A Simple and Adaptive Learning Rate for FTRL in Online Learning with Minimax Regret of $\Theta(T^{2/3})$ and its Application to Best-of-Both-Worlds

Anonymous Author(s)
Affiliation
Address
email

Abstract

Follow-the-Regularized-Leader (FTRL) is a powerful framework for various online learning problems. By designing its regularizer and learning rate to be adaptive to past observations, FTRL is known to work adaptively to various properties of an underlying environment. However, most existing adaptive learning rates are for online learning problems with a minimax regret of $\Theta(\sqrt{T})$ for the number of rounds T, and there are only a few studies on adaptive learning rates for problems with a minimax regret of $\Theta(T^{2/3})$, which include several important problems dealing with indirect feedback. To address this limitation, we establish a new adaptive learning rate framework for problems with a minimax regret of $\Theta(T^{2/3})$. Our learning rate is designed by matching the stability, penalty, and bias terms that naturally appear in regret upper bounds for problems with a minimax regret of $\Theta(T^{2/3})$. As applications of this framework, we consider two major problems dealing with indirect feedback: partial monitoring and graph bandits. We show that FTRL with our learning rate and the Tsallis entropy regularizer improves existing Best-of-Both-Worlds (BOBW) regret upper bounds, which achieve simultaneous optimality in the stochastic and adversarial regimes. The resulting learning rate is surprisingly simple compared to the existing learning rates for BOBW algorithms for problems with a minimax regret of $\Theta(T^{2/3})$.

1 Introduction

Online learning is a problem setting in which a learner interacts with an environment for T rounds with the goal of minimizing their cumulative loss. This framework includes many important online decision-making problems, such as expert problems [21, 38, 57], multi-armed bandits [6, 8, 33], linear bandits [1, 14], graph bandits [4, 42], and partial monitoring [9, 11].

For the sake of discussion in a general form, we consider the following *general online learning framework*. In this framework, a learner is initially given a finite action set $\mathcal{A} = [k] := \{1, \ldots, k\}$ and an observation set $\mathcal{O}$. At each round $t \in [T]$, the environment determines a loss function $\ell_t \colon \mathcal{A} \to [0, 1]$, and the learner selects an action $A_t \in \mathcal{A}$ based on past observations without knowing ℓ_t. The learner then suffers a loss $\ell_t(A_t)$ and observes a feedback $o_t \in \mathcal{O}$. The goal of the learner is to minimize the (pseudo-)regret Reg_T, which is defined as the expectation of the difference between the cumulative loss of the selected actions $(A_t)_{t=1}^T$ and that of an optimal action $a^* \in \mathcal{A}$ fixed in hindsight. That is, $\mathsf{Reg}_T = \mathbb{E}\big[\sum_{t=1}^T \ell_t(A_t) - \sum_{t=1}^T \ell_t(a^*)\big]$ for $a^* \in \arg\min_{a \in \mathcal{A}} \mathbb{E}\big[\sum_{t=1}^T \ell_t(a)\big]$. For example in the multi-armed bandit problem, the observation is $o_t = \ell_t(A_t)$.

Submitted to the ICML 2024 Workshop on Foundations of Reinforcement Learning and Control. Do not distribute.

Follow-the-Regularized-Leader (FTRL) is a highly powerful framework for such online learning problems. In FTRL, a probability vector q_t over $\mathcal{A}$, which is used for determining action selection probability p_t so that $A_t \sim p_t$, is obtained by solving the following convex optimization problem:

$$q_t \in \operatorname*{arg\,min}_{q \in \mathcal{P}_k} \left\{ \sum_{s=1}^{t-1} \widehat{\ell}_s(q) + \beta_t \psi(q) \right\} , \tag{1}$$

where $\mathcal{P}_k$ is the set of probability distributions over $\mathcal{A} = [k]$, $\widehat{\ell}_t \colon \mathcal{P}_k \to \mathbb{R}$ is an estimator of loss function ℓ_t, $\beta_t > 0$ is (a reciprocal of) learning rate at round t, and ψ is a convex regularizer. FTRL is known for its usefulness in various online learning problems [1, 4, 8, 27, 37]. Notably, FTRL can be viewed as a generalization of Online Gradient Descent [63] and the Hedge algorithm [21, 38, 57], and is closely related to Online Mirror Descent [36, 45].

The benefit of FTRL due to its generality is that one can design its regularizer ψ and learning rate $(\beta_t)_t$ so that it can perform adaptively to various properties of underlying loss functions. The *adaptive learning rate*, which exploits past observations, is often used to obtain such adaptivity. In order to see how it is designed, we consider the following stability–penalty decomposition, well-known in the literature [36, 45]:

$$\mathsf{Reg}_T \lesssim \underbrace{\sum_{t=1}^{T} \frac{z_t}{\beta_t}}_{\text{stability term}} + \underbrace{\beta_1 h_1 + \sum_{t=2}^{T} (\beta_t - \beta_{t-1}) h_t}_{\text{penalty term}} . \tag{2}$$

Intuitively, the *stability* term arises from the regret when the difference in FTRL outputs, x_t and x_{t+1}, is large, and the *penalty* term is due to the strength of the regularizer. For example, in the Exp3 algorithm for multi-armed bandits [8], h_t is the Shannon entropy of x_t or its upper bound, and z_t is the expectation of $(\nabla^2 \psi(x_t))^{-1}$-norm of the importance-weighted estimator $\widehat{\ell}_t$ or its upper bound.

Adaptive learning rates have been designed so that it depends on the stability or penalty. For example, the well-known AdaGrad [19, 44] and the first-order algorithm [2] depend on stability components $(z_s)_{s=1}^{t-1}$ to determine β_t. More recently, there are learning rates that depend on penalty components $(h_s)_{s=1}^{t-1}$ [25, 54] and that depend on both stability and penalty components [26, 28, 55].

However, almost all adaptive learning rates developed so far have been limited to problems with a minimax regret of $\Theta(\sqrt{T})$, and there has been limited investigation into problems with a minimax regret of $\Theta(T^{2/3})$ [25, 54]. Such online learning are primarily related to indirect feedback and includes many important problems, such as partial monitoring [9, 34], graph bandits [4], dueling bandits [51], online ranking [12], bandits with switching costs [18], and bandits with paid observations [53].

Contributions To address this limitation, we establish a new learning rate framework for online learning with a minimax regret of $\Theta(T^{2/3})$. Henceforth, we will refer to problems with a minimax regret of $\Theta(T^{2/3})$ as *hard problems* to avoid repetition, abusing the terminology of partial monitoring. For hard problems, it is common to combine FTRL with *forced exploration* [4, 17, 34, 51]. In this study, we first observe that the regret of FTRL with forced exploration rate γ_t is roughly bounded as follows:

$$\mathsf{Reg}_T \lesssim \underbrace{\sum_{t=1}^{T} \frac{z_t}{\beta_t \gamma_t}}_{\text{stability term}} + \underbrace{\beta_1 h_1 + \sum_{t=2}^{T} (\beta_t - \beta_{t-1}) h_t}_{\text{penalty term}} + \underbrace{\sum_{t=1}^{T} \gamma_t}_{\text{bias term}} . \tag{3}$$

Here, the third term, called the bias term, represents the regret incurred by forced exploration. In the aim of minimizing the RHS of (3), we will determine the exploration rate γ_t and learning rate β_t so that the above stability, penalty, and bias elements for each $t \in [T]$ are matched, where the resulting learning rate is called *Stability–Penalty–Bias matching learning rate (SPB-matching)*. This was inspired by the learning rate designed by matching the stability and penalty terms for problems with a minimax regret of $\Theta(\sqrt{T})$ [26]. Our learning rate is simultaneously adaptive to the stability component z_t and penalty component h_t, which have attracted attention in very recent years [26, 28, 55]. The SPB-matching learning rate allows us to bound the RHS of (3) from above as follows:

Theorem 1 (informal version of Theorem 6)**.** *There exists learning rate $(\beta_t)_t$ and exploration rate $(\gamma_t)_t$ for which the RHS of* (3) *is bounded by $O\big(\big(\sum_{t=1}^{T} \sqrt{z_t h_t \log(\varepsilon T)}\big)^{2/3} + \big(\sqrt{z_{\max} h_{\max}}/\varepsilon\big)^{2/3}\big)$ for any $\varepsilon \geq 1/T$, where $z_{\max} = \max_{t \in [T]} z_t$ and $h_{\max} = \max_{t \in [T]} h_t$.*

Table 1: Regret bounds for partial monitoring and graph bandits. The number of rounds is denoted as T, the number of actions as k, and the minimum suboptimality gap as Δ_{min}. The variables $c_{\mathcal{G}}$ is defined in Section 5, D is a constant dependent on the outcome distribution. The graph complexity measures δ, δ^*, satisfing $\delta^* \leq \delta$ for graphs with no self-loops, are defined in Section 6, and $\tilde{\delta}^* \leq \delta$ is the fractional weak domination number [13]. AwSB is the abbreviation of the adversarial regime with a self-bounding constraint. MS-type means that the bound in AdvSB has a form similar to the bound established by Masoudian and Seldin [43].

Setting	Ref.	Stochastic	Adversarial	AwSB
Partial monitoring (with global observability)	[30]	$D \log T$	–	–
	[37]	–	$(c_{\mathcal{G}} T)^{2/3} (\log k)^{1/3}$	–
	[54]	$\dfrac{c_{\mathcal{G}}^2 \log T \log(kT)}{\Delta_{\mathsf{min}}^2}$	$(c_{\mathcal{G}} T)^{2/3} (\log T \log(kT))^{1/3}$	✓
	[56]	$\dfrac{c_{\mathcal{G}}^2 k \log T}{\Delta_{\mathsf{min}}^2}$	$(c_{\mathcal{G}} T)^{2/3} (\log T)^{1/3}$	✓
	Ours (Cor. 9)	$\dfrac{c_{\mathcal{G}}^2 \log k \log T}{\Delta_{\mathsf{min}}^2}$	$(c_{\mathcal{G}} T)^{2/3} (\log k)^{1/3}$	✓(MS-type)
Graph bandits (with weak observability)	[4]	–	$(\delta \log k)^{1/3} T^{2/3}$	–
	[13]	–	$(\tilde{\delta}^* \log k)^{1/3} T^{2/3}$	–
	[25]	$\dfrac{\delta \log T \log(kT)}{\Delta_{\mathsf{min}}^2}$	$(\delta \log T \log(kT))^{1/3} T^{2/3}$	✓
	[15][a]	$\dfrac{\delta \log k \log T}{\Delta_{\mathsf{min}}^2}$	$(\delta \log k)^{1/3} T^{2/3}$	✓
	Ours (Cor. 11)	$\dfrac{\delta^* \log k \log T}{\Delta_{\mathsf{min}}^2}$	$(\delta^* \log k)^{1/3} T^{2/3}$	✓(MS-type)

[a]The bounds in [15] depend on δ, but their framework with the algorithm in [13] can achieve improved bounds replacing δ with $\tilde{\delta}^* \leq \delta$. The framework in [15] is a hierarchical reduction-based approach, rather than a direct FTRL method, discarding past observations as doubling-trick.

Within the general online learning framework, this theorem allows us to prove the following Best-of-Both-Worlds (BOBW) guarantee [10, 58, 61], which achieves an $O(\log T)$ regret in the stochastic regime and an $O(T^{2/3})$ regret in the adversarial regime simultaneously:

Theorem 2 (informal version of Theorem 7)**.** *Under some regularity conditions, an FTRL-based algorithm with SPB-matching achieves* $\mathsf{Reg}_T \lesssim (z_{\max} h_{\max})^{1/3} T^{2/3}$ *in the adversarial regime. In the stochastic regime, if* $\sqrt{z_t h_t} \leq \sqrt{\rho_1}(1 - q_{ta^*})$ *holds for FTRL output* $q_t \in \mathcal{P}_k$ *and* $\rho_1 > 0$ *for all* $t \in [T]$*, the same algorithm achieves* $\mathsf{Reg}_T \lesssim \rho_1 \log T / \Delta_{\mathsf{min}}^2$ *for the minimum suboptimality gap* Δ_{min}.

To assess the usefulness of the above result that holds for the general online learning framework, this study focuses on two major hard problems: partial monitoring with global observability and graph bandits with weak observability. We demonstrate that the assumptions in Theorem 2 are indeed satisfied for these problems by appropriately choosing the parameters in SPB-matching, thereby improving the existing BOBW regret upper bounds in several respects. To obtain better bounds in this analysis, we leverage the smallness of stability components z_t, which results from the forced exploration. Additionally, SPB-matching is the first unified framework to achieve a BOBW guarantee for hard online learning problems. Our learning rate is based on a surprisingly simple principle, whereas existing learning rates for graph bandits and partial monitoring are extremely complicated (see [25, Eq. (15)] and [54, Eq. (16)]). Due to its simplicity, we believe that SPB-matching will serve as a foundation for building new BOBW algorithms for a variety of hard online learning problems.

