge. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Pierre Gaillard, Gilles Stoltz, and Tim Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196. PMLR, 2014.

Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *International Conference on Machine Learning*, volume 162, pages 9173–9200, 2022.

Shinji Ito. Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits. In *Advances in Neural Information Processing Systems*, volume 34, pages 2654–2667, 2021a.

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

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In *Conference on Learning Theory*, pages 1421–1422. PMLR, 2022a.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, 2022b.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic MDPs with known transition. *Advances in Neural Information Processing Systems*, 33:16557–16566, 2020.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic MDPs with unknown transition. *Advances in Neural Information Processing Systems*, 34, 2021.

Tiancheng Jin, Junyan Liu, and Haipeng Luo. Improved best-of-both-worlds guarantees for multi-armed bandits: FTRL with general regularizers and multiple optimal arms. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Fang Kong, Canzhe Zhao, and Shuai Li. Best-of-three-worlds analysis for linear bandits with follow-the-regularized-leader algorithm. In *Conference on Learning Theory*. PMLR, 2023.

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.

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

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in Neural Information Processing Systems*, 33:15522–15533, 2020.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Saeed Masoudian and Yevgeny Seldin. Improved analysis of the tsallis-inf algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 3330–3350. PMLR, 2021.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory*, pages 244–256, 2010.

Julia Olkhovskaya, Jack Mayo, Tim van Erven, Gergely Neu, and Chen-Yu Wei. First-and second-order bounds for adversarial linear contextual bandits. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Francesco Orabona and Dávid Pál. Scale-free algorithms for online linear optimization. In *International Conference on Algorithmic Learning Theory*, pages 287–301. Springer, 2015.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Chloé Rouyer and Yevgeny Seldin. Tsallis-INF for decoupled exploration and exploitation in multi-armed bandits. In *Conference on Learning Theory*, volume 125, pages 3227–3249, 2020.

Chloé Rouyer, Yevgeny Seldin, and Nicolò Cesa-Bianchi. An algorithm for stochastic and adversarial bandits with switching costs. In *International Conference on Machine Learning*, pages 9127–9135. PMLR, 2021.

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *International Conference on Machine Learning*, pages 19011–19026. PMLR, 2022.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Best-of-both-worlds algorithms for partial monitoring. In *International Conference on Algorithmic Learning Theory*, pages 1484–1515. PMLR, 2023a.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

Julian Zimmert and Tor Lattimore. Connections between mirror descent, Thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems 32*, pages 11973–11982, 2019.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 3285–3294. PMLR, 2020.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692. PMLR, 2019.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936, 2003.

Appendix A. Additional Related Work

Online Learning using Tsallis entropy To the best of our knowledge, the use of Tsallis entropy in online learning is first considered by Audibert and Bubeck (2009); Abernethy et al. (2015), in which they showed that FTRL with Tsallis entropy can achieve an $O(\sqrt{KT})$ regret in multi-armed bandits.

After that Tsallis entropy has been employed in many online decision-making problems: FTRL with Tsallis entropy of exponent $\alpha = 1 - 1/\log(K/s)$, was used to exploit the sparsity of losses, $s = |\{i \in [K] \mid \ell_{ti} \neq 0\}|$, in multi-armed bandits (Kwon and Perchet, 2016), and FTRL with $(1-1/\log K)$-Tsallis entropy was used to obtain an improved regret bound in the strongly observable graph bandit problem (Zimmert and Lattimore, 2019).

The most relevant studies to this paper are ones aimed at constructing BOBW algorithms using FTRL with Tsallis entropy. Zimmert and Seldin (2021) showed for the first time that FTRL with $1/2$-Tsallis entropy can achieve a nearly optimal logarithmic regret, whose regret bound in stochastic regimes with adversarial corruptions is later improved by Masoudian and Seldin (2021). FTRL with $1/2$-Tsallis entropy was also proven to be powerful in combinatorial semi-bandits (Zimmert et al., 2019), in the delayed feedback setting, where the loss of the selected action is observed after a delay (Zimmert and Seldin, 2020; Masoudian et al., 2022), in multi-armed bandits with switching costs, where the learner needs to pay a cost when changing their actions (Rouyer et al., 2021; Amir et al., 2022), dueling bandits (Saha and Gaillard, 2022), and MDPs (Jin and Luo, 2020; Jin et al., 2021).

In addition to these applications, it is known that in the decoupling setting, where different actions can be chosen for exploration and exploitation, FTRL with $2/3$-Tsallis entropy can achieve a constant regret bound (Rouyer and Seldin, 2020). Interestingly, even in the setting of heavy-tailed multi-armed bandits, where the n-th moment of loss is bounded by σ^n for some $\sigma > 0$, FTRL with Tsallis entropy with exponent $1/n$ can achieve a logarithmic regret (Huang et al., 2022). Furthermore, for the weakly observable setting in graph bandits and for the globally observable setting in partial monitoring, whose minimax regrets are $\Theta(T^{2/3})$, FTRL with $1/2$-Tsallis entropy and the complement version of Tsallis entropy play key roles in achieving BOBW guarantees (Ito et al., 2022b; Tsuchiya et al., 2023a).

Adaptive Learning Rate Using an adaptive learning rate is one of the most common ways to design algorithms with a desired adaptivity. In the literature, it has been standard to determine the adaptive learning rate by relying on the stability component in (2) observed so far. Typical examples are AdaGrad (McMahan and Streeter, 2010; Duchi et al., 2011) in online convex optimization and its closely related algorithms that can achieve the first-order bounds (Allenberg et al., 2006; Abernethy et al., 2012; Wei and Luo, 2018) and the second-order bounds (Cesa-Bianchi et al., 2007; Erven et al., 2011; De Rooij et al., 2014; Gaillard et al., 2014; Orabona and Pál, 2015; Ito et al., 2022b; Olkhovskaya et al., 2023)

In contrast, some very recent studies improve the adaptivity of algorithms by designing an adaptive learning rate depending on the penalty component in (2), instead of the stability term. To our knowledge, Ito et al. (2022b) is the first attempt for such a design, where the authors aimed at constructing BOBW algorithms. A natural question that arises here is whether we can construct an adaptive learning rate that depends on both the stability and penalty terms.

The stability-penalty-adaptive (SPA) learning rate is the first adaptive learning rate that can achieve such simultaneous adaptivity (Tsuchiya et al., 2023b). With the SPA learning rate, they proved that (2) is roughly bounded by $O(\sqrt{\sum_{t=1}^{T} z_t h_{t+1} \log T})$. A comparison of the SPA learning rate and the SPM learning rate is discussed in the following.

Comparison of SPM learning rate against SPA learning rate There are several issues in the existing adaptive learning rate that depends on the penalty term. The biggest issue is that their regret upper bounds in the adversarial regime (or in the worst-case) have extra $O(\sqrt{\log T})$ factors, which is due to the loose analysis or the "ad-hoc" learning rate designs. Although the SPA learning rate is designed in a generic form so that it can be used for generic regularizers, the authors focus only on the Shannon entropy, not investigating the use of Tsallis entropy. As mentioned earlier, Tsallis entropy has been proven to be powerful in many BOBW algorithms, and our adaptive learning rate framework could be used for a wide range of online decision-making problems besides those presented in the paper.

At a high level, this study provides a non-ad-hoc, theoretically grounded adaptive learning rate design principle by rethinking the design of adaptive learning rate from the standpoint of competitive analysis. Consequently, we succeeded in constructing nearly optimal BOBW algorithms, totally removing the suboptimality caused by the existing ad-hoc design of adaptive learning rates.

Parameter-free online learning algorithms Both parameter-free algorithms (Orabona and Pál, 2016; Orabona, 2019) and ours can be interpreted as providing adaptivity to the penalty terms, but the focus is slightly different. Penalty terms are given as $h_t(u^*) := \psi(u^*) - \psi(x_t)$, where ψ is the regularizer, u^* is the comparator, and x_t is the output of the algorithm. Parameter-free algorithms achieve regret bounds depending on $h_1(u^*)$ for any u^* i.e., these are adaptive to unknown values of $\psi(u^*)$ (which corresponds to comparator norm when $\psi(x) = \|x\|_2^2$), but not to $\{\psi(x_t)\}_{t=2}^{T}$. In contrast, our study leads to regret bounds depending on $\{h_t(u^*)\}_{t=1}^{T}$ under the assumption that $\psi(u^*)$ is known (which can be justified when ψ is the negative Shannon / Tsallis entropy as $\psi(u^*) = 0$ for the standard basis), i.e., these are adaptive to $\{\psi(x_t)\}_{t=1}^{T}$ but not to $\psi(u^*)$. The adaptivity to $\{\psi(x_t)\}_{t=1}^{T}$ plays an important role in showing BOBW bounds. Designing an algorithm that achieves both adaptabilities simultaneously is an interesting future research direction to consider.

Appendix B. Lower Bound on the Competitive Ratio

This section provides a proof of Theorem 1, which provide a lower bound on the competitive ratio. Note here that we use the notation $\beta_t = 1/\eta_t$ as an alternative to η_t, as introduced in Section 2.

Proof of Theorem 1 Consider two problem instances $(z_{1:T}, h_{1:T})$ and $(z'_{1:T}, h_{1:T})$ defined as follows: $z_1 = z'_1 = 1$, $h_1 = 1$, and $z_t = 0$, $z'_t = 1$, $h_t = \xi$ for $t = 2, \ldots, T$. We then have

$$F^*(z_{1:T}, h_{1:T}) = 2, \quad F^*(z'_{1:T}, h_{1:T}) \leq \min_{\beta_1} \left\{ \frac{T}{\beta_1} + \beta_1 \right\} = 2\sqrt{T}. \tag{37}$$

For a policy π, denote $\beta_1 = \pi_1(z_1, h_1) = \pi_1(z'_1, h_1) = \pi_1(1, 1)$. We then have

$$\begin{aligned} F^\pi(z_{1:T}, h_{1:T}) &\geq \frac{1}{\beta_1} + \beta_1, \\ F^\pi(z'_{1:T}, h_{1:T}) &\geq \min_{\beta_T \geq \beta_1} \left\{ \frac{1}{\beta_1} + \beta_1 + \frac{T-1}{\beta_T} + (\beta_T - \beta_1)\xi \right\} \\ &\geq \frac{1}{\beta_1} + \beta_1 + 2\sqrt{\xi(T-1)} - \xi\beta_1 \geq 2\sqrt{\xi(T-1)} - \xi\beta_1. \end{aligned} \tag{38}$$

We hence have

$$\inf_{\pi} \sup_{z_{1:T}\in\mathbb{R}^T_{\geq 0}, h_{1:T}\in H^T_\xi} \mathrm{CR}(\pi; z_{1:T}, h_{1:T}) \geq \inf_{\pi} \max\left\{\frac{F^\pi(z_{1:T}, h_{1:T})}{F^*(z_{1:T}, h_{1:T})}, \frac{F^\pi(z'_{1:T}, h_{1:T})}{F^*(z'_{1:T}, h_{1:T})}\right\}$$
$$\geq \inf_{\beta>0} \max\left\{\frac{1}{2}\left(\frac{1}{\beta}+\beta\right), \sqrt{\frac{\xi(T-1)}{T}} - \frac{\xi\beta}{2\sqrt{T}}\right\}$$
$$\geq \inf_{\beta>0} \max\left\{\frac{\beta}{2}, \sqrt{\frac{\xi(T-1)}{T}} - \frac{\xi\beta}{2\sqrt{T}}\right\} = \frac{\sqrt{T}}{\sqrt{T}+\xi}\sqrt{\frac{\xi(T-1)}{T}} = \frac{\sqrt{T-1}}{\sqrt{T}+\xi}\sqrt{\xi}. \tag{39}$$

■

Appendix C. Omitted Proofs in Sections 2 and 3

C.1. Proof of Lemma 5

Proof Let $\tilde{h}_t = \min_{s\in[t]} h_s$. Then it is clear that $\tilde{h}_t \leq h_t$ and $\tilde{h}_{t+1} \leq \tilde{h}_t$. Further, it follows from the assumption of $h_{1:T} \in H^T_\xi$ and the definition of H^T_ξ in (8) that

$$\xi\tilde{h}_t = \min\left\{\xi h_t, \min_{s<t}\{\xi h_s\}\right\} \geq \min\{\xi h_t, h_t\} = h_t, \tag{40}$$

which completes the proof. ■

C.2. Proof of Lemma 9

Proof We first consider the case in which β_t is given by (9). We then have

$$F(\beta_{1:T}; z_{1:T}, h_{1:T}) = \sum_{t=1}^T \left(\frac{z_t}{\beta_t} + (\beta_t - \beta_{t-1})h_t\right) \leq \sum_{t=1}^T \left(\frac{z_t}{\beta_t} + (\beta_t - \beta_{t-1})\hat{h}_t\right) = 2\sum_{t=1}^T \frac{z_t}{\beta_t}. \tag{41}$$

Further, it follows from (9) that

$$\beta_t^2 = \beta_t\beta_{t-1} + \frac{z_t}{\hat{h}_t} \geq \beta_{t-1}^2 + \frac{z_t}{\hat{h}_t} = \sum_{s=1}^t \frac{z_s}{\hat{h}_s}. \tag{42}$$

By combining (41) and (42), we obtain $F(\beta_{1:T}; z_{1:T}, h_{1:T}) \leq 2G(z_{1:T}, \hat{h}_{1:T})$.