Although omitted in Theorem 2, our approach achieves a refined regret bound devised by Masoudian and Seldin [43] in the *adversarial regime with a self-bounding constraint* [61], which includes the stochastic regime, adversarial regime, and the stochastic regime with adversarial corruptions [41] as special cases. We call the refind bound *MS-type bound*, named after the author. The MS-type bound maintains an ideal form even when $C = \Theta(T)$ or $\Delta_{\mathsf{min}} = \Theta(1/\sqrt{T})$ (see [43] for details), and our bounds are the first MS-type bounds for hard problems. A comparison with existing regret bounds is summarized in Table 1.

2 Preliminaries

Notation For a natural number $n \in \mathbb{N}$, we let $[n] = \{1, \ldots, n\}$. For vector x, let x_i denote its i-th element and $\|x\|_p$ the ℓ_p-norm for $p \in [1, \infty]$. Let $\mathcal{P}_k = \{p \in [0,1]^k : \|p\|_1 = 1\}$ be the $(k-1)$-dimensional probability simplex. The vector e_i is the i-th standard basis and $\mathbf{1}$ is the all-ones vector. Let $D_\psi(x, y)$ denote the Bregman divergence from y to x induced by a differentiable convex function ψ: $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$. To simplify the notation, we sometimes write $(a_t)_{t=1}^T$ as $a_{1:T}$ and $f = O(g)$ as $f \lesssim g$. We regard function $f : \mathcal{A} = [k] \to \mathbb{R}$ as a k-dimensional vector.

General online learning framework To provide results that hold for a wide range of settings, we consider the following general online learning framework introduced in Section 1.

At each round $t \in [T] = \{1, \ldots, T\}$:
1. The environment determines a loss vector $\ell_t : \mathcal{A} \to [0, 1]$;
2. The learner selects an action $A_t \in \mathcal{A}$ based on $p_t \in \mathcal{P}_k$ without knowing ℓ_t;
3. The learner suffers a loss of $\ell_t(A_t) \in [0, 1]$ and observes a feedback $o_t \in \mathcal{O}$.

This framework includes many problems such as the expert problem, multi-armed bandits, graph bandits, partial monitoring as special cases.

Stochastic, adversarial, and their intermediate regimes Within the above general online framework, we study three different regimes for a sequence of loss functions $(\ell_t)_t$. In the stochastic regime, the sequence of loss functions is sampled from an unknown distribution $\mathcal{D}$ in an i.i.d. manner. The suboptimality gap for action $a \in \mathcal{A}$ is given by $\Delta_a = \mathbb{E}_{\ell_t \sim \mathcal{D}}[\ell_t(a) - \ell_t(a^*)]$ and the minimum suboptimality gap by $\Delta_{\mathsf{min}} = \min_{a \neq a^*} \Delta_a$. In the adversarial regime, the loss functions can be selected arbitrarily, possibly based on the past history up to round $t-1$.

We also investigate, the adversarial regime with a self-bounding constraint [61], which is an intermediate regime between the stochastic and adversarial regimes.

Definition 3. Let $\Delta \in [0,1]^k$ and $C \geq 0$. The environment is in an *adversarial regime with a* (Δ, C, T) self-bounding constraint if it holds for any algorithm that $\mathsf{Reg}_T \geq \mathbb{E}\big[\sum_{t=1}^T \Delta_{A_t} - C\big]$.

From the definition, the stochastic and adversarial regimes are special cases of this regime. Additionally, the well-known stochastic regime with adversarial corruptions [41] also falls within this regime. For the adversarial regime with a self-bounding constraint, we assume that there exists a unique optimal action a^*. This assumption is standard in the literature of BOBW algorithms (*e.g.,* [22, 39, 58]).

3 SBP-matching: Simple and adaptive learning rate for hard problems

This section designs a new learning rate framework for hard online learning problems.

3.1 Objective function that adaptive learning rate aims to minimize

In hard problems, the regret of FTRL with somewhat large exploration rate γ_t is known to be bounded in the following form [4, 25, 54]:

$$\mathsf{Reg}_T \lesssim \sum_{t=1}^T \frac{z_t}{\beta_t \gamma_t} + \sum_{t=1}^T (\beta_t - \beta_{t-1}) h_t + \sum_{t=1}^T \gamma_t \tag{4}$$

for some stability component z_t and penalty component h_t, where we set $\beta_{T+1} = \beta_T$ and $\beta_0 = 0$ for simplicity. Recall that the first term is the stability term, the second term is the penalty term, and the third term is the bias term, which arises from the forced exploration.

The goal when designing the adaptive learning rate is to minimize (4), under the constraints that $(\beta_t)_t$ is non-decreasing and β_t depends on $(z_{1:t}, h_{1:t})$ or $(z_{1:t-1}, h_{1:t})$. A naive way to choose γ_t to minimize (4) is to set $\gamma_t = \sqrt{z_t/\beta_t}$ so that the stability term and the bias term match. However, this choice does not work well in hard problems because to obtain a regret bound of (4), a lower bound of $\gamma_t \geq u_t/\beta_t$ for some $u_t > 0$ is needed. This lower bound is used to control the magnitude of the loss estimator $\widehat{\ell}_t$.[1] Therefore, we consider exploration rate of $\gamma_t = \gamma'_t + u_t/\beta_t$ for $\gamma'_t = \sqrt{z_t/\beta_t}$ and some $u_t > 0$, where γ'_t is chosen so that the stability and bias terms are matched. With these choices,

$$\begin{aligned} \text{Eq. (4)} &\le \sum_{t=1}^{T} \left(\frac{z_t}{\beta_t \gamma'_t} + (\beta_t - \beta_{t-1}) h_t + \left(\gamma'_t + \frac{u_t}{\beta_t} \right) \right) \\ &= \sum_{t=1}^{T} \left(2\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} + (\beta_t - \beta_{t-1}) h_t \right) =: F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) \,. \end{aligned} \tag{5}$$

Note that the first two terms in F, $2\sqrt{z_t/\beta_t} + u_t/\beta_t$, come from the stability and bias terms and the last term, $(\beta_t - \beta_{t-1})h_t$, is the penalty term. In the following, we investigate adaptive learning rate $(\beta_t)_{t=1}^T$ that minimizes F in (5) instead of (4).

3.2 Stability–penalty–bias matching learning rate

We consider determining $(\beta_t)_t$ by matching the stability–bias terms and the penalty term as $2\sqrt{z_t/\beta_t} + u_t/\beta_t = (\beta_t - \beta_{t-1})h_t$. Assume that when choosing β_t, we have an access to $\widehat{h}_t$ such that $h_t \le \widehat{h}_t$. Then, inspired by the above matching, we consider the following two update rules:

$$\text{(Rule 1)}\ \beta_t = \beta_{t-1} + \frac{1}{\widehat{h}_t} \left(2\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} \right), \ \text{(Rule 2)}\ \beta_t = \beta_{t-1} + \frac{1}{\widehat{h}_t} \left(2\sqrt{\frac{z_{t-1}}{\beta_{t-1}}} + \frac{u_{t-1}}{\beta_{t-1}} \right). \tag{6}$$

We call these update rules *Stability–Penalty–Bias Matching (SPB-matching)*. These are designed by following the simple principle of matching the stability, penalty, and bias elements, and Rules 1 and 2 differ only in the way indices are shifted. For the sake of convenience, we define G_1 and G_2 by

$$G_1(z_{1:T}, h_{1:T}) = \sum_{t=1}^{T} \frac{\sqrt{z_t}}{\left(\sum_{s=1}^{t} \sqrt{z_s}/h_s \right)^{1/3}} \,,\ G_2(u_{1:T}, h_{1:T}) = \sum_{t=1}^{T} \frac{u_t}{\sqrt{\sum_{s=1}^{t} u_s/h_s}} \,. \tag{7}$$

Define $z_{\max} = \max_{t \in [T]} z_t$, $u_{\max} = \max_{t \in [T]} u_t$, and $h_{\max} = \max_{t \in [T]} h_t$. Then, using SPB-matching rules in (6), we can upper-bound F in terms of G_1 and G_2 as follows:

Lemma 4. *Consider SPB-matching* (6) *and suppose that $h_t \le \widehat{h}_t$ for all $t \in [T]$. Then, Rule 1 achieves $F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) \le 3.2 G_1(z_{1:T}, \widehat{h}_{1:T}) + 2 G_2(u_{1:T}, \widehat{h}_{1:T})$ and Rule 2 achieves $F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) \le 4 G_1(z_{1:T}, \widehat{h}_{2:T+1}) + 3 G_2(u_{1:T}, \widehat{h}_{2:T+1}) + 10\sqrt{z_{\max}/\beta_1} + 5u_{\max}/\beta_1 + \beta_1 h_1$.*

The proof of Lemma 4 can be found in Appendix B.1. One can see from the proof that the effect of using $\gamma_t = \sqrt{z_t/\beta_t} + u_t/\beta_t$ instead of $\gamma_t = \sqrt{z_t/\beta_t}$ only appears in G_2, which has a less impact than G_1 when bounding F. We can further upper-bound G_1 as follows:

Lemma 5. *Let $(z_t)_{t=1}^T \subseteq \mathbb{R}_{\ge 0}$ and $(h_t)_{t=1}^T \subseteq \mathbb{R}_{>0}$ be any non-negative and positive sequences, respectively. Let $\theta_0 > \theta_1 > \cdots > \theta_J > \theta_{J+1} = 0$ and $\theta_0 \ge h_{\max}$ and define $\mathcal{T}_j = \{t \in [T] \colon \theta_{j-1} \ge h_t > \theta_j\}$ for $j \in [J]$ and $\mathcal{T}_{J+1} = \{t \in [T] \colon \theta_J \ge h_t\}$. Then, $G_1(z_{1:T}, h_{1:T}) \le \frac{3}{2} \sum_{j=1}^{J+1} \left(\sqrt{\theta_{j-1}} \sum_{t \in \mathcal{T}_j} \sqrt{z_t} \right)^{2/3}$. This implies that for all $j \in \mathbb{N}$ it holds that*

$$G_1(z_{1:T}, h_{1:T}) \le \frac{3}{2} \min\left\{ \left(\sqrt{2J} \sum_{t=1}^{T} \sqrt{z_t h_t} \right)^{\frac{2}{3}} + \left(2^{-J/2} \sqrt{z_{\max} h_{\max}} \right)^{\frac{2}{3}} T^{\frac{2}{3}} ,\ \left(\sum_{t=1}^{T} \sqrt{z_t h_{\max}} \right)^{\frac{2}{3}} \right\} .$$

Combining Lemmas 4 and 5 and the bound on G_2 in [26, Lemma 3], we obtain the following theorem.

[1] This is particularly the case when we use the Shannon entropy or Tsallis entropy regularizers, which is a weaker regularization than the log-barrier regularizer.

Algorithm 1: Best-of-both-worlds framework based on FTRL with SPB-matching learning rate and Tsallis entropy for online learning with minimax regret of $\Theta(T^{2/3})$

input: action set $\mathcal{A}$, observation set $\mathcal{O}$, exponent of Tsallis entropy α, β_1, $\bar{\beta}$
for $t = 1, 2, \ldots$ **do**
  Compute $q_t \in \mathcal{P}_k$ by (10) with a loss estimator $\widehat{y}_t$.
  Set $h_t = H_\alpha(q_t)$ and $z_t, u_t \geq 0$ defined for each problem.
  Compute action selection probability p_t from q_t by (11).
  Choose $A_t \in \mathcal{A}$ so that $\Pr[A_t = i \mid p_t] = p_{ti}$ and observe feedback $o_t \in \mathcal{O}$.
  Compute loss estimator $\widehat{\ell}_t$ based on p_t and o_t.
  Compute β_{t+1} by Rule 2 of SPB-matching in (6) with $\widehat{h}_{t+1} = h_t$.

Theorem 6. *Let $(z_t)_{t=1}^T, (u_t)_{t=1}^T \subseteq \mathbb{R}_{\geq 0}$ and $(h_t)_{t=1}^T \subseteq \mathbb{R}_{>0}$. Suppose that $\widehat{h}_t$ satisfies $h_t \leq \widehat{h}_t$ for all $t \in [T]$. Then, if β_t is given by Rule 1 in* (6)*, then for all $\varepsilon \geq 1/T$ it holds that*

$$F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) \lesssim \min\left\{\left(\sum_{t=1}^T \sqrt{z_t \widehat{h}_t \log(\varepsilon T)}\right)^{\frac{2}{3}} + \left(\sqrt{z_{\max}\widehat{h}_{\max}}\Big/\varepsilon\right)^{\frac{2}{3}}, \left(\sum_{t=1}^T \sqrt{z_t \widehat{h}_{\max}}\right)^{\frac{2}{3}}\right\}$$
$$+ \min\left\{\sqrt{\sum_{t=1}^T u_t \widehat{h}_t \log(\varepsilon T)} + \sqrt{u_{\max}\widehat{h}_{\max}/\varepsilon}\,, \sqrt{\sum_{t=1}^T u_t \widehat{h}_{\max}}\right\}. \quad (8)$$

If β_t is given by Rule 2 in (6)*, then for all $\varepsilon \geq 1/T$ it holds that*

$$F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) \lesssim \min\left\{\left(\sum_{t=1}^T \sqrt{z_t \widehat{h}_{t+1} \log(\varepsilon T)}\right)^{\frac{2}{3}} + \left(\sqrt{z_{\max}\widehat{h}_{\max}}\Big/\varepsilon\right)^{\frac{2}{3}}, \left(\sum_{t=1}^T \sqrt{z_t \widehat{h}_{\max}}\right)^{\frac{2}{3}}\right\}$$
$$+ \min\left\{\sqrt{\sum_{t=1}^T u_t \widehat{h}_{t+1} \log(\varepsilon T)} + \sqrt{u_{\max}\widehat{h}_{\max}/\varepsilon}\,, \sqrt{\sum_{t=1}^T u_t \widehat{h}_{\max}}\right\} + \sqrt{\frac{z_{\max}}{\beta_1}} + \frac{u_{\max}}{\beta_1} + \beta_1 h_1\,. \quad (9)$$

Note that these bounds are for problems with a minimax regret of $\Theta(T^{2/3})$. Roughly speaking, our bounds have an order of $\left(\sum_{t=1}^T \sqrt{z_t \widehat{h}_{t+1} \log T}\right)^{1/3}$ and differ from the existing stability-penalty-adaptive-type bounds of $\sqrt{z_t \widehat{h}_{t+1} \log T}$ for problems with a minimax regret of $\Theta(\sqrt{T})$ [26, 55]. We will see in the subsequent sections that our bounds are reasonable as they give nearly optimal regret bounds in stochastic and adversarial regimes in partial monitoring and graph bandits.

4 Best-of-both-worlds framework for hard online learning problems

Using the SPB-matching learning rate established in Section 3, this section provides a BOBW algorithm framework for hard online learning problems. We consider the following FTRL update:

$$q_t = \operatorname*{arg\,min}_{p \in \mathcal{P}_k} \left\{\sum_{s=1}^{t-1} \langle \widehat{\ell}_t, p\rangle + \beta_t(-H_\alpha(p)) + \bar{\beta}(-H_{\bar{\alpha}}(p))\right\}, \quad \alpha \in (0,1)\,,\ \bar{\alpha} = 1 - \alpha\,, \quad (10)$$

where H_α is the α-Tsallis entropy defined as $H_\alpha(p) = \frac{1}{\alpha}\sum_{i=1}^k (p_i^\alpha - p_i)$, which satisfies $H_\alpha(p) \geq 0$ and $H_\alpha(e_i) = 0$. Based on this FTRL output q_t, we set $h_t = H_\alpha(q_t)$, which satisfies $h_1 = h_{\max}$. Additionally, for q_t and some $p_0 \in \mathcal{P}_k$, we use the action selection probability $p_t \in \mathcal{P}_k$ defined by

$$p_t = (1 - \gamma_t) q_t + \gamma_t\, p_0 \quad \text{for} \quad \gamma_t = \gamma_t' + \frac{u_t}{\beta_t} = \sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t}\,, \quad (11)$$

where β_1 is chosen so that $\gamma_t \in [0, 1/2]$. Let $\kappa = \sqrt{z_{\max}/\beta_1} + u_{\max}/\beta_1 + \beta_1 h_1 + \bar{\beta}\bar{h}$ and let $\mathbb{E}_t[\,\cdot\,]$ be the expectation given all observations before round t. Then the above procedure with Rule 2 of SPB-matching in (6), summarized in Algorithm 1, achieves the following BOBW bound:

Theorem 7. *Suppose that loss function ℓ_t satisfies $\|\ell_t\|_\infty \le 1$ and the following three conditions* (i)–(iii) *are satisfied:* (i) $\mathsf{Reg}_T \le \mathbb{E}\big[\sum_{t=1}^T \langle \widehat{\ell}_t, q_t - e_{a^*}\rangle + 2\sum_{t=1}^T \gamma_t\big]$,

$$\text{(ii) } \mathbb{E}_t\Big[\langle \widehat{\ell}_t, q_t - q_{t+1}\rangle - \beta_t D_{(-H_\alpha)}(q_{t+1}, q_t)\Big] \lesssim \frac{z_t}{\beta_t \gamma_t'} , \quad \text{(iii) } h_t \lesssim h_{t-1} . \tag{12}$$

Then, in the adversarial regime, Algorithm 1 achieves

$$\mathsf{Reg}_T = O\Big((z_{\max} h_1)^{1/3} T^{2/3} + \sqrt{u_{\max} h_1 T} + \kappa\Big) . \tag{13}$$

In the adversarial regime with a (Δ, C, T)-self-bounding constraint, further suppose that

$$\sqrt{z_t h_t} \le \sqrt{\rho_1} \cdot (1 - q_{ta^*}) \quad \textit{and} \quad u_t h_t \le \rho_2 \cdot (1 - q_{ta^*}) \tag{14}$$

are satisfied for some $\rho_1, \rho_2 > 0$ for all $t \in [T]$. Then, the same algorithm achieves

$$\mathsf{Reg}_T = O\left(\frac{\rho}{\Delta_{\mathsf{min}}^2} \log\big(T\Delta_{\mathsf{min}}^2\big) + \left(\frac{C^2\rho}{\Delta_{\mathsf{min}}^2} \log\left(\frac{T\Delta_{\mathsf{min}}}{C}\right)\right)^{1/3} + \kappa'\right) \tag{15}$$

for $\rho = \max\{\rho_1, \rho_2\}$ and $\kappa' = \kappa + \big((z_{\max} h_1)^{1/3} + \sqrt{u_{\max} h_1}\big)\big(1/\Delta_{\mathsf{min}}^2 + C/\Delta_{\mathsf{min}}\big)^{2/3}$ when $T \ge 1/\Delta_{\mathsf{min}}^2 + C/\Delta_{\mathsf{min}} =: \tau$, and $\mathsf{Reg}_T = O\big((z_{\max} h_1)^{1/3} \tau^{2/3} + \sqrt{u_{\max} h_1 \tau}\big)$ when $T < \tau$.

The proof of Theorem 7 relies on Theorem 6 established in the last section and can be found in Appendix C. Note that the bound (15) becomes the bound for the stochastic regime when $C = 0$.

5 Case study (1): Partial monitoring with global observability

This section provides a new BOBW algorithm for globally observable partial monitoring games.

5.1 Problem setting and some concepts in partial monitoring

Partial monitoring games A Partial Monitoring (PM) game $\mathcal{G} = (\mathcal{L}, \Phi)$ consists of a loss matrix $\mathcal{L} \in [0,1]^{k \times d}$ and feedback matrix $\Phi \in \Sigma^{k \times d}$, where k and d are the number of actions and outcomes, respectively, and Σ is the set of feedback symbols. The game unfolds over T rounds between the learner and the environment. Before the game starts, the learner is given $\mathcal{L}$ and Φ. At each round $t \in [T]$, the environment picks an outcome $x_t \in [d]$, and then the learner chooses an action $A_t \in [k]$ without knowing x_t. Then the learner incurs an unobserved loss $\mathcal{L}_{A_t x_t}$ and only observes a feedback symbol $\sigma_t := \Phi_{A_t x_t}$. This framework can be indeed expressed as the general online learning framework in Section 2, by setting $\mathcal{O} = \Sigma$, $\ell_t(a) = \mathcal{L}_{a x_t} = e_a^\top \mathcal{L} e_{x_t}$ and $o_t = \sigma_t = \Phi_{A_t x_t}$.

We next introduce fundamental concepts for PM games. Based on the loss matrix $\mathcal{L}$, we can decompose all distributions over outcomes. For each action $a \in [k]$, the cell of action a, denoted as $\mathcal{C}_a$, is the set of probability distributions over $[d]$ for which action a is optimal. That is, $\mathcal{C}_a = \{u \in \mathcal{P}_d \colon \max_{b \in [k]} (\ell_a - \ell_b)^\top u \le 0\}$, where $\ell_a \in \mathbb{R}^d$ is the a-th row of $\mathcal{L}$.

To avoid the heavy notions and concepts of PM, we assume that the PM game has no duplicate actions $a \ne b$ such that $\ell_a = \ell_b$ and its all actions are *Pareto optimal*; that is, $\dim(\mathcal{C}_a) = d - 1$ for all $a \in [k]$. The discussion of the effect of this assumption can be found *e.g.,* in [34, 37].

Observability and loss estimation Two Pareto optimal actions a and b are *neighbors* if $\dim(\mathcal{C}_a \cap \mathcal{C}_b) = d - 2$. Then, this neighborhood relations defines *globally observable games*, for which the minimax regret of $\Theta(T^{2/3})$ is known in the litarature [9, 34]. Two neighbouring actions a and b are *globally observable* if there exists a function $w_{e(a,b)} \colon [k] \times \Sigma \to \mathbb{R}$ satisfying

$$\textstyle\sum_{c=1}^k w_{e(a,b)}(c, \Phi_{cx}) = \mathcal{L}_{ax} - \mathcal{L}_{bx} \;\text{ for all } x \in [d] , \tag{16}$$

where $e(a,b) = \{a, b\}$. A PM game is said to be globally observable if all neighboring actions are globally observable. To the end, we assume that $\mathcal{G}$ is globally observable.[2]

[2] Another representative class of PM is locally observable games, for which we can achieve a minimax regret of $\Theta(\sqrt{T})$. See [9, 36, 37] for local observability and [54, 55] for BOBW algorithms for it.

Based on the neighborhood relations, we can estimate the loss *difference* between actions, instead of estimating the loss itself. The *in-tree* is the edges of a directed tree with vertices $[k]$ and let $\mathscr{T} \subseteq [k] \times [k]$ be an in-tree over the set of actions induced by the neighborhood relations with an arbitrarily chosen root $r \in [k]$. Then, we can estimate the loss differences between Pareto optimal actions as follows. Let $G(a, \sigma)_b = \sum_{e \in \mathrm{path}_{\mathscr{T}}(b)} w_e(a, \sigma)$ for $a \in [k]$, where $\mathrm{path}_{\mathscr{T}}(b)$ is the set of edges from $b \in [k]$ to the root r on $\mathscr{T}$. Then, it is known that this G satisfies that for any Pareto optimal actions a and b, $\sum_{c=1}^{k}(G(c, \Phi_{cx})_b - G(b, \Phi_{cx})_c) = \mathcal{L}_{ax} - \mathcal{L}_{bx}$ for all $x \in [d]$ (*e.g.,* [37, Lemma 4]). From this fact, one can see that we can use $\widehat{y}_t = G(A_t, \Phi_{A_t x_t})/p_{tA_t} \in \mathbb{R}^k$ as the loss (difference) estimator, following the standard construction of the importance-weighted estimator [8, 36]. In fact, $\widehat{y}_t$ satisfies $\mathbb{E}_{A_t \sim p_t}[\widehat{y}_{ta} - \widehat{y}_{tb}] = \sum_{c=1}^{k}(G(c, \sigma_t)_a - G(c, \sigma_t)_b) = \mathcal{L}_{ax} - \mathcal{L}_{bx}$. We let $c_{\mathcal{G}} = \max\{1, k\|G\|_\infty\}$ be a game-dependent constant, where $\|G\|_\infty = \max_{a \in [k], \sigma \in \Sigma}|G(a, \sigma)|$.

5.2 Algorithm and regret upper bounds

Here, we present a new BOBW algorithm based on Algorithm 1. We use the following parameters for Algorithm 1. We use the loss (difference) estimator of $\widehat{\ell}_t = \widehat{y}_t$. We set p_0 in (11) to $p_0 = \mathbf{1}/k$. For $\tilde{I}_t \in \arg\max_{i \in [k]} q_{ti}$ and $q_{t*} = \min\{q_{t\tilde{I}_t}, 1 - q_{t\tilde{I}_t}\}$, let

$$\beta_1 \geq \frac{64 c_{\mathcal{G}}^2}{1-\alpha}, \ \bar{\beta} = \frac{32 c_{\mathcal{G}} \sqrt{k}}{(1-\alpha)^2 \sqrt{\beta_1}}, \ z_t = \frac{4 c_{\mathcal{G}}^2}{1-\alpha}\left(\sum_{i \neq \tilde{I}_t} q_{ti}^{2-\alpha} + q_{t*}^{2-\alpha}\right), \ u_t = \frac{8 c_{\mathcal{G}}}{1-\alpha} q_{t*}^{1-\alpha}. \tag{17}$$

Note that $z_{\max} = \frac{4c_{\mathcal{G}}^2}{1-\alpha}$, $u_{\max} = \frac{8c_{\mathcal{G}}}{1-\alpha}$, and $h_{\max} = h_1 = \frac{1}{\alpha} k^{1-\alpha}$. Then, we can prove the following:

Theorem 8. *In globally observable partial monitoring, for any $\alpha \in (0,1)$, Algorithm 1 with* (17) *satisfies the assumptions of Theorem 7 with $\rho_1 = \Theta\left(\frac{c_{\mathcal{G}}^2 k^{1-\alpha}}{\alpha(1-\alpha)}\right)$ and $\rho_2 = \Theta\left(\frac{c_{\mathcal{G}} k^{1-\alpha}}{\alpha(1-\alpha)}\right)$.*

The proof of Theorem 8 is given in Appendix E. Setting $\alpha = 1 - 1/(\log k)$ gives the following:

Corollary 9. *In globally observable partial monitoring with $T \geq \tau$, Algorithm 1 with* (17) *for $\alpha = 1 - 1/(\log k)$ achieves $\mathsf{Reg}_T = O\big((c_{\mathcal{G}} T)^{2/3} (\log k)^{1/3} + \kappa\big)$ in the adversarial regime and*

$$\mathsf{Reg}_T = O\left(\frac{c_{\mathcal{G}}^2 \log k}{\Delta_{\mathsf{min}}^2} \log\big(T \Delta_{\mathsf{min}}^2\big) + \left(\frac{C^2 c_{\mathcal{G}}^2 \log k}{\Delta_{\mathsf{min}}^2} \log\left(\frac{T \Delta_{\mathsf{min}}}{C}\right)\right)^{1/3} + \kappa'\right) \tag{18}$$

in the adversarial regime with a (Δ, C, T)-self-bounding constraint.