We next consider the case of (10). We then have

$$F(\beta_{1:T}; z_{1:T}, h_{1:T}) \leq \frac{z_1}{\beta_1} + \beta_1 h_1 + \sum_{t=2}^T \left(\frac{z_t}{\beta_t} + (\beta_t - \beta_{t-1})\hat{h}_t\right)$$
$$= \frac{z_1}{\beta_1} + \beta_1 h_1 + \sum_{t=2}^T \left(\frac{z_t}{\beta_t} + \frac{z_{t-1}}{\beta_{t-1}}\right)$$
$$\leq \beta_1 h_1 + 2\sum_{t=1}^T \frac{z_t}{\beta_t}. \tag{43}$$

Further, for any $t \geq 2$, it follows from (10) that

$$\beta_t^2 = \beta_{t-1}^2 + 2\frac{z_{t-1}}{\hat{h}_t} + \left(\frac{z_{t-1}^2}{\beta_{t-1}\hat{h}_t}\right)^2 \geq \beta_{t-1}^2 + 2\frac{z_{t-1}}{\hat{h}_t} = \beta_1^2 + 2\sum_{s=2}^{t}\frac{z_{s-1}}{\hat{h}_s}, \tag{44}$$

which implies that $\beta'_t := \sqrt{\beta_1^2 + 2\sum_{s=2}^{t}\frac{z_{s-1}}{\hat{h}_s}}$ satisfies $\beta'_t \leq \beta_t$. Denote $\mathcal{T} = \{t \in [T] \mid \beta'_{t+1} \leq \sqrt{2}\beta'_t\}$ and $\mathcal{T}^c = [T] \setminus \mathcal{T} = \{t \in [T] \mid \beta'_{t+1} > \sqrt{2}\beta'_t\}$. We then have

$$\begin{aligned}
\sum_{t=1}^{T}\frac{z_t}{\beta_t} \leq \sum_{t=1}^{T}\frac{z_t}{\beta'_t} &= \sum_{t\in\mathcal{T}}\frac{z_t}{\beta'_t} + \sum_{t\in\mathcal{T}^c}\frac{z_t}{\beta'_t} \\
&\leq \sqrt{2}\sum_{t\in\mathcal{T}}\frac{z_t}{\beta'_{t+1}} + \sum_{t\in\mathcal{T}^c}\frac{z_{\max}}{\beta'_t} \\
&\leq \sqrt{2}\sum_{t\in\mathcal{T}}\frac{z_t}{\beta'_{t+1}} + \sum_{s=0}^{\infty}\left(\frac{1}{\sqrt{2}}\right)^s\frac{z_{\max}}{\beta_1} \\
&\leq \sqrt{2}\sum_{t\in\mathcal{T}}\frac{z_t}{\beta'_{t+1}} + \frac{1}{1-1/\sqrt{2}}\frac{z_{\max}}{\beta_1} \leq G(z_{1:T}, \hat{h}_{2:T+1}) + (2+\sqrt{2})\frac{z_{\max}}{\beta_1}.
\end{aligned}$$

Combining this with (43), we obtain $F(\beta_{1:T}; z_{1:T}, h_{1:T}) \leq 2G(z_{1:T}, \hat{h}_{2:T+1}) + 7\frac{z_{\max}}{\beta_1} + \beta_1 h_1$. ■

C.3. Proof of Lemma 10

Proof The inequalities of the lemma can be shown as follows:

$$\begin{aligned}
G(z_{1:T}, h_{1:T}) &= \sum_{j=1}^{J+1}\sum_{t\in\mathcal{T}_j}\left(\sum_{s=1}^{t}\frac{z_s}{h_s}\right)^{-1/2} z_t \leq \sum_{j=1}^{J+1}\sum_{t\in\mathcal{T}_j}\left(\sum_{s\in\mathcal{T}_j\cap[t]}\frac{z_s}{h_s}\right)^{-1/2} z_t \\
&\leq \sum_{j=1}^{J+1}\sum_{t\in\mathcal{T}_j}\left(\sum_{s\in\mathcal{T}_j\cap[t]}\frac{z_s}{\theta_{j-1}}\right)^{-1/2} z_t = \sum_{j=1}^{J+1}\sqrt{\theta_{j-1}}\sum_{t\in\mathcal{T}_j}\frac{z_t}{\sqrt{\sum_{s\in\mathcal{T}_j\cap[t]} z_s}} \\
&\leq 2\sum_{j=1}^{J+1}\sqrt{\theta_{j-1}}\sum_{t\in\mathcal{T}_j}\frac{z_t}{\sqrt{\sum_{s\in\mathcal{T}_j\cap[t]} z_s} + \sqrt{\sum_{s\in\mathcal{T}_j\cap[t-1]} z_s}} \\
&\leq 2\sum_{j=1}^{J+1}\sqrt{\theta_{j-1}}\sum_{t\in\mathcal{T}_j}\left(\sqrt{\sum_{s\in\mathcal{T}_j\cap[t]} z_s} - \sqrt{\sum_{s\in\mathcal{T}_j\cap[t-1]} z_s}\right) \leq 2\sum_{j=1}^{J+1}\sqrt{\theta_{j-1}\sum_{t\in\mathcal{T}_j} z_t}.
\end{aligned}$$

By setting $J = 0$ and $\theta_0 = h_{\max}$, we obtain

$$G(z_{1:T}, h_{1:T}) \leq 2\sqrt{h_{\max}\sum_{t=1}^{T} z_t}. \tag{45}$$

By setting $\theta_j = 2^{-j} h_{\max}$ for $j = 0, 1, \ldots, J$, we have

$$\begin{aligned} G(z_{1:T}, h_{1:T}) &\leq 2 \sum_{j=1}^{J+1} \sqrt{\theta_{j-1} \sum_{t \in \mathcal{T}_j} z_t} \leq 2 \sum_{j=1}^{J} \sqrt{\frac{\theta_{j-1}}{\theta_j} \sum_{t \in \mathcal{T}_j} h_t z_t} + 2 \sqrt{\theta_J \sum_{t \in \mathcal{T}_J} z_t} \\ &= 2 \sum_{j=1}^{J} \sqrt{2 \sum_{t \in \mathcal{T}_j} h_t z_t} + 2 \sqrt{2^{-J} h_{\max} \sum_{t \in \mathcal{T}_J} z_t} \\ &\leq 2 \sqrt{2J \sum_{j=1}^{J} \sum_{t \in \mathcal{T}_j} h_t z_t} + 2 \sqrt{2^{-J} h_{\max} \sum_{t \in \mathcal{T}_J} z_t} \\ &\leq \sqrt{8J \sum_{t=1}^{T} h_t z_t} + 2 \sqrt{2^{-J} h_{\max} z_{\max} T}, \end{aligned} \tag{46}$$

where the second inequality follows from $h_t > \theta_j$ for $j \in \mathcal{T}_j$ and third inequality can be shown from the Cauchy-Schwarz inequality. ■

C.4. Proof of Lemma 11

Proof Define $\tau(j) = \max\{t \in [T] \mid h_t > \theta_j\}$ for $j = 1, 2, \ldots, J$ and set $\tau(0) = 0$ and $\tau(J+1) = T$. We then have $\mathcal{T}_j = \{\tau(j-1)+1, \ldots, \tau(j)\}$ for $j = 1, 2, \ldots, J+1$. For any non-decreasing sequence $\beta_{1:T} \in \mathbb{R}^T_{>0}$, we have

$$\begin{aligned} F(\beta_{1:T}; z_{1:T}, h_{1:T}) &= \sum_{t=1}^{T} \left(\frac{z_t}{\beta_t} + (\beta_t - \beta_{t-1}) h_t \right) \geq \sum_{j=1}^{J} \sum_{t=\tau(j-1)+1}^{\tau(j)} \left(\frac{z_t}{\beta_t} + (\beta_t - \beta_{t-1}) h_t \right) \\ &\geq \sum_{j=1}^{J} \sum_{t=\tau(j-1)+1}^{\tau(j)} \left(\frac{z_t}{\beta_{\tau(j)}} + (\beta_t - \beta_{t-1}) \theta_j \right) \\ &= \sum_{j=1}^{J} \left(\frac{1}{\beta_{\tau(j)}} \sum_{t=\tau(j-1)+1}^{\tau(j)} z_t + \left(\beta_{\tau(j)} - \beta_{\tau(j-1)} \right) \theta_j \right) \\ &\geq \sum_{j=1}^{J} \left(\frac{1}{\beta_{\tau(j)}} \sum_{t=\tau(j-1)+1}^{\tau(j)} z_t + \beta_{\tau(j)} (\theta_j - \theta_{j+1}) \right) + \beta_{\tau(J)} \theta_{J+1} - \beta_{\tau(0)} \theta_1 \\ &\geq 2 \sum_{j=1}^{J} \sqrt{(\theta_j - \theta_{j+1}) \sum_{t=\tau(j-1)+1}^{\tau(j)} z_t} = 2 \sum_{j=1}^{J} \sqrt{(\theta_j - \theta_{j+1}) \sum_{t \in \mathcal{T}_j} z_t}, \end{aligned} \tag{47}$$

where the last inequality follows from the AM-GM inequality and the fact that $\beta_{\tau(0)} = \beta_0 = 0$. ■

C.5. Proof of Lemma 13

Proof We first suppose that $h_{1:T}$ is monotone non-increasing. Then, from Lemma 11 with $\theta_j = h_{\max} 2^{-j}$, we have

$$F^*(z_{1:T}, h_{1:T}) \geq 2 \sum_{j=1}^{\infty} \sqrt{(\theta_j - \theta_{j+1}) \sum_{t \in \mathcal{T}_j} z_t} = 2 \sum_{j=1}^{\infty} \sqrt{(2^{-1}\theta_{j-1} - 2^{-2}\theta_{j-1}) \sum_{t \in \mathcal{T}_j} z_t}$$
$$\geq \sum_{j=1}^{\infty} \sqrt{\theta_{j-1} \sum_{t \in \mathcal{T}_j} z_t} = H(z_{1:T}, h_{1:T}). \tag{48}$$

Further, as Lemma 10 implies $G(z_{1:T}, h_{1:T}) \leq 2H(z_{1:T}, h_{1:T})$, we have $G(z_{1:T}, h_{1:T}) \leq 2F^*(z_{1:T}, h_{1:T})$ for non-increasing sequence $h_{1:T}$.

We next consider the case in which $h_{1:T}$ is α-approximately non-increasing. Define $\tilde{h}_t = \min_{s \in [t]} h_t$. Then $\tilde{h}_{1:T}$ is monotone non-increasing and it holds for any t that $\tilde{h}_t \leq h_t \leq \alpha \tilde{h}_t$. We hence have

$$G(z_{1:T}, h_{1:T}) \leq G(z_{1:T}, \alpha\tilde{h}_{1:T}) \leq 2F^*(z_{1:T}, \alpha\tilde{h}_{1:T}) = 2\sqrt{\alpha} F^*(z_{1:T}, \tilde{h}_{1:T}) \leq 2\sqrt{\alpha} F^*(z_{1:T}, h_{1:T}), \tag{49}$$

which completes the proof. ■

C.6. Proof of Lemma 14

Proof Denote $\tau(j) = \max\{t \in [T] \mid h_t > \theta_j\}$ and $\tau'(j) = \max\{t \in [T] \mid h_{t+1} > \theta_j\}$ for $j \geq 1$ and $\tau(0) = \tau'(0) = 0$. We then have $\tau'(j) \leq \tau(j) \leq \tau'(j) + 1$. We hence have

$$H(z_{1:T}, h_{1:T}) = \sum_{j=1}^{\infty} \sqrt{\theta_{j-1} \sum_{t=\tau(j-1)+1}^{\tau(j)} z_t} \leq \sum_{j=1}^{\infty} \sqrt{\theta_{j-1} \sum_{t=\tau'(j-1)+1}^{\tau'(j)+1} z_t}$$
$$\leq \sum_{j=1}^{\infty} \left(\sqrt{\theta_{j-1} \sum_{t=\tau'(j-1)+1}^{\tau'(j)} z_t} + \sqrt{\theta_{j-1} z_{\max}} \right)$$
$$= H(z_{1:T}, h_{2:T+1}) + \sqrt{h_{\max} z_{\max}} \sum_{j=1}^{\infty} \sqrt{2^{1-j}}$$
$$\leq H(z_{1:T}, h_{2:T+1}) + 4\sqrt{h_{\max} z_{\max}}. \tag{50}$$

■

C.7. Proof of Theorem 7

Proof The bounds on F of (12) and (13) immediately follow from Lemmas 9 and 10. In the following, we show bounds that depend on F^*. Suppose $h_{1:T} \in H_{\xi}^{T}$. Then, $\tilde{h}_t := \min_{s \in [t]} h_s$

satisfies $\tilde{h}_t \leq h_t \leq \xi \tilde{h}_t \leq \xi \tilde{h}_{t-1}$ and $\tilde{h}_{1:T} \in H_1^T$, i.e., $\tilde{h}_t \geq \tilde{h}_{t+1}$. Hence, if $\beta_{1:T}$ is given by (9) with $\hat{h}_t = h_t$, we have

$$F(\beta_{1:T}; z_{1:T}, h_{1:T}) \leq 2G(z_{1:T}, h_{1:T}) \leq 4\sqrt{\xi} F^*(z_{1:T}, h_{1:T}), \tag{51}$$

where the first and second inequalities follow from Lemmas 9 and 13, respectively. This means that (a) in Theorem 7 holds. We next see that (c) holds. If $\beta_{1:T}$ is given by (9) with (11) we then have

$$\begin{aligned}
F(\beta_{1:T}; z_{1:T}, h_{1:T}) &\leq 2G(z_{1:T}, \hat{h}_{1:T}) && \text{(Lemma 9)} \\
&= 2G(z_{1:T}, \xi \tilde{h}_{0:T-1}) = 2\sqrt{\xi} G(z_{1:T}, \tilde{h}_{0:T-1}) && \text{(Definitions of } \tilde{h}_t \text{ and } G \text{ in (15))} \\
&\leq 4\sqrt{\xi} H(z_{1:T}, \tilde{h}_{0:T-1}) && \text{(Lemma 10)} \\
&\leq 4\sqrt{\xi} \left(H(z_{1:T}, \tilde{h}_{1:T}) + 4\sqrt{h_{\max} z_{\max}} \right) && \text{(Lemma 14)} \\
&\leq 4\sqrt{\xi} \left(F^*(z_{1:T}, \tilde{h}_{1:T}) + 4\sqrt{h_{\max} z_{\max}} \right) && \text{(Lemma 11)} \\
&\leq 4\sqrt{\xi} \left(F^*(z_{1:T}, h_{1:T}) + 4\sqrt{h_{\max} z_{\max}} \right), && \text{(Definition of } F^* \text{ and } \tilde{h}_t \leq h_t)
\end{aligned}$$

which completes the proof of (c) in Theorem 7. We next show (b) and (d) by considering the case in which β_t is given by (10). Denote $\kappa = \frac{z_{\max}}{\beta_1} + \beta_1 h_1$. If β_t is given by (10) with $\hat{h}_t = h_t$, we have

$$\begin{aligned}
F(\beta_{1:T}; z_{1:T}, h_{1:T}) &\leq 2G(z_{1:T}, h_{2:T+1}) + O(\kappa) && \text{(Lemma 9)} \\
&\leq 2G(z_{1:T}, \xi \tilde{h}_{1:T}) + O(\kappa) && (h_t \leq \xi \tilde{h}_{t-1}) \\
&= 2\sqrt{\xi} G(z_{1:T}, \tilde{h}_{1:T}) + O(\kappa) && \text{(Definition (15) of } G) \\
&\leq 2\sqrt{\xi} F^*(z_{1:T}, \tilde{h}_{1:T}) + O(\kappa) && \text{(Lemmas 10 and 11)} \\
&\leq 2\sqrt{\xi} F^*(z_{1:T}, h_{1:T}) + O(\kappa). && \text{(Definition of } F^* \text{ and } \tilde{h}_t \leq h_t)
\end{aligned}$$

This means that (b) in Theorem 7 holds. We next show (d). If $\beta_{1:T}$ is given by (10) with (11), we have

$$\begin{aligned}
F(\beta_{1:T}; z_{1:T}, h_{1:T}) &\leq 2G(z_{1:T}, \hat{h}_{2:T+1}) + O(\kappa) && \text{(Lemma 9)} \\
&\leq 2G(z_{1:T}, \xi \tilde{h}_{1:T}) + O(\kappa) && \text{(Definition of } \hat{h}_t) \\
&= 2\sqrt{\xi} G(z_{1:T}, \tilde{h}_{1:T}) + O(\kappa) && \text{(Definition (15) of } G) \\
&\leq 2\sqrt{\xi} F^*(z_{1:T}, \tilde{h}_{1:T}) + O(\kappa) && \text{(Lemmas 10 and 11)} \\
&\leq 2\sqrt{\xi} F^*(z_{1:T}, h_{1:T}) + O(\kappa), && \text{(Definition of } F^* \text{ and } \tilde{h}_t \leq h_t)
\end{aligned}$$

which completes the proof of (d) of Theorem 7. ■

Appendix D. Analysis for Algorithm 1: FTRL with SPM Learning Rates

D.1. Facts on FTRL

Lemma 19 *Suppose q_t is given by (21). Then, it holds for any $p^* \in \mathcal{P}(K)$ that*

$$\begin{aligned}&\sum_{t=1}^{T} \left\langle \hat{\ell}_t, q_t - p^* \right\rangle \\ &\leq \sum_{t=1}^{T} \left(\left\langle \hat{\ell}_t, q_t - q_{t+1} \right\rangle - \beta_t D(q_{t+1}, q_t) + (\beta_t - \beta_{t-1})(\psi(p^*) - \psi(q_t)) \right) + \bar{\psi}(p^*) - \bar{\psi}(q_1),\end{aligned}$$

where $D(p,q)$ denotes the Bregman divergence associated with ψ.

Proof We can apply a standard analytical technique, e.g., in the proof of Lemma 1 by Ito (2021a), as follows:

$$\begin{aligned}&\left\langle \sum_{t=1}^{T} \hat{\ell}_t, p^* \right\rangle + \beta_T \psi(p^*) + \bar{\psi}(p^*) \\ &\geq \left\langle \sum_{t=1}^{T} \hat{\ell}_t, q_{T+1} \right\rangle + \beta_T \psi(q_{T+1}) + \bar{\psi}(q_{T+1}) \\ &= \left\langle \sum_{t=1}^{T-1} \hat{\ell}_t, q_{T+1} \right\rangle + \left\langle \hat{\ell}_T, q_{T+1} \right\rangle + \beta_T \psi(q_{T+1}) + \bar{\psi}(q_{T+1}) \\ &\geq \left\langle \sum_{t=1}^{T-1} \hat{\ell}_t, q_T \right\rangle + \left\langle \hat{\ell}_T, q_{T+1} \right\rangle + \beta_T \psi(q_T) + \bar{\psi}(q_T) + \beta_T D(q_{T+1}, q_T) \\ &\geq \sum_{t=1}^{T} \left(\left\langle \hat{\ell}_t, q_{t+1} \right\rangle + \beta_t D(q_{t+1}, q_t) + (\beta_{t-1} - \beta_t)\, \psi(q_t) \right) + \bar{\psi}(q_1),\end{aligned}$$

which implies that the desired inequality holds. ■

D.2. Facts on Tsallis entropy

When ψ is given by (22), then the Bregman divergence associated with ψ is given by

$$D(p,q) = \frac{1}{\alpha} \sum_{i=1}^{K} \left(q_i^\alpha + \alpha(p_i - q_i) q_i^{\alpha-1} - p_i^\alpha \right) = \sum_{i=1}^{K} d(p_i, q_i), \tag{52}$$

where we define

$$d(p,q) := \alpha^{-1} q^\alpha + (p-q) q^{\alpha-1} - \alpha^{-1} p^\alpha \leq \frac{1-\alpha}{2} \left(\min\{p,q\}\right)^{\alpha-2} (p-q)^2. \tag{53}$$

Lemma 20 (stability for one dimensional case) *Let $p, q \in (0,1)$. Suppose $\ell \geq -\frac{1-\alpha}{2} q^{\alpha-1}$. We then have*

$$\ell \cdot (q-p) - d(p,q) \leq \frac{2q^{2-\alpha}\ell^2}{1-\alpha}. \tag{54}$$

Proof For any given q and ℓ, the left-hand side of (54) is concave in p. Hence, this is maximized when

$$\frac{\mathrm{d}}{\mathrm{d}p}\left(\ell \cdot (q-p) - d(p,q)\right) = -\ell - q^{\alpha-1} + p^{\alpha-1} = 0. \tag{55}$$

For such p, we have

$$p = (q^{\alpha-1} + \ell)^{\frac{1}{\alpha-1}} \leq \left(q^{\alpha-1} - \frac{1-\alpha}{2} q^{\alpha-1}\right)^{\frac{1}{\alpha-1}} = q\left(1 - \frac{1-\alpha}{2}\right)^{\frac{1}{\alpha-1}} \tag{56}$$

$$= q \exp\left(\frac{1}{\alpha-1} \log\left(1 + \frac{\alpha-1}{2}\right)\right) \leq q \exp(\log 2) = 2q, \tag{57}$$

where the first equality follows from (55) and the first inequality follows from the assumption of $\ell \geq -\frac{1-\alpha}{2} q^{\alpha-1}$. Further, from the intermediate value theorem and the fact that $p^{\alpha-2}$ is monotone decreasing in p, we have

$$\begin{aligned}
|\ell| &= |p^{\alpha-1} - q^{\alpha-1}| \\
&\geq \min\left\{|(\alpha-1)p^{\alpha-2}|, |(\alpha-1)q^{\alpha-2}|\right\} |p-q| \\
&= (1-\alpha) \max\{p,q\}^{\alpha-2} |p-q|,
\end{aligned}$$

where the first equality follows from (55) and the inequality follows from the fact that $p \mapsto p^{\alpha-1}$ is a convex function. This implies

$$|p-q| \leq \frac{1}{1-\alpha} \cdot \max\{p,q\}^{2-\alpha} |\ell|. \tag{58}$$

As we have $\ell \cdot (q-p) = d(p,q) + d(q,p)$ for p satisfying (55), we have

$$\ell \cdot (q-p) - d(p,q) = d(q,p) \leq \frac{1-\alpha}{2} (\min\{p,q\}^{\alpha-2})(p-q)^2 \tag{59}$$

$$\leq \frac{1}{2(1-\alpha)} \min\{p,q\}^{\alpha-2} (\max\{p,q\}^{2-\alpha}\ell)^2, \tag{60}$$

where the first inequality follows from (53) and the second inequality follows from (58). If $p \geq q$, as we have $p \leq 2q$ from (56), it holds that

$$\ell \cdot (q-p) - d(p,q) \leq \frac{1}{2(1-\alpha)} \min\{p,q\}^{\alpha-2} (\max\{p,q\}^{2-\alpha}\ell)^2 \tag{61}$$

$$\leq \frac{1}{2(1-\alpha)} q^{\alpha-2} ((2q)^{2-\alpha}\ell)^2 = \frac{2q^{2-\alpha}\ell^2}{1-\alpha}. \tag{62}$$

If $p < q$, we have

$$\ell \cdot (q-p) - d(p,q) \leq \ell \cdot (q-p) \leq \frac{1}{1-\alpha} \cdot \max\{p,q\}^{2-\alpha} \ell^2 = \frac{q^{2-\alpha}\ell^2}{1-\alpha} \tag{63}$$

where the first inequality follows from $d(p,q) \geq 0$, the second inequality follows form (58), and the equality follows from the assumption of $p < q$. As (62) holds if $p \geq q$ and (63) holds otherwise, we have (54) for all p. ■

Lemma 21 (stability for probability simplex) *Fix arbitrary $i^* \in [K]$ and $q \in \mathcal{P}(K)$. If $\ell_i \geq -\frac{1-\alpha}{4} q_i^{\alpha-1}$ for all $i \in [K]$, we then have*

$$\langle \ell, q - p \rangle - D(p,q) \leq \frac{4}{1-\alpha} \left(\sum_{i=1}^{K} q_i^{2-\alpha} \ell_i^2 \right) \tag{64}$$

for any $p \in \mathcal{P}(K)$. If $\ell_i \geq -\frac{1-\alpha}{4} q_i^{\alpha-1}$ for all $i \in [K] \setminus \{i^*\}$ and $\ell_{i^*} \leq \frac{1-\alpha}{4}(1-q_{i^*})^{\alpha-1}$, we then have

$$\langle \ell, q - p \rangle - D(p,q) \leq \frac{4}{1-\alpha} \left(\sum_{i \neq i^*} q_i^{2-\alpha} \ell_i^2 + (1-q_{i^*})^{2-\alpha} \ell_{i^*}^2 \right) \tag{65}$$

for any $p \in \mathcal{P}(K)$.