This regret upper bound is better than the bound in [54, 56] in both stochastic and adversarial regimes, notably by a factor of $\log T$ or k in the stochastic regime. The bound for the adversarial regime with a (Δ, C, T)-self-bounding constraint is the first MS-type bound in PM.

6 Case study (2): Graph bandits with weak observability

This section presents a new BOBW algorithm for weakly observable graph bandits.

6.1 Problem setting and some concepts in graph bandits

Problem setting In the graph bandit problem, the learner is given a directed feedback graph $G = (V, E)$ with $V = [k]$ and $E \subseteq V \times V$. For each $i \in V$, let $N^{\mathsf{in}}(i) = \{j \in V : (j, i) \in E\}$ and $N^{\mathsf{out}}(i) = \{j \in V : (i, j) \in E\}$ be the in-neighborhood and out-neighborhood of vertex $i \in V$, respectively. The game proceeds as the general online learning framework provided in Section 2, with action set $\mathcal{A} = V$, loss function $\ell_t \colon V \to [0, 1]$, and observation $o_t = \{\ell_t(j) \colon j \in N^{\mathsf{out}}(I_t)\}$.

Observability and domination number Similar to partial monitoring, the minimax regret of graph bandits is characterized by the properties of the feedback graph G [4]. A graph G is *observable* if it contains no self-loops, $N^{\mathsf{in}}(i) \neq \emptyset$ for all $i \in V$. A graph G is *strongly observable* if $i \in N^{\mathsf{in}}(i)$ or $V \setminus \{i\} \subseteq N^{\mathsf{in}}(i)$ for all $i \in V$. Then, a graph G is *weakly observable* if it is observable but not strongly observable.[3] The minimax regret of the weakly observable is known to be $\Theta(T^{2/3})$.

[3]Similar to the locally observable games of partial monitoring, we can achieve an $O(\sqrt{T})$ regret for graph bandits with strong observability. See *e.g.,* [4] for details.

The weak domination number characterizes precisely the minimax regret. The *weakly dominating set* $D \subseteq V$ is a set of vertices such that $\{i \in V : i \notin N^{\mathsf{out}}(i)\} \subseteq \bigcup_{i \in D} N^{\mathsf{out}}(i)$. Then, the *weak domination number* $\delta(G)$ of graph G is the size of the smallest weakly dominating set. For weakly observable G, the minimax regret of $\tilde{\Theta}(\delta^{1/3} T^{2/3})$ is known [4]. Instead, our bound depends on the *fractional domination number* $\delta^*(G)$, defined by the optimal value of the following linear program:

$$\text{minimize } \sum_{i \in V} x_i \quad \text{subject to} \quad \sum_{i \in N^{\mathsf{in}}(j)} x_i \geq 1 \ \forall j \in V \,,\ 0 \leq x_i \leq 1 \ \forall i \in V \,. \tag{19}$$

We use $(x_i^*)_{i \in V}$ to denote the optimal solution of (19) and define its normalized version $u \in \mathcal{P}_k$ by $u_i = x_i^* / \sum_{j \in V} x_j^*$. The advantage of using the fractional domination number mainly lies in its computational complexity; further details are provided in Appendix F.1.

6.2 Algorithm and regret analysis

Here, we present a new BOBW algorithm based on Algorithm 1. We use the following parameters for Algorithm 1. We use the estimator $\widehat{\ell}_t \in \mathbb{R}^k$ defined by $\widehat{\ell}_{ti} = \frac{\ell_{ti}}{P_{ti}} \mathbb{1}[i \in N^{\mathsf{out}}(I_t)]$ for $P_{ti} = \sum_{j \in N^{\mathsf{in}}(i)} p_{tj}$, which is unbiased and has been employed in the literature [4, 13]. We set p_0 in (11) to $p_0 = u$. For $\tilde{I}_t \in \arg\max_{i \in [k]} q_{ti}$ and $q_{t*} = \min\{q_{t\tilde{I}_t}, 1 - q_{t\tilde{I}_t}\}$, let

$$\beta_1 \geq \frac{64\delta^*}{1-\alpha} \,,\ \bar{\beta} = \frac{32\sqrt{k\delta^*}}{(1-\alpha)^2 \sqrt{\beta_1}} \,,\ z_t = \frac{4\delta^*}{1-\alpha} \left(\sum_{i \in V \setminus \{\tilde{I}_t\}} q_{ti}^{2-\alpha} + q_{t*}^{2-\alpha} \right) ,\ u_t = \frac{8\delta^*}{1-\alpha} q_{t*}^{1-\alpha} \,. \tag{20}$$

Note that $z_{\max} = \frac{4\delta^*}{1-\alpha}$, $u_{\max} = \frac{8\delta^*}{1-\alpha}$, and $h_{\max} = h_1 = \frac{1}{\alpha} k^{1-\alpha}$. Then, we can prove the following:

Theorem 10. *In the weakly observable graph bandit problem, for any $\alpha \in (0,1)$, Algorithm 1 with* (20) *satisfies the assumptions of Theorem 7 with $\rho_1 = \rho_2 = \Theta\left(\frac{\delta^* k^{1-\alpha}}{\alpha(1-\alpha)}\right)$.*

The proof of Theorem 10 is given in Appendix F. Setting $\alpha = 1 - 1/(\log k)$ gives the following:

Corollary 11. *In weakly observable graph bandits with $T \geq \max\{\delta^* (\log k)^2, \tau\}$, Algorithm 1 with* (20) *for $\alpha = 1 - 1/(\log k)$ achieves $\mathsf{Reg}_T = O\big({\delta^*}^{1/3} T^{2/3} (\log k)^{1/3} + \kappa\big)$ in adversarial regime and*

$$\mathsf{Reg}_T = O\left(\frac{\delta^* \log k}{\Delta_{\mathsf{min}}^2} \log\big(T\Delta_{\mathsf{min}}^2\big) + \left(\frac{C^2 \delta^* \log k}{\Delta_{\mathsf{min}}^2} \log\left(\frac{T\Delta_{\mathsf{min}}}{C} \right) \right)^{1/3} + \kappa' \right) \tag{21}$$

in the adversarial regime with a (Δ, C, T)-self-bounding constraint.

Our bound is the first BOBW FTRL-based algorithm with the $O(\log T)$ bound in the stochastic regime, improving the existing best FTRL-based algorithm in [25]. Compared to the reduction-based approach in [15], the dependences on T are the same. However, our bound unfortunately depends on the fractional domination number δ^* instead of the weak domination number δ, which can be smaller than δ^*. Roughly speaking, this comes from the use of Tsallis entropy instead of Shannon entropy employed for the existing BOBW bound [25]. The technical challenges of making our bound depend on δ instead of δ^* or the weak fractional domination number $\tilde{\delta}^*$ are further discussed in Appendix F.3. Still, we believe that our algorithm can perform better since the reduction-based algorithm discards past observations as the doubling trick. Furthermore, the bound for the adversarial regime with a (Δ, C, T)-self-bounding constraint is the first MS-type bound in weakly observable graph bandits.

7 Conclusion and future work

In this work, we investigated hard online learning problems, that is online learning with a minimax regret of $\Theta(T^{2/3})$, and established a simple and adaptive learning rate framework called stability–penalty–bias matching (SPB-matching). We showed that FTRL with this framework and the Tsallis entropy regularization improves the existing BOBW regret bounds based on FTRL for two typical hard problems, partial monitoring with global observability and graph bandits with weak observability. Interestingly, the optimal exponent of Tsallis entropy in both settings is $1 - 1/(\log k)$, suggesting the reasonableness of using Shannon entropy in existing algorithms for partial monitoring [37] and graph bandits [4]. Our learning rate is surprisingly simple compared to existing ones for hard problems [25, 54]. Hence, it is important future work to investigate whether this simplicity can be leveraged to apply SPB-matching to other hard problems, such as bandits with switching costs [18] or with paid observations [53] and dueling bandits with Borda winner [51].

References

[1] Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *The 21st Annual Conference on Learning Theory*, pages 263–274, 2008.

[2] 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.

[3] 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.

[4] Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 23–35, 2015.

[5] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, volume 7, pages 1–122, 2009.

[6] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[7] Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *29th Annual Conference on Learning Theory*, volume 49, pages 116–120, 2016.

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

[9] Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19, pages 133–154, 2011.

[10] Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 42.1–42.23, 2012.

[11] Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

[12] Sougata Chaudhuri and Ambuj Tewari. Online learning to rank with top-k feedback. *Journal of Machine Learning Research*, 18(103):1–50, 2017.

[13] Houshuang Chen, zengfeng Huang, Shuai Li, and Chihao Zhang. Understanding bandits with graph feedback. In *Advances in Neural Information Processing Systems*, volume 34, pages 24659–24669, 2021.

[14] Varsha Dani, Thomas P. Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *The 21st Annual Conference on Learning Theory*, volume 2, pages 355–366, 2008.

[15] Chris Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pages 5503–5570, 2023.

[16] Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(37):1281–1316, 2014.

[17] Ofer Dekel, Ambuj Tewari, and Raman Arora. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1747–1754, 2012.

[18] Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: $T^{2/3}$ regret. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, pages 459–467, 2014.

[19] 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.

[20] Liad Erez and Tomer Koren. Towards best-of-all-worlds online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 34, pages 28511–28521, 2021.

[21] 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.

[22] Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 176–196, 2014.

[23] Pratik Gajane and Tanguy Urvoy. Utility-based dueling bandits as a partial monitoring game. *arXiv preprint arXiv:1507.02750*, 2015.

[24] Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 2552–2583, 2021.

[25] 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, pages 28631–28643, 2022.

[26] Shinji Ito, Taira Tsuchiya, and Junya Honda. Adaptive learning rate for follow-the-regularized-leader: Competitive analysis and best-of-both-worlds. *arXiv preprint arXiv:2403.00715*, 2024.

[27] Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic MDPs with unknown transition. In *Advances in Neural Information Processing Systems*, volume 34, pages 20491–20502, 2021.

[28] 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, pages 30918–30978, 2023.

[29] Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *the 44th Annual IEEE Symposium on Foundations of Computer Science*, pages 594–605, 2003.

[30] Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *Advances in Neural Information Processing Systems*, volume 28, pages 1792–1800, 2015.

[31] Fang Kong, Yichi Zhou, and Shuai Li. Simultaneously learning stochastic and adversarial bandits with general graph feedback. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 11473–11482, 2022.

[32] 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.

[33] T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[34] Tor Lattimore and Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98, pages 529–556, 2019.

[35] Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *the 32nd Annual Conference on Learning Theory*, volume 99, pages 2111–2139, 2019.

[36] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[37] Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 2488–2515, 2020.

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

[39] Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: AdaNormalHedge. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 1286–1304, 2015.

[40] Haipeng Luo, Chen-Yu Wei, and Kai Zheng. Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems*, volume 31, pages 8235–8245, 2018.

[41] Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

[42] Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, volume 24, pages 684–692, 2011.

[43] Saeed Masoudian and Yevgeny Seldin. Improved analysis of the Tsallis-INF algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 3330–3350, 2021.

[44] H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *The 23rd Conference on Learning Theory*, pages 244–256, 2010.

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

[46] Aldo Pacchiano, Christoph Dann, and Claudio Gentile. Best of both worlds model selection. In *Advances in Neural Information Processing Systems*, volume 35, pages 1883–1895, 2022.

[47] Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss (extended abstract). In *Computational Learning Theory*, pages 208–223, 2001.

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

[49] Chloé Rouyer, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 35035–35048, 2022.

[50] Aldo Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1): 224–243, 1999.

[51] Aadirupa Saha, Tomer Koren, and Yishay Mansour. Adversarial dueling bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 9235–9244, 2021.

[52] Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1287–1295, 2014.

[53] Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 280–287, 2014.

[54] Taira Tsuchiya, Shinji Ito, and Junya Honda. Best-of-both-worlds algorithms for partial monitoring. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, pages 1484–1515, 2023.

[55] 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, 2023.

[56] Taira Tsuchiya, Shinji Ito, and Junya Honda. Exploration by optimization with hybrid regularizers: Logarithmic regret with adversarial robustness in partial monitoring. *arXiv preprint arXiv:2402.08321*, 2024.

[57] Vladimir Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383, 1990.

[58] Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 1263–1291, 2018.

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

[60] Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 467–475, 2019.

[61] 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.

[62] Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 7683–7692, 2019.

[63] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *the Twentieth International Conference on Machine Learning*, pages 928–935, 2003.