Proof From the definition of the Bregman divergence, we have

$$\begin{aligned}
&\langle \ell, q - p \rangle - D(p,q) \\
&= \frac{1}{2} \sum_{i \neq i^*} \left(2\ell_i \cdot (q_i - p_i) - d(p_i, q_i) \right) + \frac{1}{2} \left(2\ell_{i^*} \cdot (q_{i^*} - p_{i^*}) - d(p_{i^*}, q_{i^*}) - \sum_{i \neq i^*} d(p_i, q_i) \right) \\
&\leq \frac{1}{2} \sum_{i \neq i^*} \left(2\ell_i \cdot (q_i - p_i) - d(p_i, q_i) \right) \\
&\quad + \frac{1}{2} \min \left\{ 2\ell_{i^*} \cdot (q_{i^*} - p_{i^*}) - d(p_{i^*}, q_{i^*}), 2\ell_{i^*} \cdot (q_{i^*} - p_{i^*}) - \sum_{i \neq i^*} d(p_i, q_i) \right\}.
\end{aligned} \tag{66}$$

From Lemma 20, if $\ell_i \geq -\frac{1-\alpha}{4} q_i^{\alpha-1}$, we have

$$2\ell_i \cdot (q_i - p_i) - d(p_i, q_i) \leq \frac{8 q_i^{2-\alpha} \ell_i^2}{1-\alpha}. \tag{67}$$

Hence, if it holds for all $i \in [K]$ that $\ell_i \geq -\frac{1-\alpha}{4} q_i^{\alpha-1}$, we have (64). Further, we have

$$q_{i^*} - p_{i^*} = (1 - p_{i^*}) - (1 - q_{i^*}) = \sum_{i \in [K] \setminus \{i^*\}} (p_i - q_i). \tag{68}$$

As we have $(1-q_{i^*})^{\alpha-1} \leq q_i^{\alpha-1}$ for any $i \in [K] \setminus \{i^*\}$, if $\ell_{i^*} \leq \frac{1-\alpha}{4}(1-q_{i^*})^{\alpha-1}$, we then have $-\ell_{i^*} \geq -\frac{1-\alpha}{4} q_i^{\alpha-1}$ for any $i \in [K] \setminus \{i^*\}$. Hence, Lemma 20 implies

$$2\ell_{i^*} \cdot (q_{i^*} - p_{i^*}) - \sum_{i \in [K] \setminus \{i^*\}} d(p_i, q_i) = \sum_{i \in [K] \setminus \{i^*\}} \left(-2\ell_{i^*} \cdot (q_i - p_i) - d(p_i, q_i) \right) \tag{69}$$

$$\leq \frac{2}{1-\alpha} \sum_{i \in [K] \setminus \{i^*\}} (2\ell_{i^*})^2 q_i^{2-\alpha} \leq \frac{8}{1-\alpha} \ell_{i^*}^2 \left(\sum_{i \in [K] \setminus \{i^*\}} q_i \right)^{2-\alpha} = \frac{8}{1-\alpha} (1-q_{i^*})^{2-\alpha} \ell_{i^*}^2. \tag{70}$$

By combining this with (66) and (67) for $i \in [K] \setminus \{i^*\}$, we obtain (65). ■

Lemma 22 *Fix arbitrary $q \in \mathcal{P}(K)$ and let $i^* \in \arg\max_{i \in [K]} q_i$. If $|\ell_i| \leq \frac{1-\alpha}{4} \min\{q_{i^*}, (1 - q_{i^*})\}^{\alpha-1}$ holds for all $i \in [K]$, we have*

$$\langle \ell, q - p \rangle - D(p, q) \leq \frac{4}{1-\alpha} \left(\sum_{i \neq i^*} q_i^{2-\alpha} \ell_i^2 + \min\{q_{i^*}, (1 - q_{i^*})\}^{2-\alpha} \ell_{i^*}^2 \right) \tag{71}$$

for any $p \in \mathcal{P}(K)$.

Proof As we have $q_i \leq q_{i^*}$ and $q_i \leq 1 - q_{i^*}$ holds for any $i \in [K] \setminus \{i^*\}$, we have $|\ell_i| \leq \frac{1-\alpha}{4} \min\{q_{i^*}, 1 - q_{i^*}\}^{\alpha-1} \leq \frac{1-\alpha}{4} q_i^{\alpha-1}$ for all $i \in [K] \setminus \{i^*\}$. If $q_{i^*} \leq 1 - q_{i^*}$, from (64) in Lemma 21, we have

$$\langle \ell, q - p \rangle - D(p, q) \leq \frac{4}{1-\alpha} \left(\sum_{i=1}^{K} q_i^{2-\alpha} \ell_i^2 \right) \leq \frac{4}{1-\alpha} \left(\sum_{i \neq i^*} q_i^{2-\alpha} \ell_i^2 + \min\{q_{i^*}, 1 - q_{i^*}\}^{2-\alpha} \ell_{i^*}^2 \right). \tag{72}$$

If $q_{i^*} > 1 - q_{i^*}$, from (65) in Lemma 21, we have

$$\begin{aligned} \langle \ell, q - p \rangle - D(p, q) &\leq \frac{4}{1-\alpha} \left(\sum_{i \neq i^*} q_i^{2-\alpha} \ell_i^2 + (1 - q_{i^*})^{2-\alpha} \ell_{i^*}^2 \right) \\ &\leq \frac{4}{1-\alpha} \left(\sum_{i \neq i^*} q_i^{2-\alpha} \ell_i^2 + \min\{q_{i^*}, 1 - q_{i^*}\}^{2-\alpha} \ell_{i^*}^2 \right). \end{aligned} \tag{73}$$

■

Lemma 23 *Fix arbitrary $\omega > 1$. For $q, r \in \mathcal{P}(K)$, suppose that $r_i \leq \omega q_i$ holds for all i. We then have $-\psi(r) \leq -(1 + (\omega - 1)\alpha)\psi(q) \leq -\omega\psi(q)$.*

Proof As $\psi(x)$ is a convex function, we have

$$\begin{aligned} \psi(q) - \psi(r) \leq \langle \nabla\psi(q), q - r \rangle &= -\frac{1}{\alpha} \sum_{i=1}^{K} (\alpha q_i^{\alpha-1} - 1)(q_i - r_i) \\ &= -\sum_{i=1}^{K} (q_i^{\alpha-1} - 1)(q_i - r_i) = \sum_{i=1}^{K} (q_i^{\alpha-1} - 1)(r_i - q_i) \\ &\leq (\omega - 1) \sum_{i=1}^{K} (q_i^{\alpha-1} - 1) q_i = -(\omega - 1)\alpha\psi(q) \leq -(\omega - 1)\psi(q), \end{aligned} \tag{74}$$

where the second inequality follows from the assumption of $r_i \leq \omega q_i$. This implies that $-\psi(r) \leq -(1 + (\omega - 1)\alpha)\psi(q) \leq -\omega\psi(q)$. ■

Lemma 24 *Let $L, \ell \in \mathbb{R}^K$ and $\omega = \sqrt{2}$. Suppose $q, r \in \mathcal{P}(K)$ are given by*

$$q \in \mathop{\arg\min}_{p \in \mathcal{P}(K)} \left\{ \langle L, p \rangle + \beta \psi(p) + \bar{\beta} \bar{\psi}(p) \right\}, \tag{75}$$

$$r \in \mathop{\arg\min}_{p \in \mathcal{P}(K)} \left\{ \langle L + \ell, p \rangle + \beta' \psi(p) + \bar{\beta} \bar{\psi}(p) \right\} \tag{76}$$

with $0 < \beta \leq \beta'$ and $0 \leq \bar{\beta}$, and

$$\psi(p) = -\frac{1}{\alpha} \sum_{i=1}^{K} (p_i^{\alpha} - p_i), \quad \bar{\psi}(p) = -\frac{1}{\bar{\alpha}} \sum_{i=1}^{K} (p_i^{\bar{\alpha}} - p_i), \tag{77}$$

where $0 \leq \bar{\alpha} < \alpha < 1$. Denote $q_* = \min\left\{1 - \max_{i \in [K]} q_i, \max_{i \in [K]} q_i\right\}$. We also assume

$$\|\ell\|_\infty \leq \max\left\{ \frac{1 - \omega^{\alpha - 1}}{2} \beta q_*^{\alpha - 1}, \frac{1 - \omega^{\bar{\alpha} - 1}}{2} \bar{\beta} q_*^{\bar{\alpha} - 1} \right\}, \tag{78}$$

$$0 \leq \beta' - \beta \leq \max\left\{ (1 - \omega^{\alpha - 1})\beta, \frac{1 - \omega^{\bar{\alpha} - 1}}{\omega} \bar{\beta} q_*^{\bar{\alpha} - \alpha} \right\}. \tag{79}$$

We then have $r_i \leq 2 q_i$ for all $i \in [K]$.

Proof Let $i^* \in \arg\max_{i \in [K]} q_i$. We then have $q_* = \min\{q_{i^*}, 1 - q_{i^*}\}$. For any $i \in [K] \setminus \{i^*\}$, we have $q_i \leq q_{i^*}$ and $q_i = 1 - \sum_{i' \in [K] \setminus \{i\}} q_{i'} \leq 1 - q_{i^*}$, which implies $q_i \leq q_*$. If $q_{i^*} > q_*$, we have $q_{i^*} > 1 - q_{i^*}$, which means $q_{i^*} > 1/2$. Hence, we can see that it suffices to show $r_i \leq 2q_i$ for all $i \in [K]$ such that $q_i \leq q_*$. In fact, if $q_i > q_*$, such i must be i^* and $q_{i^*} > 1/2$, and therefore it is qlear that $r_i \leq 1 \leq 2q_i$. In the following, we focus on i such that $q_i \leq q_*$.

We define a monotone decreasing function $g : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ by

$$g(x) = \beta x^{\alpha - 1} + \bar{\beta} x^{\bar{\alpha} - 1}. \tag{80}$$

and define

$$s \in \mathop{\arg\min}_{p \in \mathcal{P}(K)} \left\{ \langle L + \ell, p \rangle + \beta \psi(p) + \bar{\beta} \bar{\psi}(p) \right\}. \tag{81}$$

We first show that $\omega^{-1} q_i \leq s_i \leq \omega q_i$ holds for all i such that $q_i \leq q_*$. From the first-order optimality condition, there exists $\lambda \in \mathbb{R}$ such that

$$g(s_i) = g(q_i) + \ell_i + \lambda \tag{82}$$

holds for all $i \in [K]$. If $\lambda < -\|\ell\|_\infty$, we have $g(s_i) < g(q_i)$ for all $i \in [K]$. Then, as g is monotone decreasing, we have $s_i > q_i$ for all $i \in [K]$, which contradicts to $\|s\|_1 = \|q\|_1 = 1$. Hence, we have $\lambda \geq -\|\ell\|_\infty$. Similarly, we can see $\lambda \leq \|\ell\|_\infty$. We hence have

$$g(q_i) - 2\|\ell\|_\infty \leq g(s_i) \leq g(q_i) + 2\|\ell\|_\infty \tag{83}$$

for all $i \in [K]$. This implies that $\omega^{-1} q_i \leq s_i \leq \omega q_i$ for all i such that $q_i \leq q_*$. In fact, we have

$$\begin{aligned} g(\omega q_i) &= \beta(\omega q_i)^{\alpha - 1} + \bar{\beta}(\omega q_i)^{\bar{\alpha} - 1} = \beta q_i^{\alpha - 1} + \bar{\beta} q_i^{\bar{\alpha} - 1} - \beta(1 - \omega^{\alpha - 1}) q_i^{\alpha - 1} - \bar{\beta}(1 - \omega^{\alpha - 1}) q_i^{\bar{\alpha} - 1} \\ &\leq g(q_i) - \beta(1 - \omega^{\alpha - 1}) q_*^{\alpha - 1} - \bar{\beta}(1 - \omega^{\alpha - 1}) q_*^{\bar{\alpha} - 1} \leq g(q_i) - 2\|\ell\|_\infty \leq g(s_i), \end{aligned} \tag{84}$$

$$\begin{aligned} g(\omega^{-1} q_i) &= \beta(\omega^{-1} q_i)^{\alpha - 1} + \bar{\beta}(\omega^{-1} q_i)^{\bar{\alpha} - 1} = \beta q_i^{\alpha - 1} + \bar{\beta} q_i^{\bar{\alpha} - 1} + \beta(\omega^{1 - \alpha} - 1) q_i^{\alpha - 1} + \bar{\beta}(\omega^{1 - \alpha} - 1) q_i^{\bar{\alpha} - 1} \\ &\geq g(q_i) + \beta(1 - \omega^{\alpha - 1}) q_*^{\alpha - 1} + \bar{\beta}(1 - \omega^{\alpha - 1}) q_*^{\bar{\alpha} - 1} \geq g(q_i) + 2\|\ell\|_\infty \geq g(s_i). \end{aligned} \tag{85}$$

Since g is a decreasing function, these implies that $\omega^{-1} q_i \leq s_i \leq \omega q_i$.

We next show that $r_i \leq \omega s_i$ holds for all i such that $q_i \leq q_*$. From the first-order optimality condition, there exists $\lambda \in \mathbb{R}$ such that

$$g(r_i) + (\beta' - \beta) r_i^{\alpha - 1} = g(s_i) + \lambda \tag{86}$$

holds for all $i \in [K]$. If $\lambda < 0$, we have $g(r_i) = g(s_i) + \lambda - (\beta' - \beta) r_i^{\alpha-1} < g(s_i)$, which contradicts to $\|r\|_1 = \|s\|_1 = 1$. We hence have $\lambda \geq 0$, which implies

$$g(r_i) + (\beta' - \beta) r_i^{\alpha - 1} = g(s_i) + \lambda \geq g(s_i). \tag{87}$$

For $i \in [K]$ such that $q_i \leq q_*$, we have

$$\begin{aligned}
&g(\omega s_i) + (\beta' - \beta)(\omega s_i)^{\alpha-1} = \beta \omega^{\alpha-1} s_i^{\alpha-1} + \bar{\beta} \omega^{\bar{\alpha}-1} s_i^{\bar{\alpha}-1} + (\beta' - \beta) s_i^{\alpha-1} \\
&= g(s_i) + \beta(\omega^{\alpha-1} - 1) s_i^{\alpha-1} + \bar{\beta}(\omega^{\bar{\alpha}-1} - 1) s_i^{\bar{\alpha}-1} + (\beta' - \beta) s_i^{\alpha-1} \\
&\leq g(s_i) + \beta(\omega^{\alpha-1} - 1) s_i^{\alpha-1} + \bar{\beta}(\omega^{\bar{\alpha}-1} - 1) s_i^{\bar{\alpha}-1} + \left((1 - \omega^{\alpha-1}) \beta + \frac{1 - \omega^{\bar{\alpha}-1}}{\omega} \bar{\beta} q_*^{\bar{\alpha}-\alpha} \right) s_i^{\alpha-1} \\
&= g(s_i) + \bar{\beta}(\omega^{\bar{\alpha}-1} - 1) s_i^{\bar{\alpha}-1} + \left(\frac{q_*}{s_i} \right)^{\bar{\alpha}-\alpha} \frac{1 - \omega^{\bar{\alpha}-1}}{\omega} \bar{\beta} s_i^{\bar{\alpha}-1} \\
&= g(s_i) + \bar{\beta}(\omega^{\bar{\alpha}-1} - 1) s_i^{\bar{\alpha}-1} + (1 - \omega^{\bar{\alpha}-1}) \bar{\beta} s_i^{\bar{\alpha}-1} = g(s_i) \leq g(r_i) + (\beta' - \beta) r_i^{\alpha-1}.
\end{aligned}$$

This implies that $r_i \leq \omega s_i$ since the function of $x \mapsto g(x) + (\beta' - \beta) x^{\alpha-1}$ is monotone decreasing.

We hence have $r_i \leq \omega s_i \leq \omega^2 q_i = 2 q_i$ for all $i \in [K]$ such that $q_i \leq q_*$, which completes the proof. ■

Lemma 25 *Fix arbitrary $L \in \mathbb{R}^K$ and $\omega \in (1, 2]$. Let $G = (V = [K], E)$ be an arbitrary undirected graph such that $(i, i) \in E$ holds for all $i \in V$, and let $N(i)$ denote the neighborhood of i, i.e., $N(i) = \{j \in V \mid (i, j) \in E\}$. Suppose $q, r \in \mathcal{P}(K)$ are given by*

$$q \in \operatorname*{arg\,min}_{p \in \mathcal{P}(K)} \left\{ \langle L, p \rangle + \beta \psi(p) + \bar{\beta} \bar{\psi}(p) \right\}, \tag{88}$$

$$r \in \operatorname*{arg\,min}_{p \in \mathcal{P}(K)} \left\{ \langle L + \ell, p \rangle + \beta \psi(p) + \bar{\beta} \bar{\psi}(p) \right\} \tag{89}$$

with

$$\psi(p) = -\frac{1}{\alpha} \sum_{i=1}^{K} (p_i^{\alpha} - p_i), \quad \bar{\psi}(p) = -\frac{1}{\bar{\alpha}} \sum_{i=1}^{K} (p_i^{\bar{\alpha}} - p_i), \tag{90}$$

where $0 \leq \bar{\alpha} < \alpha < 1$, $\beta \geq \frac{K}{(\omega-1)(1-\omega^{\alpha-1})}$, and $\bar{\beta} \geq 0$. Suppose ℓ is given by

$$\ell_i = \frac{\mathbf{1}[i' \in N(j)]}{\sum_{i' \in N(j)} q_{i'}} \ell'_i \tag{91}$$

for some j and $\ell' \in [0, 1]^K$. We then have $r_i \leq \omega q_i$ for all $i \in [K]$.

Proof Denote $Q_j = \sum_{i' \in N(j)} q_{i'}$. Define

$$g(x) = \beta x^{\alpha-1} + \bar{\beta} x^{\bar{\alpha}-1}. \tag{92}$$

From the first-order optimality condition, there exists $\lambda \in \mathbb{R}$ such that

$$g(r_i) = g(q_i) + \ell_i - \lambda \tag{93}$$

holds for all $i \in [K]$. As g is monotone decreasing and $\|r\|_1 = \|q\|_1 = 1$, we have $0 \leq \lambda \leq \|\ell\|_\infty$ We also have $\|\ell\|_\infty \leq Q_j$ fron the assumption of (91). Suppose $Q_j \geq \varepsilon$ with $\varepsilon := \frac{1}{\beta(1-\omega^{\alpha-1})} \leq \frac{\omega-1}{K} \leq \frac{1}{K}$. We then have $\lambda \leq 1/Q_j \leq 1/\varepsilon$. For $i \in [K]$, we have

$$g(r_i) \geq g(q_i) - \lambda \geq g(q_i) - 1/\varepsilon \geq g(\omega q_i) + \beta(1-\omega^{\alpha-1}) q_i^{\alpha-1} - 1/\varepsilon \geq g(\omega q_i), \tag{94}$$

which implies $r_i \leq \omega q_i$. Suppose $Q_j < \varepsilon$. Then, noting that $\varepsilon \leq 1/K$, we can see that $i^* \in \arg\max_{i \in [K]} q_i$ is not included in $N(j)$ as $q_{i^*} \geq 1/K$. As we have $r_i \geq q_i$ for all $i \in [K] \setminus N(j)$, we have $r_{i^*} - q_{i^*} \leq \sum_{i \in [K] \setminus N(j)} (r_i - q_i) = \sum_{i \in N(j)} (q_i - r_i) \leq Q_j < \varepsilon$. Denote $a := r_{i^*}/q_{i^*} \geq 1$. We then have $a = 1 + (r_{i^*} - q_{i^*})/q_{i^*} < 1 + K\varepsilon \leq \omega$. In addition, we have

$$g(q_i) - g(r_i) = \lambda - \ell_i \leq \lambda = g(q_{i^*}) - g(r_{i^*}) = g(q_{i^*}) - g(a q_{i^*}) \leq g(q_i) - g(a q_i), \tag{95}$$

where the last inequality follows from the fact that the function of $x \mapsto g(x) - g(ax)$ is monotone non-increasing for $a \geq 1$. This means that $g(aq_i) \leq g(r_i)$, which implies $r_i \leq aq_i$ as g is monotone decreasing. By combining this with $a \leq \omega$, we obtain $r_i \leq \omega q_i$ for all $i \in [K]$. ■

Lemma 26 *Fix arbitrary $L \in \mathbb{R}^K$ and $\omega > 1$. Suppose $q, r \in \mathcal{P}(K)$ are given by*

$$q \in \operatorname*{arg\,min}_{p \in \mathcal{P}(K)} \left\{ \langle L, p \rangle + \beta \psi(p) + \bar{\beta} \bar{\psi}(p) \right\}, \tag{96}$$

$$r \in \operatorname*{arg\,min}_{p \in \mathcal{P}(K)} \left\{ \langle L + \ell, p \rangle + \beta \psi(p) + \bar{\beta} \bar{\psi}(p) \right\} \tag{97}$$

with

$$\psi(p) = -\frac{1}{\alpha} \sum_{i=1}^K (p_i^\alpha - p_i), \quad \bar{\psi}(p) = -\frac{1}{\bar{\alpha}} \sum_{i=1}^K (p_i^{\bar{\alpha}} - p_i), \tag{98}$$

where $0 \leq \bar{\alpha} < \alpha < 1$, and $\bar{\beta} \geq 0$. Suppose $\ell \in \mathbb{R}^K_{\geq 0}$ and

$$\sum_{i=1}^K q_i \ell_i \leq \frac{1}{K} \left((1-\omega^{\alpha-1})\beta + (1-\omega^{\bar{\alpha}-1})\bar{\beta} \right) \tag{99}$$

We then have $r_i \leq \omega q_i$ for all $i \in [K]$.

Proof Define

$$g(x) = \beta x^{\alpha-1} + \bar{\beta} x^{\bar{\alpha}-1}. \tag{100}$$

From the first-order optimality condition, there exists $\lambda \in \mathbb{R}$ such that

$$g(r_i) = g(q_i) + \ell_i - \lambda \tag{101}$$

As g is monotone decreasing and $\|r\|_1 = \|q\|_1 = 1$, we have $0 \leq \lambda \leq \|\ell\|_\infty$. Let $g'(x) = (\alpha - 1)\beta x^{\alpha-2} + (\bar{\alpha} - 1)\bar{\beta} x^{\bar{\alpha}-2} < 0$ denote the derivative of $g(x)$. As g is a convex function, we have

$$g(r_i) \geq g(q_i) + g'(q_i)(r_i - q_i). \tag{102}$$

Combining (101) and (102), we obtain

$$\ell_i - \lambda \geq g'(q_i)(r_i - q_i), \tag{103}$$

which implies

$$\sum_{i=1}^K (g'(q_i))^{-1}(\ell_i - \lambda) \leq \sum_{i=1}^K (r_i - q_i) = 1 - 1 = 0. \tag{104}$$

We hence have

$$\lambda \leq \left(\sum_{i=1}^K \left(-g'(q_i)\right)^{-1} \right)^{-1} \sum_{i=1}^K \left(-g'(q_i)\right)^{-1} \ell_i. \tag{105}$$

Further, since it holds for any $x \in (0,1)$ that

$$\begin{aligned} \left((1-\alpha)\beta + (1-\bar{\alpha})\bar{\beta}\right) x^{-1} &\leq (1-\alpha)\beta x^{\alpha-2} + (1-\bar{\alpha})\bar{\beta} x^{\bar{\alpha}-2} = -g'(x) \\ &\leq \left((1-\alpha)\beta + (1-\bar{\alpha})\bar{\beta}\right) x^{-2}, \end{aligned}$$

we have

$$\sum_{i=1}^K \left(-g'(q_i)\right)^{-1} \geq \sum_{i=1}^K \left((1-\alpha)\beta + (1-\bar{\alpha})\bar{\beta}\right)^{-1} q_i^2 \geq \left((1-\alpha)\beta + (1-\bar{\alpha})\bar{\beta}\right)^{-1} \frac{1}{K} \tag{106}$$

and

$$\begin{aligned} \sum_{i=1}^K \left(-g'(q_i)\right)^{-1} \ell_i &\leq \sum_{i=1}^K \left((1-\alpha)\beta + (1-\bar{\alpha})\bar{\beta}\right)^{-1} q_i \ell_i & (107) \\ &\leq \left((1-\alpha)\beta + (1-\bar{\alpha})\bar{\beta}\right)^{-1} \frac{1}{K} (1 - \omega^{\alpha-1})\beta + (1 - \omega^{\bar{\alpha}-1})\bar{\beta}, & (108) \end{aligned}$$

where the second inequality follows from the assumption of (99). Combining (105), (106) and (107), we obtain

$$\lambda \leq (1 - \omega^{\alpha-1})\beta + (1 - \omega^{\bar{\alpha}-1})\bar{\beta}. \tag{109}$$