A Additional related work

Best-of-both-worlds algorithms The study of BOBW algorithms was initiated by Bubeck and Slivkins [10], who focused on multi-armed bandits. The motivation arises from the difficulty of determining in advance whether the underlying environment is stochastic or adversarial in real-world problems. Since then, BOBW algorithms have been extensively studied [7, 16, 22, 40, 46, 52], and recently, FTRL is the common approach for developing BOBW algorithms [24, 28, 60, 62]. One reason is by appropriately designing the learning rate and regularizer of FTRL, we can prove a BOBW guarantee for various problem settings. Another reason is that FTRL-based approaches not only perform well in both stochastic and adversarial regimes but also achieve favorable regret bounds in the adversarial regime with a self-bounding constraint, intermediate settings including stochastically constrained adversarial regime [58] and stochastic regime with adversarial corruptions [41]. This intermediate regime is particularly useful, considering that real-world problems often lie between purely stochastic and purely adversarial regimes.

This study is closely related to FTRL with the Tsallis entropy regularization. Tsallis entropy in online learning was introduced in [3, 5], and its significance for BOBW algorithms was established in [61]. In the multi-armed bandit problem, using the exponent of Tsallis entropy $\alpha = 1/2$ provides optimal upper bounds, up to logarithmic factors, in both stochastic and adversarial regimes [61]. However, in the graph bandits, where the dependence on k is critical or in decoupled settings, optimal upper bounds can be achieved with $\alpha \neq 1/2$ [26, 32, 48, 59]. In this work, we demonstrate that using the exponent tofo $\alpha = 1 - 1/(\log k)$ for the number of actions k results in favorable regret bounds, as shown in Corollaries 9 and 11.

Partial monitoring Partial monitoring [11, 47, 50] is a very general online decision-making framework and includes a wide range of problems such as multi-armed bandits, (utility-based) dueling bandits [23], online ranking [12], and dynamic pricing [29]. The characterization of the minimax regret in partial monitoring has been progressively understood through various studies. It is known that all partial monitoring games can be classified into trivial, easy, hard, and hopeless games, where their minimax regrets are 0, $\Theta(\sqrt{T})$, $\Theta(T^{2/3})$ and $\Omega(T)$. For comprehensive literature, refer to [9] and the improved results presented in [34, 35]. The games for which we can achieve a regret bound of $O(T^{2/3})$ correspond to globally observable games.

There is limited research on BOBW algorithms for partial monitoring with global observability [54, 56]. The existing bounds exhibit suboptimal dependencies on k and T, particularly in the stochastic regime, which comes from the use of the Shannon entropy or the log-barrier regularization. By employing Tsallis entropy, our algorithm is the first to achieve ideal dependencies on both k and T. It remains uncertain whether our upper bound in the stochastic regime is optimal with respect to variables other than T. While there is an asymptotic lower bound for the stochastic regime [30], its coefficient is expressed as a complex optimization problem. Investigating this lower bound further is important future work.

Graph bandits The study on the graph bandit problem, which is also known as online learning with feedback graphs, was initiated by [42]. This problem includes several important problems such as the expert setting, multi-armed bandits, and label-efficient prediction. For example, considering a feedback graph with only self-loops, one can see that this corresponds to the multi-armed bandit problem. One of the most seminal studies on the graph bandit problem is by Alon et al. [4], who elucidated how the structure of the feedback graph influences its minimax regret. They demonstrated that the minimax regret is characterized by the observability of the feedback graph, introducing the notions of weakly observable graphs and strongly observable graphs. Of particular relevance to this study is the minimax regret of $\tilde{O}(\delta T^{2/3})$ for weakly observable graphs, where δ is the weak domination number and $\tilde{O}(\cdot)$ ignores logarithmic factors. Recently, this upper bound was improved to $\tilde{O}(\delta^* T^{2/3})$ by replacing the weak domination number with the fractional weak domination number $\tilde{\delta}^*$ [13].

There are several BOBW algorithms for graph bandits [15, 20, 25, 31, 49]. However, only a few of these studies consider the weakly observable setting [15, 25, 31]. The existing results based on FTRL rely on the domination number rather than the weak domination number [31] or exhibit poor dependence on T [25, 31], and the best regret bound of them still exhibited a dependence on T of

$(\log T)^2$ [25]. Our algorithm is the first FTRL-based algorithm in the weakly observable setting that achieves an $O(\log T)$ stochastic bound.

B Proofs for SPB-matching learning rate (Section 3)

B.1 Proof of Lemma 4

Proof of Lemma 4. We first consider Rule 1 in (6). The learning rate β_t is lower-bounded as

$$\beta_t^{3/2} \geq \beta_t^{1/2}\left(\beta_{t-1} + \frac{2}{\widehat{h}_t}\sqrt{\frac{z_t}{\beta_t}}\right) \geq \beta_{t-1}^{3/2} + \frac{2\sqrt{z_t}}{\widehat{h}_t} \geq 2\sum_{s=1}^{t} \frac{\sqrt{z_s}}{\widehat{h}_s}\,, \tag{22}$$

where the first inequality follows from the definition of β_t in (6) and the second inequality from the fact that $(\beta_t)_t$ is non-decreasing. We also have

$$\beta_t^2 \geq \beta_t\left(\beta_{t-1} + \frac{1}{\widehat{h}_t}\frac{u_t}{\beta_t}\right) \geq \beta_{t-1}^{3/2} + \frac{u_t}{\widehat{h}_t} \geq \sum_{s=1}^{t} \frac{u_s}{\widehat{h}_s}\,. \tag{23}$$

Using the last two lower bounds on β_t, we can bound F in (5) as

$$\begin{aligned}
F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) &\leq \sum_{t=1}^{T}\left(2\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} + (\beta_t - \beta_{t-1})\widehat{h}_t\right) \\
&\leq \sum_{t=1}^{T}\left(4\sqrt{\frac{z_t}{\beta_t}} + 2\frac{u_t}{\beta_t}\right) \\
&\leq 4\sum_{t=1}^{T}\sqrt{\frac{z_t}{\left(2\sum_{s=1}^{t}\sqrt{z_s}/\widehat{h}_s\right)^{1/3}}} + 2\sum_{t=1}^{T}\frac{u_t}{\sqrt{\sum_{s=1}^{t} u_t/\widehat{h}_t}} \\
&= 3.2 G_1(z_{1:T}, \widehat{h}_{1:T}) + 2G_2(u_{1:T}, \widehat{h}_{1:T})\,,
\end{aligned} \tag{24}$$

where the secoind inequality follows from the definition of β_t in (6) and the third inequality from (22) and (23). This completes the proof of the first statement in Lemma 4.

We next consider Rule 2 in (6). In this case, we can bound F as follows:

$$\begin{aligned}
F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) &\leq 2\sqrt{\frac{z_1}{\beta_1}} + \frac{u_1}{\beta_1} + \beta_1 h_1 + \sum_{t=2}^{T}\left(2\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} + (\beta_t - \beta_{t-1})\widehat{h}_t\right) \\
&= 2\sqrt{\frac{z_1}{\beta_1}} + \frac{u_1}{\beta_1} + \beta_1 h_1 + \sum_{t=2}^{T}\left(2\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} + 2\sqrt{\frac{z_{t-1}}{\beta_{t-1}}} + \frac{u_{t-1}}{\beta_{t-1}}\right) \\
&\leq \beta_1 h_1 + \sum_{t=1}^{T}\left(4\sqrt{\frac{z_t}{\beta_t}} + 2\frac{u_t}{\beta_t}\right),
\end{aligned} \tag{25}$$

where the equality follows from (6).

We then first consider bounding $\sum_{t=1}^{T}\sqrt{z_t/\beta_t}$. We can lower-bound $\beta_t^{3/2}$ as

$$\beta_t^{3/2} \geq \beta_t^{1/2}\left(\beta_{t-1} + \frac{2}{\widehat{h}_t}\sqrt{\frac{z_{t-1}}{\beta_{t-1}}}\right) \geq \beta_{t-1}^{3/2} + \frac{2\sqrt{z_{t-1}}}{\widehat{h}_t} \geq \beta_1^{3/2} + 2\sum_{s=2}^{t}\frac{\sqrt{z_{s-1}}}{\widehat{h}_s} =: \left(\beta_t^{(1)}\right)^{3/2}, \tag{26}$$

where we define

$$\beta_t^{(1)} = \left(\beta_1^{3/2} + 2\sum_{s=2}^{t}\frac{\sqrt{z_{s-1}}}{\widehat{h}_s}\right)^{2/3} = \left(\beta_1^{3/2} + 2\sum_{s=1}^{t-1}\frac{\sqrt{z_s}}{\widehat{h}_{s+1}}\right)^{2/3} \leq \beta_t\,. \tag{27}$$

In the following, we will upper-bound $\sum_{t=1}^T \sqrt{z_t/\beta_t} \le \sum_{t=1}^T \sqrt{z_t/\beta_t^{(1)}}$. Let $c = (1+\delta)^2$ for $\delta > 0$ and and we then define $\mathcal{S} = \{t \in [T] \colon \beta_{t+1}^{(1)} \le c^2 \beta_t^{(1)}\}$ and $\mathcal{S}^\mathsf{c} = [T] \setminus \mathcal{S} = \{t \in [T] \colon \beta_{t+1}^{(1)} > c^2 \beta_t^{(1)}\}$. From these definitions, we have

$$\sum_{t \in \mathcal{S}^\mathsf{c}} \sqrt{\frac{z_t}{\beta_t^{(1)}}} \le \sum_{t \in \mathcal{S}^\mathsf{c}} \sqrt{\frac{z_{\max}}{\beta_t^{(1)}}} \le \sum_{s=0}^{\infty} \left(\frac{1}{c}\right)^s \sqrt{\frac{z_{\max}}{\beta_1}} \le \frac{1}{1-1/c} \sqrt{\frac{z_{\max}}{\beta_1}} \,. \tag{28}$$

Hence, using the last inequality, we obtain

$$\begin{aligned}
\sum_{t=1}^T \sqrt{\frac{z_t}{\beta_t}} &\le \sum_{t \in \mathcal{S}} \sqrt{\frac{z_t}{\beta_t^{(1)}}} + \sum_{t \in \mathcal{S}^\mathsf{c}} \sqrt{\frac{z_t}{\beta_t^{(1)}}} \\
&\le c \sum_{t \in \mathcal{S}} \sqrt{\frac{z_t}{\beta_{t+1}^{(1)}}} + \frac{1}{1-1/c} \sqrt{\frac{z_{\max}}{\beta_1}} \\
&\le c \sum_{t \in \mathcal{S}} \sqrt{\frac{z_t}{\left(2 \sum_{s=1}^t \sqrt{z_s}/\widehat{h}_{s+1}\right)^{2/3}}} + \frac{1}{1-1/c} \sqrt{\frac{z_{\max}}{\beta_1}} \\
&= \frac{c}{2^{1/3}} \, G_1(z_{1:T}, \widehat{h}_{2:T+1}) + \frac{c}{c-1} \sqrt{\frac{z_{\max}}{\beta_1}} \,,
\end{aligned} \tag{29}$$

where the third inequality follows from the definition of $\beta^{(1)}$ in (26).

We next bound $\sum_{t=1}^T u_t/\beta_t$. We can lower-bound β_t^2 as

$$\beta_t^2 \ge \beta_t \left(\beta_{t-1} + \frac{1}{\widehat{h}_t} \frac{u_{t-1}}{\beta_{t-1}}\right) \ge \beta_{t-1}^2 + \frac{u_{t-1}}{\widehat{h}_t} \ge \beta_1^2 + \sum_{s=2}^t \frac{u_{s-1}}{\widehat{h}_s} =: \left(\beta_t^{(2)}\right)^2 \,, \tag{30}$$

where we define

$$\beta_t^{(2)} = \sqrt{\beta_1^2 + \sum_{s=2}^t \frac{u_{s-1}}{\widehat{h}_s}} = \sqrt{\beta_1^2 + \sum_{s=1}^{t-1} \frac{u_s}{\widehat{h}_{s+1}}} \le \beta_t \,. \tag{31}$$

In the following, we will upper-bound $\sum_{t=1}^T u_t/\beta_t \le \sum_{t=1}^T u_t/\beta_t^{(2)}$. Let us define $\mathcal{T} = \left\{t \in [T] \colon \beta_{t+1}^{(2)} \le c\beta_t^{(2)}\right\}$ and $\mathcal{T}^\mathsf{c} = [T] \setminus \mathcal{T} = \left\{t \in [T] \colon \beta_{t+1}^{(2)} > c\beta_t^{(2)}\right\}$. From these definitions, we have

$$\sum_{t \in \mathcal{T}^\mathsf{c}} \frac{u_t}{\beta_t^{(2)}} \le \sum_{t \in \mathcal{T}^\mathsf{c}} \frac{u_{\max}}{\beta_t^{(2)}} \le \sum_{s=0}^{\infty} \left(\frac{1}{c}\right)^s \frac{u_{\max}}{\beta_1} \le \frac{1}{1-1/c} \frac{u_{\max}}{\beta_1} \,. \tag{32}$$

Hence, using the last inequality, we obtain

$$\begin{aligned}
\sum_{t=1}^T \frac{u_t}{\beta_t} &\le \sum_{t \in \mathcal{T}} \frac{u_t}{\beta_t^{(2)}} + \sum_{t \in \mathcal{T}^\mathsf{c}} \frac{u_t}{\beta_t^{(2)}} \\
&\le c \sum_{t \in \mathcal{T}} \frac{u_t}{\beta_{t+1}^{(2)}} + \frac{1}{1-1/c} \frac{u_{\max}}{\beta_1} \\
&\le c \sum_{t \in \mathcal{T}} \frac{u_t}{\sqrt{\sum_{s=1}^t u_s/\widehat{h}_{s+1}}} + \frac{1}{1-1/c} \frac{u_{\max}}{\beta_1} \\
&= c \, G_2(u_{1:T}, \widehat{h}_{2:T+1}) + \frac{c}{c-1} \frac{z_{\max}}{\beta_1} \,.
\end{aligned} \tag{33}$$

Finally, combining (25) with (29) and (33), we obtain

$$\begin{aligned}
F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) \le{}& 3.2c \, G_1(z_{1:T}, \widehat{h}_{2:T+1}) + 2c \, G_2(u_{1:T}, \widehat{h}_{2:T+1}) \\
&+ \frac{c}{c-1} \left(2\sqrt{\frac{z_{\max}}{\beta_1}} + \frac{u_{\max}}{\beta_1}\right) + \beta_1 h_1 \,.
\end{aligned} \tag{34}$$

Setting $c = 1.25$ completes the proof. $\square$

B.2 Proof of Lemma 5

Before proving Lemma 5, we prepare the following lemma, a variant of [45, Lemma 4.13].