Therefore, we have

$$\begin{aligned} g(\omega q_i) = \beta(\omega q_i)^{\alpha-1} + \bar{\beta}(\omega q_i)^{\bar{\alpha}-1} &= g(q_i) - (1 - \omega^{\alpha-1})\beta q_i^{\alpha-1} - (1 - \omega^{\bar{\alpha}-1})\bar{\beta} q_i^{\bar{\alpha}-1} & (110) \\ &\leq g(q_i) - (1 - \omega^{\alpha-1})\beta - (1 - \omega^{\bar{\alpha}-1})\bar{\beta} \leq g(q_i) - \lambda \leq g(r_i) & (111) \end{aligned}$$

for any $i \in [K]$, where the second and the last inequalities follow from (109) and (101) with $\ell_i \geq 0$. Hence, as g is monotone decreasing, we have $r_i \leq \omega q_i$. ■

D.3. Proof of Proposition 16

Proof Fix arbitrary $i^* \in [K]$. Let $p^* \in \{0,1\}^K$ denote the indicator vector of i^*, i.e., $p^*_{i^*} = 1$ and $p^*_i = 0$ for all $i \in [K] \setminus \{i^*\}$. From the definition (23) of p_t and the assumption that $\hat{\ell}_t$ is an unbiased estimator of ℓ_t, we have

$$\begin{aligned} R_T(i^*) &= \mathbf{E}\left[\sum_{t=1}^T \ell_{t,I(t)} - \sum_{t=1}^T \ell_{t,i^*}\right] \\ &= \mathbf{E}\left[\sum_{t=1}^T \langle \ell_t, p_t - p^* \rangle\right] \\ &= \mathbf{E}\left[\sum_{t=1}^T \langle \ell_t, q_t - p^* \rangle + \sum_{t=1}^T \gamma_t \langle \ell_t, p_0 - q_t \rangle\right] \\ &\leq \mathbf{E}\left[\sum_{t=1}^T \left\langle \hat{\ell}_t, q_t - p^* \right\rangle + 2\sum_{t=1}^T \gamma_t\right]. \end{aligned} \tag{112}$$

From Lemma 19, we have

$$\sum_{t=1}^T \left\langle \hat{\ell}_t, q_t - p^* \right\rangle \leq \sum_{t=1}^T \left(\left\langle \hat{\ell}_t, q_t - q_{t+1} \right\rangle - \beta_t D(q_{t+1}, q_t) + (\beta_t - \beta_{t-1}) h_t + \bar{\beta}\bar{h}\right), \tag{113}$$

where $D(p,q)$ represents the Bregman divergence associated with $\psi(p)$, i.e., $D(p,q) = \psi(p) - \psi(q) - \langle \nabla\psi(q), p - q \rangle$, and we denote $h_t = -\psi(q_t)$ and $\bar{h} = -\bar{\psi}(q_1) \leq \frac{1}{\bar{\alpha}} K^{1-\bar{\alpha}}$. By combining these inequalities, we obtain

$$\begin{aligned} R_T(i^*) &\leq \mathbf{E}\left[\sum_{t=1}^T \left(2\gamma_t + \left\langle \hat{\ell}_t, q_t - q_{t+1} \right\rangle - \beta_t D(q_{t+1}, q_t) + (\beta_t - \beta_{t-1}) h_t\right) + \bar{\beta}\bar{h}\right] \\ &\leq O\left(\mathbf{E}\left[\sum_{t=1}^T \left(\frac{z_t}{\beta_t} + (\beta_t - \beta_{t-1}) h_{t-1}\right) + \bar{\beta}\bar{h}\right]\right) \\ &\leq O\left(\mathbf{E}\left[F(\beta_{1:T}; z_{1:T}, h_{0:T-1})\right] + \bar{\beta}\bar{h}\right), \end{aligned} \tag{114}$$

where the second inequality follows from the assumption of (25) and we define $h_0 = h_1$. From Theorem 7, if β_t is given by (10) with $\hat{h}_t = h_{t-1}$ (which is clearly an upper bound on h_{t-1}), we have

$$F(\beta_{1:T}; z_{1:T}, h_{0:T-1}) = O\left(\sqrt{h_1 \sum_{t=1}^T z_t} + \frac{z_{\max}}{\beta_1} + \beta_1 \hat{h}_1\right), \tag{115}$$

$$F(\beta_{1:T}; z_{1:T}, h_{0:T-1}) = O\left(\inf_{\varepsilon \geq \frac{1}{T}} \left\{\sqrt{\sum_{t=1}^T z_t h_t \log(\varepsilon T)} + \frac{z_{\max} h_1}{\varepsilon}\right\} + \frac{z_{\max}}{\beta_1} + \beta_1 \hat{h}_1\right). \tag{116}$$

By combining (114) and (115), we obtain $R_T = O\left(\mathbf{E}\left[\sqrt{h_1 \sum_{t=1}^T z_t} + \kappa\right]\right) \leq O\left(\sqrt{h_1 z_{\max} T} + \kappa\right)$ in adversarial regimes.

We next consider the case of adversarial regimes with self-bounding constraints. By combining (114), (116), and Jensen's inequality, we obtain

$$R_T = O\left(\sqrt{\mathbf{E}\left[\sum_{t=1}^{T} z_t h_t\right]\log(\varepsilon T) + \frac{z_{\max}h_1}{\varepsilon}} + \kappa\right) \tag{117}$$

for any $\varepsilon \geq 1/T$. Under the condition of adversarial regimes with a (Δ, C, T) self-bounding constraint, we have

$$\mathbf{E}\left[\sum_{t=1}^{T} z_t h_t\right] \leq \omega(\Delta)\,\mathbf{E}\left[\sum_{t=1}^{T}\langle \Delta, q_t\rangle\right] \leq 2\omega(\Delta)\,\mathbf{E}\left[\sum_{t=1}^{T}\langle \Delta, p_t\rangle\right] \leq 2\omega(\Delta)(R_T + 2C), \tag{118}$$

where the first inequality follows from (27), the second inequality follows from $p_{ti} = (1-\gamma_t)q_{ti} + \gamma_t p_{0i} \geq \frac{1}{2}q_{ti}$, and the last inequality follows from the assumption of self-bounding constraints given in Definition 15. We hence have

$$R_T = O\left(\sqrt{\omega(\Delta)(R_T + C)\log(\varepsilon T) + \frac{z_{\max}h_1}{\varepsilon}} + \kappa\right), \tag{119}$$

which implies

$$R_T = O\left(\omega(\Delta)\log(\varepsilon T) + \sqrt{C\omega(\Delta)\log(\varepsilon T) + \frac{z_{\max}h_1}{\varepsilon}} + \kappa\right). \tag{120}$$

We here used the fact that $X = O(\sqrt{AX} + B)$ implies $X = O(A + B)$ for any $X, A, B \geq 0$. By setting

$$\varepsilon = \frac{z_{\max}h_1}{\omega(\Delta)^2 + C\omega(\Delta)}, \tag{121}$$

we obtain

$$R_T = O\left(\omega(\Delta)\log_+\left(\frac{z_{\max}h_1 T}{\omega(\Delta)^2 + C\omega(\Delta)}\right) + \sqrt{C\omega(\Delta)\log_+\left(\frac{z_{\max}h_1 T}{\omega(\Delta)^2 + C\omega(\Delta)}\right)} + \kappa\right).$$

■

D.4. Multi-Armed Bandit: Proof of Theorem 17

From Proposition 16, it suffices to verify that conditions (25) and (27) hold.

Verifying condition (25) In the following, we denote

$$\tilde{I}(t) \in \arg\max_{i \in [K]} q_{ti}. \tag{122}$$

We then have $p_{t,\tilde{I}(t)} = q_{t,\tilde{I}(t)} \geq 1/K$, and hence $\hat{\ell}_{t,\tilde{I}(t)} \leq \frac{\ell_{t,\tilde{I}}}{p_{t,\tilde{I}(t)}} \leq K \leq \frac{(1-\alpha)\beta_1}{4} \leq \frac{(1-\alpha)\beta_t}{4}$. Hence, from Lemma 21 with $i^* = \tilde{I}(t)$, we have

$$\begin{aligned} \mathbf{E}\left[\left\langle \hat{\ell}_t, q_t - q_{t+1} \right\rangle - \beta_t D(q_{t+1}, q_t) | \mathcal{H}_{t-1}\right] &\leq \frac{4}{(1-\alpha)\beta_t} \left(\sum_{i \in [K] \setminus \{\tilde{I}(t)\}} q_{ti}^{1-\alpha} + q_{t*}^{1-\alpha} \right) \\ &\leq \frac{8}{(1-\alpha)\beta_t} \sum_{i \in [K] \setminus \{\tilde{I}(t)\}} q_{ti}^{1-\alpha} = O\left(\frac{z_t}{\beta_t}\right), \end{aligned} \tag{123}$$

which implies that the second part of (25) holds.

We next show that the first part of (25) holds. Define $q'_t \in \arg\min_{p \in \mathcal{P}(K)} \left\{ \left\langle \sum_{s=1}^{t-1} \hat{\ell}_s, p \right\rangle + \beta_{t+1}\psi(p) + \bar{\beta}\bar{\psi}(p) \right\}$. We show $q'_{ti} \leq 2q_{ti}$ and $q_{t+1,i} \leq 2q'_{ti}$ by using Lemmas 24 and 25, respectively. The condition for Lemma 24 can be verified as follows: From the definition of z_t, we have

$$z_t \leq \frac{K}{1-\alpha} q_{t*}^{1-\alpha} \tag{124}$$

and

$$h_t = -\psi(q_t) \geq \frac{q_{t*}^{\alpha}}{\alpha}(1 - 2^{\alpha-1}) \geq \frac{(1-\alpha)q_{t*}^{\alpha}}{4\alpha}. \tag{125}$$

We hence have

$$\beta_{t+1} - \beta_t = \frac{z_t}{\beta_t \hat{h}_{t+1}} = \frac{z_t}{\beta_t h_t} \leq \frac{4\alpha K q_{t*}^{1-2\alpha}}{\beta_1 (1-\alpha)^2}. \tag{126}$$

Therefore, from the definition of β_1 in (30), if $\alpha \leq 1/2$, we have

$$\frac{4\alpha K q_{t*}^{1-2\alpha}}{\beta_1(1-\alpha)^2} \leq \frac{4K}{\beta_1(1-\alpha)} \leq 1 \leq \frac{1-\alpha}{4}\beta_1 \leq (1 - \sqrt{2}^{\alpha-1})\beta_t \tag{127}$$

and hence the condition (79) in Lemma 24 holds. If $\alpha > 1/2$, as we have $\bar{\alpha} = 1 - \alpha$, from the definition of $\bar{\beta}$ in (31), we obtain

$$\frac{4\alpha K q_{t*}^{1-2\alpha}}{\beta_1(1-\alpha)^2} \leq \frac{\alpha}{8}\bar{\beta} q_{t*}^{1-2\alpha} = \frac{\alpha}{8}\bar{\beta} q_{t*}^{\bar{\alpha}-\alpha} \leq \frac{1 - \sqrt{2}^{\bar{\alpha}-1}}{\sqrt{2}} \bar{\beta} q_{t*}^{\bar{\alpha}-\alpha}, \tag{128}$$

which implies the condition (79) in Lemma 24 holds. Hence, by applying Lemma 24 with $\ell = 0$, $\beta = \beta_t$, $\beta' = \beta_{t+1}$, and $\bar{\alpha} = 1 - \alpha$, we obtain $q'_{ti} \leq 2q_{ti}$ for all $i \in [K]$. Further, as we have $\beta_{t+1} \geq \beta_1 \geq \frac{4K}{1-\alpha} \geq \frac{2K}{1-2^{\alpha-1}}$, we can apply Lemma 25 with $\omega = 2$, $E = \{(i,i) \mid i \in [K]\}$, $\ell = \hat{\ell}_t$, and $\beta = \beta_{t+1}/2$ to obtain $q_{t+1,i} \leq 2q'_{ti}$ for all $i \in [K]$. We hence have $q_{t+1,i} \leq 4q_{ti}$ for all $i \in [K]$. Therefore, from Lemma 23, we obtain $h_{t+1} = O(h_t)$, which means that the first part of (25) holds.