Lemma 12. *Let $\mathcal{T} \subseteq [T] = \{1, \dots, T\}$ and $(x_t)_{t\in\mathcal{T}}$ be a non-negative sequence. Then,*

$$\sum_{t\in\mathcal{T}} \frac{x_t}{\left(\sum_{s\in[t]\cap\mathcal{T}} x_s\right)^{1/3}} \le \frac{3}{2}\left(\sum_{t\in\mathcal{T}} x_t\right)^{2/3}. \tag{35}$$

Proof. Let $S_t = \sum_{s\in[t]\cup\mathcal{T}} x_s$. Then,

$$\frac{x_t}{\left(\sum_{s\in[t]\cap\mathcal{T}} x_s\right)^{1/3}} = \frac{x_t}{S_t^{1/3}} = \int_{S_{t-1}}^{S_t} S_t^{-1/3}\mathrm{d}z \le \int_{S_{t-1}}^{S_t} z^{-1/3}\mathrm{d}z = \frac{3}{2}\left(S_t^{2/3} - S_{t-1}^{2/3}\right). \tag{36}$$

Summing up the last inequality over $\mathcal{T}$, we obtain

$$\sum_{t\in\mathcal{T}} \frac{x_t}{\left(\sum_{s\in[t]\cap\mathcal{T}} x_s\right)^{1/3}} = \frac{3}{2}\sum_{t\in\mathcal{T}}\left(S_t^{2/3} - S_{t-1}^{2/3}\right) \le \frac{3}{2} S_T^{2/3}, \tag{37}$$

where the last inequality follows from the telescoping argument with the assumption that $x_t \ge 0$. □

Proof of Lemma 5. We upper-bound G_1 as follows:

$$\begin{aligned} G_1(z_{1:T}, h_{1:T}) &= \sum_{t=1}^{T} \frac{\sqrt{z_t}}{\left(\sum_{s=1}^{t}\sqrt{z_s}/h_s\right)^{1/3}} = \sum_{j=1}^{J+1}\sum_{t\in\mathcal{T}_j} \frac{\sqrt{z_t}}{\left(\sum_{s=1}^{t}\sqrt{z_s}/h_s\right)^{1/3}} \\ &\le \sum_{j=1}^{J+1}\sum_{t\in\mathcal{T}_j} \frac{\sqrt{z_t}}{\left(\sum_{s\in\mathcal{T}_j\cap[t]}\sqrt{z_s}/h_s\right)^{1/3}} \le \sum_{j=1}^{J+1}\sum_{t\in\mathcal{T}_j} \frac{\sqrt{z_t}}{\left(\sum_{s\in\mathcal{T}_j\cap[t]}\sqrt{z_s}/\theta_{j-1}\right)^{1/3}} \\ &= \sum_{j=1}^{J+1}\theta_{j-1}^{1/3}\sum_{t\in\mathcal{T}_j} \frac{\sqrt{z_t}}{\left(\sum_{s\in\mathcal{T}_j\cap[t]}\sqrt{z_s}\right)^{1/3}} \le \frac{3}{2}\sum_{j=1}^{J+1}\left(\sqrt{\theta_{j-1}}\sum_{t\in\mathcal{T}_j}\sqrt{z_t}\right)^{2/3}, \end{aligned} \tag{38}$$

where the last inequality follows from Lemma 12. This completes the proof of the first statement in Lemma 5. Setting $J = 0$ and $\theta_0 = h_{\max}$ in (38) yields that

$$G_1(z_{1:T}, h_{1:T}) \le \frac{3}{2}\left(\sum_{t=1}^{T}\sqrt{z_t h_{\max}}\right)^{2/3}. \tag{39}$$

Setting $\theta_j = 2^{-j}h_{\max}$ for $j \in \{0\}\cup[J]$ in (38) also gives

$$\begin{aligned} G_1(z_{1:T}, h_{1:T}) &\le \frac{3}{2}\sum_{j=1}^{J+1}\left(\sqrt{\theta_{j-1}}\sum_{t\in\mathcal{T}_j}\sqrt{z_t}\right)^{2/3} \\ &\le \frac{3}{2}\sum_{j=1}^{J}\left(\sqrt{\frac{\theta_{j-1}}{\theta_j}}\sum_{t\in\mathcal{T}_j}\sqrt{z_t h_t}\right)^{2/3} + \frac{3}{2}\left(\sqrt{\theta_J}\sum_{t\in\mathcal{T}_J}\sqrt{z_t}\right)^{2/3} \\ &= \frac{3}{2}\sum_{j=1}^{J}\left(\sqrt{2}\sum_{t\in\mathcal{T}_j}\sqrt{z_t h_t}\right)^{2/3} + \frac{3}{2}\left(2^{-J/2}\sum_{t\in\mathcal{T}_J}\sqrt{z_t h_{\max}}\right)^{2/3} \\ &\le \frac{3}{2}\left(\sqrt{2J}\sum_{j=1}^{J}\sum_{t\in\mathcal{T}_j}\sqrt{z_t h_t}\right)^{2/3} + \frac{3}{2}\left(2^{-J/2}\sum_{t\in\mathcal{T}_J}\sqrt{z_t h_{\max}}\right)^{2/3} \end{aligned}$$

(Hölder's inequality)

$$\leq \frac{3}{2}\left(\sqrt{2J}\sum_{t=1}^{T}\sqrt{z_t h_t}\right)^{2/3} + \frac{3}{2}\left(2^{-J/2}\sqrt{z_{\max} h_{\max}}\right)^{2/3} T^{2/3}\,, \tag{40}$$

where the second inequality follows from $(x+y)^{2/3} \leq x^{2/3} + y^{2/3}$ for $x, y \geq 0$. Combining the last inequality and (39) completes the proof of the second statement in Lemma 5. □

C Proof for best-of-both-worlds analysis in general online learning framework (Theorem 7, Section 4)

This section provides the proof of Theorem 7.

Proof. From Assumption (i), the regret is bounded as

$$\mathsf{Reg}_T \leq \mathbb{E}\left[\sum_{t=1}^{T}\langle \widehat{\ell}_t, q_t - e_{a^*}\rangle + 2\sum_{t=1}^{T}\gamma_t\right]. \tag{41}$$

From the standard FTRL analysis in [36, Exercise 28.12], we obtain

$$\sum_{t=1}^{T}\langle \widehat{\ell}_t, q_t - e_{a^*}\rangle \leq \sum_{t=1}^{T}\left(\left\langle \widehat{\ell}_t, q_t - q_{t+1}\right\rangle - \beta_t D_{(-H_\alpha)}(q_{t+1}, q_t) + (\beta_t - \beta_{t-1})h_t\right) + \bar{\beta}\bar{h}\,. \tag{42}$$

Combining the last two inequalities, we obtain

$$\begin{aligned}
\mathsf{Reg}_T &\leq \mathbb{E}\left[\sum_{t=1}^{T}\left(\left\langle \widehat{\ell}_t, q_t - q_{t+1}\right\rangle - \beta_t D_{(-H_\alpha)}(q_{t+1}, q_t) + (\beta_t - \beta_{t-1})h_t + 2\gamma_t\right) + \bar{\beta}\bar{h}\right] \\
&\lesssim \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{z_t}{\beta_t\gamma'_t} + (\beta_t - \beta_{t-1})h_t + \gamma_t\right) + \bar{\beta}\bar{h}\right] && \text{(Assumption (ii) in (12))} \\
&\lesssim \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{z_t}{\beta_t\gamma'_t} + (\beta_t - \beta_{t-1})h_t + \gamma'_t + \frac{u_t}{\beta_t}\right) + \bar{\beta}\bar{h}\right] && \text{(definition of } \gamma_t \text{ in (11))} \\
&\lesssim \mathbb{E}\left[\sum_{t=1}^{T}\left(\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} + (\beta_t - \beta_{t-1})h_{t-1}\right) + \bar{\beta}\bar{h}\right] && \text{(definition of } \gamma'_t \text{ and Assumption (iii))} \\
&\lesssim \mathbb{E}[F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{0:T-1})] + \bar{\beta}\bar{h}\,,
\end{aligned} \tag{43}$$

where the last inequality follows from (5). Now, since β_t follows Rule 2 in (6) with $\widehat{h}_t = h_{t-1}$, Eq. (9) in Theorem 6 gives

$$F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{0:T-1}) \lesssim \left(\sum_{t=1}^{T}\sqrt{z_t h_1}\right)^{\frac{2}{3}} + \sqrt{\sum_{t=1}^{T} u_t h_1} + \sqrt{\frac{z_{\max}}{\beta_1}} + \frac{u_{\max}}{\beta_1} + \beta_1 h_1\,, \tag{44}$$

$$\begin{aligned}
F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{0:T-1}) \lesssim \inf_{\varepsilon \geq 1/T}\Bigg\{&\left(\sum_{t=1}^{T}\sqrt{z_t h_t \log(\varepsilon T)}\right)^{\frac{2}{3}} + \left(\frac{\sqrt{z_{\max} h_1}}{\varepsilon}\right)^{\frac{2}{3}} \\
&+ \sqrt{\sum_{t=1}^{T} u_t h_t \log(\varepsilon T)} + \sqrt{\frac{u_{\max} h_1}{\varepsilon}}\Bigg\} + \sqrt{\frac{z_{\max}}{\beta_1}} + \frac{u_{\max}}{\beta_1} + \beta_1 h_1\,.
\end{aligned} \tag{45}$$

Hence, in the adversarial regime, combining (43) and (44) gives

$$\mathsf{Reg}_T \lesssim \mathbb{E}\left[\left(\sum_{t=1}^{T}\sqrt{z_t h_1}\right)^{2/3} + \sqrt{\sum_{t=1}^{T} u_t h_1}\right] + \kappa \leq (z_{\max} h_1)^{1/3} T^{2/3} + \sqrt{u_{\max} h_1 T} + \kappa\,, \tag{46}$$

where we recall that $\kappa = \sqrt{z_{\max}/\beta_1} + u_{\max}/\beta_1 + \beta_1 h_1 + \bar{\beta}\bar{h}$. This completes the proof of (13).