Verifying condition (27) For any $i^* \in [K]$, we have

$$\begin{aligned} z_t = \frac{1}{1-\alpha}\sum_{i=1}^{K} \tilde{q}_{ti}^{1-\alpha} &\leq \frac{1}{1-\alpha}\left(\sum_{i \in [K] \setminus \{\tilde{I}(t)\}} q_{ti}^{1-\alpha} + (1 - q_{t,\tilde{I}(t)})^{1-\alpha} \right) \\ &\leq \frac{2}{1-\alpha}\left(\sum_{i \in [K] \setminus \{\tilde{I}(t)\}} q_{ti}^{1-\alpha} \right) \leq \frac{2}{1-\alpha}\left(\sum_{i \in [K] \setminus \{i^*\}} q_{ti}^{1-\alpha} \right) \leq \frac{2(K-1)^{\alpha}}{1-\alpha} \end{aligned} \tag{129}$$

and

$$h_t = \frac{1}{\alpha}\left(\sum_{i=1}^{K} q_{ti}^{\alpha} - 1\right) \leq \frac{1}{\alpha}\left(\sum_{i=1}^{K} q_{ti}^{\alpha} - q_{t,i^*}^{\alpha}\right) = \frac{1}{\alpha}\sum_{i\in[K]\setminus\{i^*\}} q_{ti}^{\alpha} \leq \frac{(K-1)^{1-\alpha}}{\alpha}. \tag{130}$$

We hence have $h_1 z_{\max} \leq \frac{2(K-1)}{(1-\alpha)\alpha}$. Further, from Hölder's inequality, we have

$$\begin{aligned} z_t &\leq \frac{2}{1-\alpha}\left(\sum_{i\in[K]\setminus\{i^*\}} q_{ti}^{1-\alpha}\right) = \frac{2}{1-\alpha}\left(\sum_{i\in[K]\setminus\{i^*\}} \frac{1}{\Delta_i^{1-\alpha}}(\Delta_i q_{ti})^{1-\alpha}\right) \\ &\leq \frac{2}{1-\alpha}\left(\sum_{i\in[K]\setminus\{i^*\}} \frac{1}{\Delta_i^{\frac{1-\alpha}{\alpha}}}\right)^{\alpha}\left(\sum_{i\in[K]\setminus\{i^*\}} \Delta_i q_{ti}\right)^{1-\alpha} \end{aligned} \tag{131}$$

and

$$\begin{aligned} h_t &\leq \frac{1}{\alpha}\left(\sum_{i\in[K]\setminus\{i^*\}} q_{ti}^{\alpha}\right) = \frac{1}{\alpha}\left(\sum_{i\in[K]\setminus\{i^*\}} \frac{1}{\Delta_i^{\alpha}}(\Delta_i q_{ti})^{\alpha}\right) \\ &\leq \frac{1}{\alpha}\left(\sum_{i\in[K]\setminus\{i^*\}} \frac{1}{\Delta_i^{\frac{\alpha}{1-\alpha}}}\right)^{1-\alpha}\left(\sum_{i\in[K]\setminus\{i^*\}} \Delta_i q_{ti}\right)^{\alpha}. \end{aligned} \tag{132}$$

We hence have

$$h_t z_t \leq \frac{2}{\alpha(1-\alpha)}\left(\sum_{i\in[K]\setminus\{i^*\}} \Delta_i^{-\frac{\alpha}{1-\alpha}}\right)^{1-\alpha}\left(\sum_{i\neq[K]\setminus\{i^*\}} \Delta_i^{-\frac{1-\alpha}{\alpha}}\right)^{\alpha}\langle \Delta, q_t\rangle, \tag{133}$$

which means that (27) holds with $\omega(\Delta)$ defined by (32).

D.5. Linear Bandit: Proof of Theorem 18

From Proposition 16, it suffices to verify that conditions (25) and (27) hold.

Verifying condition (25) From (35) and (34), we have

$$|\hat{\ell}_{ti}| \leq \frac{cd}{\gamma_t} \leq \frac{(1-\alpha)q_{t*}^{\alpha-1}}{4}\beta_t. \tag{134}$$

Hence, we can apply Lemma 22 to obtain the following:

$$\begin{aligned}
&\mathbf{E}\left[\left\langle \hat{\ell}_t, q_t - q_{t+1} \right\rangle - \beta_t D(q_{t+1}, q_t) | \mathcal{H}_{t-1}\right] \leq \frac{4}{(1-\alpha)\beta_t} \mathbf{E}\left[\sum_{i=1}^{K} \hat{\ell}_{ti}^2 \tilde{q}_{ti}^{2-\alpha} | \mathcal{H}_{t-1}\right] \\
&\leq \frac{4}{(1-\alpha)\beta_t} \mathbf{E}\left[\sum_{i=1}^{K} \phi_i^\top S(p_t)^{-1} \phi_{I(t)} \phi_{I(t)}^\top S(p_t)^{-1} \phi_i \tilde{q}_{ti}^{2-\alpha} | \mathcal{H}_{t-1}\right] \\
&= \frac{4}{(1-\alpha)\beta_t} \sum_{i=1}^{K} \phi_i^\top S(q_t)^{-1} \phi_i \tilde{q}_{ti}^{2-\alpha} \leq \frac{8}{(1-\alpha)\beta_t} \sum_{i=1}^{K} \phi_i^\top S(q_t)^{-1} \phi_i \tilde{q}_{ti}^{2-\alpha} \\
&= \frac{8}{(1-\alpha)\beta_t} \mathrm{tr}\left(S(q_t)^{-1} \sum_{i=1}^{K} \phi_i \phi_i^\top \tilde{q}_{ti}^{2-\alpha} \right) \leq \frac{8}{(1-\alpha)\beta_t} \mathrm{tr}\left(S(q_t)^{-1} \sum_{i=1}^{K} \phi_i \phi_i^\top q_{ti} \right) q_{t*}^{1-\alpha} \\
&= \frac{8}{(1-\alpha)\beta_t} \mathrm{tr}(I_d) q_{t*}^{1-\alpha} = \frac{8d}{(1-\alpha)\beta_t} q_{t*}^{1-\alpha} = O\left(\frac{z_t}{\beta_t}\right), \qquad (135)
\end{aligned}$$

where $\mathrm{tr}(M)$ represents the trace of a matrix M and $I_d \in \mathbb{R}^{d\times d}$ denotes the identity matrix of size d. As it is clear from the definition of γ_t in (35) that $\gamma_t = O(z_t/\beta_t)$, we can verify that the second part of (25) holds. We next see that $h_{t+1} = O(h_t)$. From (125) and the definition of z_t, we have

$$\beta_{t+1} - \beta_t = \frac{z_t}{\beta_t \hat{h}_{t+1}} = \frac{z_t}{\beta_t h_t} \leq \frac{4\alpha d q_{t*}^{1-2\alpha}}{\beta_1 (1-\alpha)^2} \leq \frac{\bar{\beta}\alpha q_{t*}^{1-2\alpha}}{8}, \qquad (136)$$

where the first equality comes from (10), the second equality follows from the definition $\hat{h}_t$ in Algorithm 1, the first inequality follows from (125) and the definition of z_t in (35), and the last inequality follows from the condition on $\bar{\beta}$ in (35). Thus, we can apply Lemma 24 with $\ell = \hat{\ell}_t$, $\beta = \beta_t$, and $\beta' = \beta_{t+1}$ to obtain $h_{t+1} = O(h_t)$. Therefore, it has been confirmed that condition (25) is satisfied.

Verifying condition (27) From the definition of z_t in (35), and from (130), we have $h_1 z_{\max} \leq \frac{d}{\alpha(1-\alpha)} K^{1-\alpha}$. In addition, for any $i^* \in [K]$ we have

$$\begin{aligned}
z_t &\leq \frac{d}{1-\alpha} \left(1 - q_{t,\tilde{I}(t)}\right)^{1-\alpha} \leq \frac{d}{1-\alpha} (1 - q_{t,i^*})^{1-\alpha} \\
&\leq \frac{d}{(1-\alpha)\Delta_{\min}^{1-\alpha}} \left(\Delta_{\min} \sum_{i\in[K]\setminus\{i^*\}} q_{ti} \right)^{1-\alpha} \leq \frac{d}{(1-\alpha)\Delta_{\min}^{1-\alpha}} \left(\langle \Delta, q_t \rangle\right)^{1-\alpha}. \qquad (137)
\end{aligned}$$

By combining this with (132), we obtain

$$h_t z_t \leq \frac{d}{\alpha(1-\alpha)} \Delta_{\min}^{\alpha-1} \left(\sum_{i\neq i^*} \Delta_i^{-\frac{\alpha}{1-\alpha}} \right)^{1-\alpha} \langle \Delta, q_t \rangle, \qquad (138)$$

which implies that (27) holds with $\omega(\Delta)$ defined by (36).

D.6. Graph bandit

In the *graph bandit* problems, the player is given *feedback graph* $G = (V, E)$, where $V = [K]$ is the set of vertices and $E \subseteq V \times V$ is the set of edges. In this paper, we assume that the graph is undirected and that every vertex has a self-loop, i.e., $(i, j) \in E$ if $(j, i) \in E$ and $(i, i) \in E$ for all $i, j \in V$. Denote $N(i) = \{j \in [K] \mid (i, j) \in E\}$. The feedback from the environment is the values of losses for vertices adjacent to the chosen vertex, i.e., the player can observe ℓ_{ti} for all $i \in N(I(t))$, after incurring the loss of $\ell_{t,I(t)}$. Let $P_{ti} \in [0, 1]$ denote the probability that ℓ_{ti} is observed, i.e., let $P_{ti} = \sum_{j \in N(i)} p_{tj}$. Let $\zeta \geq 1$ denote the independence number of the feedback graph G.

In applying Algorithm 1 to graph bandit problems, we choose arbitrary $\alpha \in (0, 1)$ and set parameters as

$$\beta_1 \geq \frac{4K}{1-\alpha}, \quad z_t = \frac{1}{1-\alpha} \sum_{i=1}^{K} \frac{\tilde{q}_{ti}^{2-\alpha}}{P_{ti}}, \quad \gamma_t = 0, \quad \hat{\ell}_{ti} = \frac{\mathbf{1}[i \in N(I(t))]}{P_{ti}} \ell_{ti}. \tag{139}$$

We also set $\bar{\beta} \geq 0$ by (31). Then, (25) holds under the conditions of (31) and (139). In addition, we can show that $h_t = -\psi(q_t)$ and z_t in (30) satisfy $h_1 z_t \leq 2 \frac{\zeta}{\alpha(1-\alpha)} \left(\frac{K}{\zeta}\right)^{1-\alpha}$ and that (27) holds with $\omega(\Delta)$ defined by

$$\omega(\Delta) = \frac{2\zeta^{\alpha}}{\alpha(1-\alpha)} \Delta_{\min}^{\alpha-1} \left(\sum_{i \neq i^*} \Delta_i^{-\frac{\alpha}{1-\alpha}} \right)^{1-\alpha} \leq \frac{2\zeta}{\alpha(1-\alpha)\Delta_{\min}} \left(\frac{K}{\zeta}\right)^{1-\alpha}. \tag{140}$$

Hence, Proposition 16 leads to the following regret bounds:

Theorem 27 *Let $G = (V = [K], E)$ be an undirected graph, of which all vertices have self-loops, with the independence number $\zeta \geq 1$. For the graph bandit problem associated with G, Algorithm 1 with* (139) *and* (31) *achieves BOBW regret bounds in Proposition 16 with $h_1 z_{\max} = O\left(\frac{\zeta}{\alpha(1-\alpha)} \left(\frac{K}{\zeta}\right)^{1-\alpha}\right)$ and $\omega(\Delta)$ given by* (140).

Proof From Proposition 16, it suffices to verify that conditions (25) and (27) hold.