We next consider the adversarial regime with a (Δ, C, T)-self-bounding constraint. For any $\varepsilon \geq 1/T$, combining (43) and (45) gives

$$\begin{aligned}
\mathsf{Reg}_T &\lesssim \mathbb{E}\left[\left(\sum_{t=1}^{T}\sqrt{z_t h_t \log(\varepsilon T)}\right)^{\frac{2}{3}} + \sqrt{\sum_{t=1}^{T} u_t h_t \log(\varepsilon T)}\right] + \left(\frac{\sqrt{z_{\max}}h_1}{\varepsilon}\right)^{\frac{2}{3}} + \sqrt{\frac{u_{\max}h_1}{\varepsilon}} + \kappa \\
&\leq \left(\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{z_t h_t}\right]\sqrt{\log(\varepsilon T)}\right)^{\frac{2}{3}} + \sqrt{\mathbb{E}\left[\sum_{t=1}^{T} u_t h_t\right]\log(\varepsilon T) + \left(\frac{\sqrt{z_{\max}}h_1}{\varepsilon}\right)^{\frac{2}{3}}} + \sqrt{\frac{u_{\max}h_1}{\varepsilon}} + \kappa\,,
\end{aligned} \tag{47}$$

where the last inequality follows from Jensen's inequality. Now, using the assumption (14) and defining $Q(a^*) = \mathbb{E}\big[\sum_{t=1}^{T}(1-q_{ta^*})\big] \in [0,T]$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{z_t h_t}\right] \leq \sqrt{\rho_1}\,\mathbb{E}\left[\sum_{t=1}^{T}(1-q_{ta^*})\right] = \sqrt{\rho_1}\,Q(a^*)\,, \tag{48}$$

$$\mathbb{E}\left[\sum_{t=1}^{T}u_t h_t\right] \leq \rho_2\,\mathbb{E}\left[\sum_{t=1}^{T}(1-q_{ta^*})\right] = \rho_2\,Q(a^*)\,. \tag{49}$$

Since we consider the adversarial regime with a (Δ, C, T)-self-bounding constraint, the regret is lower-bounded as

$$\begin{aligned}
\mathsf{Reg}_T &\geq \mathbb{E}\left[\sum_{t=1}^{T}\langle \Delta, p\rangle\right] - C \geq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\langle \Delta, q\rangle\right] - C \\
&\geq \frac{1}{2}\Delta_{\mathsf{min}}\mathbb{E}\left[\sum_{t=1}^{T}(1-q_{ta^*})\right] - C = \frac{1}{2}\Delta_{\mathsf{min}}Q(a^*) - C\,,
\end{aligned} \tag{50}$$

where the second inequality follows from $p = (1-\gamma_t)q_t + \gamma_t p_0 \geq q_t/2$. Hence, combining (47) with (48), (49) and (50), we can bound the regret for any $\lambda \in (0,1]$ as follows:

$$\begin{aligned}
&\mathsf{Reg}_T = (1+\lambda)\mathsf{Reg}_T - \lambda\mathsf{Reg}_T \\
&\lesssim (1+\lambda)\Big(\sqrt{\rho_1}Q(a^*)\sqrt{\log(\varepsilon T)}\Big)^{2/3} - \frac{\lambda}{4}\Delta_{\mathsf{min}}Q(a^*) + (1+\lambda)\sqrt{\rho_2 Q(a^*)\log(\varepsilon T)} - \frac{\lambda}{4}\Delta_{\mathsf{min}}Q(a^*) \\
&\quad + (1+\lambda)\left(\left(\frac{\sqrt{z_{\max}}h_1}{\varepsilon}\right)^{2/3} + \sqrt{\frac{u_{\max}h_1}{\varepsilon}} + \kappa\right) + \lambda C \\
&\lesssim \frac{(1+\lambda)^3}{\lambda^2}\frac{\rho_1\log(\varepsilon T)}{\Delta_{\mathsf{min}}^2} + \frac{(1+\lambda)^2}{\lambda}\frac{\rho_2\log(\varepsilon T)}{\Delta_{\mathsf{min}}} + \left(\frac{\sqrt{z_{\max}}h_1}{\varepsilon}\right)^{2/3} + \sqrt{\frac{u_{\max}h_1}{\varepsilon}} + \kappa + \lambda C \\
&\lesssim \frac{\rho_1\log(\varepsilon T)}{\Delta_{\mathsf{min}}^2} + \frac{\rho_2\log(\varepsilon T)}{\Delta_{\mathsf{min}}} + \frac{1}{\lambda^2}\left(\frac{\rho_1\log(\varepsilon T)}{\Delta_{\mathsf{min}}^2} + \frac{\rho_2\log(\varepsilon T)}{\Delta_{\mathsf{min}}}\right) + \left(\frac{\sqrt{z_{\max}}h_1}{\varepsilon}\right)^{2/3} + \sqrt{\frac{u_{\max}h_1}{\varepsilon}} + \kappa + \lambda C \\
&\lesssim \frac{\rho\log(\varepsilon T)}{\Delta_{\mathsf{min}}^2} + \frac{1}{\lambda^2}\frac{\rho\log(\varepsilon T)}{\Delta_{\mathsf{min}}^2} + \left(\frac{\sqrt{z_{\max}}h_1}{\varepsilon}\right)^{2/3} + \sqrt{\frac{u_{\max}h_1}{\varepsilon}} + \kappa + \lambda C\,,
\end{aligned} \tag{51}$$

where in the first inequality we used (47) with (48), (49), (50), and Jensen's inequality, in the second inequality we used $ax^2 - bx^3 \leq 4a^3/(27b^2)$ for $a \geq 0, b > 0$ and $x \geq 0$ and $ax - bx^2 \leq a^2/(4b)$ for $a \geq 0, b > 0$ and $x \geq 0$ and in the third inequality we used $\lambda \in (0,1]$. Setting $\lambda = \Theta\big((\rho\log(\varepsilon T)/C)^{1/3}\big)$ in the last inequality, we obtain

$$\mathsf{Reg}_T \lesssim \frac{\rho\log(\varepsilon T)}{\Delta_{\mathsf{min}}^2} + \left(\frac{C^2\rho\log(\varepsilon T)}{\Delta_{\mathsf{min}}^2}\right)^{1/3} + \left(\frac{\sqrt{z_{\max}}h_1}{\varepsilon}\right)^{2/3} + \sqrt{\frac{u_{\max}h_1}{\varepsilon}} + \kappa\,.$$

Finally, when $T \geq \tau = 1/\Delta_{\mathsf{min}}^2 + C/\Delta_{\mathsf{min}}$, setting

$$\varepsilon = \frac{1}{\rho^2/\Delta_{\mathsf{min}}^2 + C\rho/\Delta_{\mathsf{min}}} \geq \frac{1}{T} \tag{52}$$

yields that

$$\begin{aligned}
\mathsf{Reg}_T &\lesssim \frac{\rho}{\Delta_{\min}^2}\log_+\left(\frac{T}{1/\Delta_{\min}^2 + C/\Delta_{\min}}\right) + \left(\frac{C^2\rho}{\Delta_{\min}^2}\log_+\left(\frac{T}{1/\Delta_{\min}^2 + C/\Delta_{\min}}\right)\right)^{1/3} \\
&\quad + (z_{\max}h_1)^{1/3}\left(\frac{1}{\Delta_{\min}^2} + \frac{C}{\Delta_{\min}}\right)^{2/3} + \sqrt{u_{\max}h_1}\sqrt{\frac{1}{\Delta_{\min}^2} + \frac{C}{\Delta_{\min}}} + \kappa \\
&\lesssim \frac{\rho}{\Delta_{\min}^2}\log_+\big(T\Delta_{\min}^2\big) + \left(\frac{C^2\rho}{\Delta_{\min}^2}\log_+\left(\frac{T\Delta_{\min}}{C}\right)\right)^{1/3} \\
&\quad + \left((z_{\max}h_1)^{1/3} + \sqrt{u_{\max}h_1}\right)\left(\frac{1}{\Delta_{\min}^2} + \frac{C}{\Delta_{\min}}\right)^{2/3} + \kappa\,, \qquad (53)
\end{aligned}$$

which completes the proof. □

D Auxiliary lemmas

This section provides auxiliary lemmas useful for proving the BOBW gurantee.

Lemma 13. *Let $\alpha \in (0,1)$ and $i^* \in [k]$. Then, the α-Tsallis entropy H_α is bounded from above as*

$$H_\alpha(q) = \frac{1}{\alpha}\sum_{i=1}^k (q_i^\alpha - q_i) \leq \frac{1}{\alpha}(k-1)^\alpha(1-q_{i^*})^\alpha \qquad (54)$$

for any $q \in \mathcal{P}_k$.

Proof. From Jensen's inequality and the fact that $x \mapsto x^\alpha$ is concave for $\alpha \in (0,1)$,

$$\begin{aligned}
\sum_{i=1}^k (q_i^\alpha - q_i) &\leq \sum_{i \neq i^*} q_i^\alpha = (k-1)\sum_{i\neq i^*}\frac{1}{k-1}q_i^\alpha \leq (k-1)\left(\frac{1}{k-1}\sum_{i\neq i^*} q_i\right)^\alpha \\
&= (k-1)^{1-\alpha}\left(\sum_{i\neq i^*} q_i\right)^\alpha = (k-1)^{1-\alpha}(1-q_{i^*})^\alpha\,, \qquad (55)
\end{aligned}$$

which completes the proof. □

Lemma 14 ([26, Lemma 10])**.** *Let $q \in \mathcal{P}_k$ and $\tilde{I} \in \arg\max_{i\in[k]} q_i$. For $\ell \in \mathbb{R}^k$, if $|\ell_i| \leq \frac{1-\alpha}{4}\frac{1}{\min\{q_{\tilde{I}}, 1-q_{\tilde{I}}\}^{1-\alpha}}$ for all $i \in [k]$, it holds that*

$$\max_{p\in\mathcal{P}_k}\big\{\langle \ell, q-p\rangle - D_{(-H_\alpha)}(p,q)\big\} \leq \frac{4}{1-\alpha}\left(\sum_{i\neq\tilde{I}} q_i^{2-\alpha}\ell_i^2 + \min\{q_{\tilde{I}}, 1-q_{\tilde{I}}\}^{2-\alpha}\ell_{\tilde{I}}^2\right). \qquad (56)$$

Lemma 15 ([26, Lemmas 11 and 12])**.** *Let $L \in \mathbb{R}^k$ and $\ell \in \mathbb{R}^k$ and suppose that $q, r \in \mathcal{P}_k$ are given by*

$$\begin{aligned}
q &\in \operatorname*{arg\,min}_{p\in\mathcal{P}_k}\big\{\langle L, p\rangle + \beta(-H_\alpha(p)) + \bar{\beta}(-H_{\bar{\alpha}}(p))\big\} \\
r &\in \operatorname*{arg\,min}_{p\in\mathcal{P}_k}\big\{\langle L+\ell, p\rangle + \beta'(-H_\alpha(p)) + \bar{\beta}(-H_{\bar{\alpha}}(p))\big\} \qquad (57)
\end{aligned}$$

for the Tsallis entropy H_α and $H_{\bar{\alpha}}$, $0 < \beta \leq \beta'$. Suppose also that

$$\|\ell\|_\infty \leq \max\left\{\frac{1-(\sqrt{2})^{\alpha-1}}{2}q_*^{\alpha-1}\beta, \frac{1-(\sqrt{2})^{\bar{\alpha}-1}}{2}q_*^{\bar{\alpha}-1}\bar{\beta}\right\}, \qquad (58)$$

$$0 \leq \beta' - \beta \leq \max\left\{\left(1-(\sqrt{2})^{\alpha-1}\right)\beta, \frac{1-(\sqrt{2})^{\bar{\alpha}-1}}{\sqrt{2}}q_*^{\bar{\alpha}-\alpha}\bar{\beta}\right\}. \qquad (59)$$

Then, it holds that $H_\alpha(r) \leq 2H_\alpha(q)$.

E Proof for partial monitoring (Theorem 8, Section 5)

This section provides the proof of Theorem 8.

Proof of Theorem 8. It suffices to prove that assumptions in Theorem 7 are satified. We first vertify Assumptions (i)–(iii) in (12). Let us start from checking Assumption (i). From the definition of the loss difference estimator $\widehat{y}_t$, the regret is bounded as

$$\begin{aligned}
\mathsf{Reg}_T &= \mathbb{E}\left[\sum_{t=1}^{T}(\mathcal{L}_{A_t x_t} - \mathcal{L}_{a^* x_t})\right] = \mathbb{E}\left[\sum_{t=1}^{T}\langle p_t - e_{a^*}, \mathcal{L}e_{x_t}\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\langle q_t - e_{a^*}, \mathcal{L}e_{x_t}\rangle + \sum_{t=1}^{T}\gamma_t\left\langle \frac{1}{k}\mathbf{1} - q_t, \mathcal{L}e_{x_t}\right\rangle\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\langle q_t - e_{a^*}, \mathcal{L}e_{x_t}\rangle + \sum_{t=1}^{T}\gamma_t\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{k} q_{ta}(\mathcal{L}_{a x_t} - \mathcal{L}_{a^* x_t}) + \sum_{t=1}^{T}\gamma_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{k} q_{ta}(\widehat{y}_{ta} - \widehat{y}_{ta^*}) + \sum_{t=1}^{T}\gamma_t\right] = \mathbb{E}\left[\sum_{t=1}^{T}\langle q_t - e_{a^*}, \widehat{y}_t\rangle + \sum_{t=1}^{T}\gamma_t\right],
\end{aligned} \tag{60}$$

where the inequality holds since $\mathcal{L} \in [0,1]^{k\times d}$, This implies that Assumption (i) is indeed satisfied.