Verifying condition (25) As $\hat{\ell}_{t,\tilde{I}(t)} \leq \frac{\ell_{t,\tilde{I}(t)}}{p_{t,\tilde{I}(t)}} \leq K$ we can apply Lemma 21 with $i^* = \tilde{I}(t)$ to obtain

$$\begin{aligned} &\mathbf{E}\left[\left\langle \hat{\ell}_t, q_t - q_{t+1} \right\rangle - \beta_t D(q_{t+1}, q_t) | \mathcal{H}_{t-1}\right] \leq \frac{4}{(1-\alpha)\beta_t} \mathbf{E}\left[\sum_{i=1}^{K} \hat{\ell}_{ti}^2 \tilde{q}_{ti}^{2-\alpha} | \mathcal{H}_{t-1}\right] \\ &\leq \frac{4}{(1-\alpha)\beta_t} \mathbf{E}\left[\sum_{i=1}^{K} \frac{\mathbf{1}\{I_t \in N(i)\}}{P_{ti}^2} \tilde{q}_{ti}^{2-\alpha} | \mathcal{H}_{t-1}\right] = \frac{4}{(1-\alpha)\beta_t} \sum_{i=1}^{K} \frac{\tilde{q}_{ti}^{2-\alpha}}{P_{ti}} = O\left(\frac{z_t}{\beta_t}\right). \end{aligned} \tag{141}$$

Further, $h_{t+1} = O(h_t)$ can be shown following the approach outlined in Section D.4. Thus, it has been confirmed that condition (25) is satisfied.

Verifying condition (27) We can obtain a bound on z_t from Lemma 1 by Eldowa et al. (2023) as follows:

Lemma 28 *Let $\zeta \geq 1$ be the independence number of G. We then have*

$$\sum_{i=1}^{K} \frac{\tilde{q}_{ti}^{2-\alpha}}{P_{ti}} \leq (1 + \zeta^\alpha) \left(1 - q_{t,\tilde{I}(t)}\right)^{1-\alpha}. \tag{142}$$

Proof From the proof of Lemma 1 by Eldowa et al. (2023), there exists an independent set $S \subseteq [K] \setminus \{\tilde{I}(t)\}$ such that

$$\sum_{i \in [K] \setminus \{\tilde{I}(t)\}} \frac{q_{ti}^{2-\alpha}}{P_{ti}} \leq \sum_{i \in S} q_{ti}^{1-\alpha}. \tag{143}$$

From Hölder's inequality, we have

$$\sum_{i \in S} q_{ti}^{1-\alpha} \leq |S|^\alpha \left(\sum_{i \in S} q_{ti}\right)^{1-\alpha}. \tag{144}$$

As S is an independent set of G and $S \subseteq [K] \setminus \{\tilde{I}(t)\}$, we have

$$|S|^\alpha \left(\sum_{i \in S} q_{ti}\right)^{1-\alpha} \leq \zeta^\alpha \left(\sum_{i \in [K] \setminus \{\tilde{I}(t)\}} q_{ti}\right)^{1-\alpha} = \zeta^\alpha \left(1 - q_{t,\tilde{I}(t)}\right)^{1-\alpha}. \tag{145}$$

In addition, we have

$$\frac{q_{t*}^{2-\alpha}}{P_{t,\tilde{I}(t)}} \leq \frac{q_{t*}^{2-\alpha}}{q_{t*}} \leq q_{t*}^{1-\alpha} \leq \left(1 - q_{t,\tilde{I}(t)}\right)^{1-\alpha}. \tag{146}$$

Combining these inequalities, we obtain (142). ■

From this lemma, we have

$$z_t \leq \frac{1 + \zeta^\alpha}{1 - \alpha} \left(1 - q_{t,\tilde{I}(t)}\right)^{1-\alpha}. \tag{147}$$

From this and (130), we have $h_1 z_{\max} \leq \frac{2}{\alpha(1-\alpha)} \zeta^\alpha K^{1-\alpha}$. In addition, for any $i^* \in [K]$ we have

$$\begin{aligned} z_t &\leq \frac{1 + \zeta^\alpha}{1 - \alpha} \left(1 - q_{t,\tilde{I}(t)}\right)^{1-\alpha} \leq \frac{1 + \zeta^\alpha}{1 - \alpha} \left(1 - q_{t,i^*}\right)^{1-\alpha} \\ &\leq \frac{1 + \zeta^\alpha}{(1 - \alpha)\Delta_{\min}^{1-\alpha}} \left(\Delta_{\min} \sum_{i \in [K] \setminus \{i^*\}} q_{ti}\right)^{1-\alpha} \leq \frac{1 + \zeta^\alpha}{(1 - \alpha)\Delta_{\min}^{1-\alpha}} \left(\langle \Delta, q_t \rangle\right)^{1-\alpha}. \end{aligned} \tag{148}$$

By combining this with (132), we obtain

$$h_t z_t \leq \frac{2\zeta^\alpha}{\alpha(1 - \alpha)} \Delta_{\min}^{\alpha-1} \left(\sum_{i \neq i^*} \Delta_i^{-\frac{\alpha}{1-\alpha}}\right)^{1-\alpha} \langle \Delta, q_t \rangle, \tag{149}$$

which implies that (27) holds with $\omega(\Delta)$ defined by (140). ■

Note that we can obtain $\frac{\zeta}{\alpha(1-\alpha)}\left(\frac{K}{\zeta}\right)^{1-\alpha} = O\left(\zeta \log\left(1+\frac{K}{\zeta}\right)\right)$ by setting $\alpha = 1-\frac{1}{2(1+\log(K/\zeta))}$, which recovers the minimax regret upper bound shown by Eldowa et al. (2023).

D.7. Contextual bandit

In the *contextual bandit* problems, or the bandit problems with expert advices, each action i is associated with an *expert*, which provides an *advice* $\phi_{ti} \in \mathcal{P}(M)$ in each round t. After choosing an expert $I(t) \in [K]$, the player can observe the advices ϕ_{ti} of all experts $i \in [K]$, and pick $J(t) \in [M]$ following the distribution of $\phi_{t,I(t)}$. Then the player gets feedback of the incurred loss $\ell'_{t,J(t)}$, where $\ell'_t \in [0,1]^M$ is chosen by the environment before the player chooses $I(t)$. Let $P_t \in \mathcal{P}(M)$ denote the distribution that $J(t)$ follows given p_t and $\{\phi_{ti}\}_{i=1}^K$, i.e., $P_{tj} = \sum_{i=1}^K p_{ti}\phi_{tij}$.

Let $\alpha \geq 1/2$ and set

$$\beta_1 \geq \frac{8K}{1-\alpha}, \; \bar{\beta} \geq \frac{32M}{(1-\alpha)^2\beta_1}, \; z_t = \frac{Mq_{t*}^{1-\alpha}}{1-\alpha}, \; \gamma_t = 0, \; \hat{\ell}_{ti} = \frac{\ell'_{t,J(t)}\phi_{ti,J(t)}}{P_{t,J(t)}}. \tag{150}$$

If parameters are given by (150), then (25) holds. Further, $h_t = -\psi(q_t)$ and z_t in (150) satisfy $h_1 z_t \leq \frac{M}{\alpha(1-\alpha)}K^{1-\alpha}$ and (27) with $\omega(\Delta)$ defined as

$$\omega(\Delta) = \frac{M}{\alpha(1-\alpha)}\Delta_{\min}^{\alpha-1}\left(\sum_{i\neq i^*}\Delta_i^{-\frac{\alpha}{1-\alpha}}\right)^{1-\alpha} \leq \frac{MK^{1-\alpha}}{\alpha(1-\alpha)\Delta_{\min}}. \tag{151}$$

Hence, Proposition 16 leads to the following regret bounds:

Theorem 29 *For contextual bandit problems of M arms with K experts, Algorithm 1 with parameters given by (150) achieves BOBW regret bounds in Proposition 16 with $h_1 z_{\max} = O\left(\frac{MK^{1-\alpha}}{\alpha(1-\alpha)}\right)$ and $\omega(\Delta)$ given by (151).*

Proof From Proposition 16, it suffices to verify that conditions (25) and (27) hold.

Verifying condition (25) As $q_{t,\tilde{I}(t)} \geq 1/K$, we have

$$\hat{\ell}_{t,\tilde{I}(t)} = \frac{\ell'_{t,J(t)}\phi_{t,\tilde{I}(t),J(t)}}{P_{t,J(t)}} \leq \frac{\phi_{t,\tilde{I}(t),J(t)}}{\sum_{i=1}^K q_{ti}\phi_{ti,J(t)}} \leq \frac{\phi_{t,\tilde{I}(t),J(t)}}{q_{t,\tilde{I}(t)}\phi_{t,\tilde{I}(t),J(t)}} \leq \frac{1}{q_{t,\tilde{I}(t)}} \leq K. \tag{152}$$

Hence, we can apply Lemma 21 with $i^* = \tilde{I}(t)$ to obtain

$$\begin{aligned}
&\mathbf{E}\left[\left\langle \hat{\ell}_t, q_t - q_{t+1} \right\rangle - \beta_t D(q_{t+1}, q_t) | \mathcal{H}_{t-1}\right] \leq \frac{4}{(1-\alpha)\beta_t} \mathbf{E}\left[\sum_{i=1}^{K} \hat{\ell}_{ti}^2 \tilde{q}_{ti}^{2-\alpha} | \mathcal{H}_{t-1}\right] \\
&\leq \frac{4}{(1-\alpha)\beta_t} \mathbf{E}\left[\sum_{i=1}^{K} \frac{\phi_{ti,J(t)}^2}{P_{t,J(t)}^2} \tilde{q}_{ti}^{2-\alpha} | \mathcal{H}_{t-1}\right] \leq \frac{4}{(1-\alpha)\beta_t} \mathbf{E}\left[\sum_{i=1}^{K} \frac{\phi_{ti,J(t)}}{P_{t,J(t)}^2} \tilde{q}_{ti}^{2-\alpha} | \mathcal{H}_{t-1}\right] \\
&= \frac{4}{(1-\alpha)\beta_t} \sum_{j=1}^{M} \sum_{i=1}^{K} \frac{\phi_{tij}}{P_{tj}} \tilde{q}_{ti}^{2-\alpha} \leq \frac{4}{(1-\alpha)\beta_t} \sum_{j=1}^{M} \sum_{i=1}^{K} \frac{q_{ti}\phi_{tij}}{P_{tj}} q_{t*}^{1-\alpha} \\
&= \frac{4}{(1-\alpha)\beta_t} \sum_{j=1}^{M} \frac{P_{tj}}{P_{tj}} q_{t*}^{1-\alpha} = \frac{4M q_{t*}^{1-\alpha}}{(1-\alpha)\beta_t} = O\left(\frac{z_t}{\beta_t}\right). \qquad (153)
\end{aligned}$$

Further, $h_{t+1} = O(h_t)$ can be shown following the approach outlined in Section D.4. In fact, as we have $\hat{\ell}_{ti} \geq 0$ and $\sum_{i=1}^{K} q_{ti}\hat{\ell}_{ti} \leq 1$ and $\beta_t \geq \frac{8K}{1-\alpha}$, we can apply Lemma 26 with $\omega = 2$, $\ell = \hat{\ell}_t$. In addition, (152) and the definition of β and $\bar{\beta}$ in (150) ensure that we can apply Lemma 24 with $\omega = 2$, $\ell = 0$, and $i^* = \tilde{I}(t)$. Thus, it has been confirmed that condition (25) is satisfied.

Verifying condition (27) From the definition of z_t in (150), and from (130), we have $h_1 z_{\max} \leq \frac{d}{\alpha(1-\alpha)} K^{1-\alpha}$. In addition, for any $i^* \in [K]$ we have

$$\begin{aligned}
z_t &\leq \frac{M}{1-\alpha} \left(1 - q_{t,\tilde{I}(t)}\right)^{1-\alpha} \leq \frac{M}{1-\alpha} \left(1 - q_{t,i^*}\right)^{1-\alpha} \\
&\leq \frac{M}{(1-\alpha)\Delta_{\min}^{1-\alpha}} \left(\Delta_{\min} \sum_{i \in [K] \setminus \{i^*\}} q_{ti}\right)^{1-\alpha} \leq \frac{M}{(1-\alpha)\Delta_{\min}^{1-\alpha}} \left(\langle \Delta, q_t \rangle\right)^{1-\alpha}.
\end{aligned}$$

By combining this with (132), we obtain

$$h_t z_t \leq \frac{M}{\alpha(1-\alpha)} \Delta_{\min}^{\alpha-1} \left(\sum_{i \neq i^*} \Delta_i^{-\frac{\alpha}{1-\alpha}}\right)^{1-\alpha} \langle \Delta, q_t \rangle,$$

which implies that (27) holds with $\omega(\Delta)$ defined by (151). ■

Note that we obtain $\frac{MK^{1-\alpha}}{\alpha(1-\alpha)} = O(M \log K)$ by setting $\alpha = 1 - \frac{1}{4 \log K}$, which recovers the regret upper bound by Dann et al. (2023, Corollary 13).