We next check Assumption (ii) in (12). For any $b \in [k]$ we have

$$\left|\frac{\widehat{y}_{tb}}{\beta_t}\right| = \left|\frac{G(A_t,\sigma_t)_b}{\beta_t p_{tA_t}}\right| \le \frac{|G(A_t,\sigma_t)_b| k}{\beta_t\gamma_t} \le \frac{c_{\mathcal{G}}}{\beta_t\gamma_t} \le \frac{c_{\mathcal{G}}}{u_t} = \frac{1-\alpha}{8}\frac{1}{\left(\min\{q_{t\tilde{I}_t}, 1 - q_{t\tilde{I}_t}\}\right)^{1-\alpha}}\,, \tag{61}$$

where the third inequality follows from $\gamma_t \ge u_t/\beta_t$ and the last equality follows from the defintition of u_t in (17). Hence, from Lemma 14 the LHS of Assumption (ii) is bounded as

$$\begin{aligned}
&\mathbb{E}_t\left[\langle \widehat{y}_t, q_t - q_{t+1}\rangle - \beta_t\, D_{(-H_\alpha)}(q_{t+1}, q_t)\right] = \beta_t\mathbb{E}_t\left[\left\langle\frac{\widehat{y}_t}{\beta_t}, q_t - q_{t+1}\right\rangle - D_{(-H_\alpha)}(q_{t+1}, q_t)\right] \\
&\le \mathbb{E}_t\left[\frac{4}{\beta_t(1-\alpha)}\left(\sum_{i\ne\tilde{I}_t} q_{ti}^{2-\alpha}\widehat{y}_{ti}^2 + \left(\min\{q_{t\tilde{I}_t}, 1 - q_{t\tilde{I}_t}\}\right)^{2-\alpha}\widehat{y}_{t\tilde{I}_t}^2\right)\right] \\
&= \frac{4}{\beta_t(1-\alpha)}\left(\sum_{i\ne\tilde{I}_t} q_{ti}^{2-\alpha}\mathbb{E}_t\left[\widehat{y}_{ti}^2\right] + q_{t*}^{2-\alpha}\mathbb{E}_t\left[\widehat{y}_{t\tilde{I}_t}^2\right]\right).
\end{aligned} \tag{62}$$

Since the variance of $\widehat{y}_t$ is bounded from above as

$$\mathbb{E}_t\left[\widehat{y}_{ti}^2\right] = \sum_{a=1}^{k} p_{ta}\frac{G(a,\sigma_t)_i^2}{p_{ta}^2} \le \sum_{a=1}^{k}\frac{k\|G\|_\infty^2}{\gamma_t} = \frac{c_{\mathcal{G}}^2}{\gamma_t} \tag{63}$$

for any $i \in [k]$, the LHS of Assumption (ii) is further bounded as

$$\mathbb{E}_t[\langle \widehat{y}_t, q_t - q_{t+1}\rangle - \beta_t D_{\psi_t}(q_{t+1}, q_t)] \le \frac{4c_{\mathcal{G}}^2}{\beta_t\gamma_t(1-\alpha)}\left(\sum_{i\ne\tilde{I}_t} q_{ti}^{2-\alpha} + q_{t*}^{2-\alpha}\right) = \frac{z_t}{\beta_t\gamma_t} \le \frac{z_t}{\beta_t\gamma_t'}\,, \tag{64}$$

which implies that Assumption (ii) in (12) is satisfied.

Next, we will prove $h_{t+1} \lesssim h_t$ of Assumption (iii) in (12). To prove this, we will check the condition in Lemma 15. For any $a \in [k]$,

$$|\widehat{y}_{ta}| \le \frac{\|G\|_\infty}{p_{tA_t}} \le \frac{k\|G\|_\infty}{\gamma_t} \le \frac{c_{\mathcal{G}}\beta_t}{u_t} \le \frac{1-\alpha}{8}\frac{\beta_t}{q_{t*}^{1-\alpha}} \le \frac{1-(\sqrt{2})^{\alpha-1}}{2}\frac{\beta_t}{q_{t*}^{1-\alpha}}\,, \tag{65}$$

where the second inequality follows from $p_{ta} \ge \gamma_t/k$, the third inequality from $\gamma_t \ge u_t/\beta_t$, and the last inequality from the fact that $(1-x)/4 \le 1 - (\sqrt{2})^{x-1}$ for $x \in [0,1]$. Thus, the condition (58) is satisfied.

We next check the condition (59). Recalling $q_{t*} = \min\{q_{t\tilde{I}_t}, 1 - q_{t\tilde{I}_t}\}$, the parameters z_t and u_t satisfy

$$\sqrt{z_t} = \frac{2c_{\mathcal{G}}}{\sqrt{1-\alpha}} \sqrt{\sum_{i \neq \tilde{I}_t} q_{ti}^{2-\alpha} + q_{t*}^{2-\alpha}} \leq \frac{2\sqrt{k}c_{\mathcal{G}}}{\sqrt{1-\alpha}} q_{t*}^{1-\frac{1}{2}\alpha}, \quad u_t = \frac{8c_{\mathcal{G}}}{1-\alpha} q_{t*}^{1-\alpha}, \tag{66}$$

where the inequality follows from $q_{ti} \leq q_{t*}$ for $i \neq \tilde{I}_t$. The penalty component h_t is lower-bounded as

$$h_t = H_\alpha(q_t) = \frac{1}{\alpha} \sum_{i=1}^{k} (q_{ti}^\alpha - q_{ti}) \geq \frac{1 - (1/2)^{1-\alpha}}{\alpha} q_{t*}^\alpha \geq \frac{1-\alpha}{4\alpha} q_{t*}^\alpha, \tag{67}$$

where the last inequality in (67) follows from $1 - (1/2)^{1-x} \geq (1-x)/4$ for $x \leq 0$, and the first inequality can be proven as folows: when $q_{t\tilde{I}_t} \leq 1/2$, it holds that $\sum_{i=1}^k (q_{ti}^\alpha - q_{ti}) \geq q_{t\tilde{I}_t}^\alpha - q_{t\tilde{I}_t} = q_{t\tilde{I}_t}^\alpha (1 - q_{t\tilde{I}_t}^{1-\alpha}) \geq q_{t\tilde{I}_t}^\alpha \big(1 - (1/2)^{1-\alpha}\big) = q_{t*}^\alpha (1 - (1/2)^{1-\alpha})$, and when $q_{t\tilde{I}_t} > 1/2$, it holds that $\sum_{i=1}^k (q_{ti}^\alpha - q_{ti}) \geq \sum_{i=1}^k q_{ti}^\alpha - 1 \geq \sum_{i \neq \tilde{I}_t} q_{ti}^\alpha + (1/2)^\alpha - 1 \geq (\sum_{i \neq \tilde{I}_t} q_{ti})^\alpha + (1/2)^\alpha - 1 = (1 - q_{t\tilde{I}_t})^\alpha + (1/2)^\alpha - 1 = q_{t*}^\alpha + (1/2)^\alpha - 1 \geq q_{t*}^\alpha (1 - (1/2)^{1-\alpha})$. Using the bounds on z_t, u_t, and h_t in (66) and (67), we have

$$\begin{aligned}
\beta_{t+1} - \beta_t &= \frac{1}{\widehat{h}_{t+1}} \left(2\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} \right) = \frac{2}{h_t} \sqrt{\frac{z_t}{\beta_t}} + \frac{1}{h_t} \frac{u_t}{\beta_t} \\
&\leq \frac{16\alpha c_{\mathcal{G}} \sqrt{k}}{\sqrt{\beta_1}(1-\alpha)^{3/2}} q_{t*}^{1-\frac{3}{2}\alpha} + \frac{32\alpha c_{\mathcal{G}}}{\sqrt{\beta_1}(1-\alpha)^2} q_{t*}^{1-2\alpha} \\
&\leq \alpha \bar{\beta} q_{t*}^{1-\frac{3}{2}\alpha} + \alpha \bar{\beta} q_{t*}^{1-2\alpha} \\
&\leq 2(1-\bar{\alpha}) \bar{\beta} q_{t*}^{\bar{\alpha}-\alpha} \leq 2 \frac{1 - (\sqrt{2})^{\bar{\alpha}-1}}{\sqrt{2}} \bar{\beta} q_{t*}^{\bar{\alpha}-\alpha},
\end{aligned} \tag{68}$$

where the first inequality follows from (66), (67), and the fact that $\beta_t \geq \beta_1 \geq 1$, the second inequality from the definition of $\bar{\beta}$ in (17), the third inequality from $\min\{1 - \frac{3}{2}\alpha, 1 - 2\alpha\} \geq \bar{\alpha} - \alpha$ since $\bar{\alpha} = 1 - \alpha$, and the last inequality from $1 - x \leq (1 - (\sqrt{2})^{x-1})/\sqrt{2}$ for $x \leq 1$. Therefore, the condition (59) is satified. Hence, from Lemma 15, we have $h_{t+1} = H_\alpha(q_{t+1}) \leq 2H_\alpha(q_t) = 2h_t$, which implies that Assumption (iii) in (12) is satisfied.

Finally, we check the assumption (14) in Theorem 7. We first consider the first inequality in (14). From the definition of z_t and the fact that $q_{ti} \leq q_{t\tilde{I}_t}$ for $i \neq \tilde{I}_t$, the stability component z_t is bounded as

$$\begin{aligned}
z_t &= \frac{4c_{\mathcal{G}}^2}{1-\alpha} \left\{ \sum_{i \neq \tilde{I}_t} q_{ti}^{2-\alpha} + \big(\min\{q_{t\tilde{I}_t}, 1 - q_{t\tilde{I}_t}\}\big)^{2-\alpha} \right\} \\
&\leq \frac{4c_{\mathcal{G}}^2}{1-\alpha} \left\{ \sum_{i \neq \tilde{I}_t} q_{ti}^{2-\alpha} + \left(\sum_{i \neq \tilde{I}_t} q_{ti} \right)^{2-\alpha} \right\} \\
&\leq \frac{8c_{\mathcal{G}}^2}{1-\alpha} \left(\sum_{i \neq \tilde{I}_t} q_{ti} \right)^{2-\alpha} \leq \frac{8c_{\mathcal{G}}^2}{1-\alpha} \left(\sum_{i \neq a^*} q_{ti} \right)^{2-\alpha} = \frac{8c_{\mathcal{G}}^2}{1-\alpha} (1 - q_{ta^*})^{2-\alpha},
\end{aligned} \tag{69}$$

where the second inequality holds from the inequality $x^a + y^a \leq (x+y)^a$ for $x, y \geq 0$ and $a \in [0,1]$, and the third inequality from $q_{ti} \leq q_{t\tilde{I}_t}$ for $i \neq \tilde{I}_t$. From Lemma 13, we also obtain that

$$h_t = H_\alpha(q_t) \leq \frac{1}{\alpha} (k-1)^{1-\alpha} (1 - q_{ta^*})^\alpha. \tag{70}$$

Hence, combining this with (69), we obtain

$$z_t h_t \leq \frac{8c_{\mathcal{G}}^2}{1-\alpha} (1 - q_{ta^*})^{2-\alpha} \cdot \frac{1}{\alpha} (k-1)^{1-\alpha} (1 - q_{ta^*})^\alpha = \underbrace{\frac{8c_{\mathcal{G}}^2 (k-1)^{1-\alpha}}{\alpha(1-\alpha)}}_{=\rho_1} (1 - q_{ta^*})^2. \tag{71}$$

We next consider the second inequality in (14). We can bound u_t from above as

$$u_t = \frac{8c_{\mathcal{G}}}{1-\alpha}\left(\min\{q_{t\tilde{I}_t}, 1-q_{t\tilde{I}_t}\}\right)^{1-\alpha} \leq \frac{8c_{\mathcal{G}}}{1-\alpha}\left(\sum_{i\neq\tilde{I}_t} q_{ti}\right)^{1-\alpha}$$
$$\leq \frac{8c_{\mathcal{G}}}{1-\alpha}\left(\sum_{i\neq a^*} q_{ti}\right)^{1-\alpha} = \frac{8c_{\mathcal{G}}}{1-\alpha}(1-q_{ta^*})^{1-\alpha}\,, \tag{72}$$

where the second inequality follows from $q_{t\tilde{I}_t} \geq q_{ti}$ for all $i \in [k]$. Hence, combining the last two inequality and (70),

$$u_t h_t \leq \underbrace{\frac{4c_{\mathcal{G}}(k-1)^{1-\alpha}}{\alpha(1-\alpha)}}_{=\rho_2}(1-q_{ta^*})\,. \tag{73}$$

Hence, the assumption (14) is satified with above ρ_1 and ρ_2, and thus we have completed the proof. $\square$

F Proof for graph bandits (Theorem 10, Section 6)

This section provides the missing detail of Section 6.

F.1 Fractional domination number

Before introducing the fractional domination number, we define the domination number $\tilde{\delta} \leq \delta$. A *dominating set* $D \subseteq V$ is a set of vertices such that $V \subseteq \bigcup_{i\in D} N^{\mathsf{out}}(i)$. The *domination number* $\tilde{\delta}(G)$ of graph G is the size of the smallest dominating set. From the definition, the domination number $\tilde{\delta}$ can also be written as the optimal value of the following optimization problem:

$$\text{minimize} \sum_{i\in V} x_i \quad \text{subject to} \quad \sum_{i\in N^{\mathsf{in}}(j)} x_i \geq 1 \;\forall j \in V\,,\; x_i \in \{0,1\} \;\forall i \in V\,, \tag{74}$$

where $x_i \in \{0,1\}$ a binary variable indicating whether vertex i is in the dominating set ($x_i = 1$) or not ($x_i = 0$).

Then, one can see that the fractional domination number δ^* is defined as the optimal value of the following optimization problem, in which the variables $(x_i)_{i\in V}$ are allowed to take values in $[0,1]$ instead of $\{0,1\}$:

$$\text{minimize} \sum_{i\in V} x_i \quad \text{subject to} \quad \sum_{i\in N^{\mathsf{in}}(j)} x_i \geq 1 \;\forall j \in V\,,\; 0 \leq x_i \leq 1 \;\forall i \in V\,, \tag{75}$$

which is the linear program provided in (19). From the definitions, the fractional domination number is less than or equal to the domination number, $\delta^* \leq \tilde{\delta}$. Another advantage of using δ^* instead of $\tilde{\delta}$ is that the fractional domination number δ^* can be computed in polynomial time, while the computation of the domination number $\tilde{\delta}$ is NP-hard. See [13] for more benefits of using the fractional version of the (weak) domination number.

F.2 Proof of Theorem 10

Here, we provide the proof of Theorem 10.

Proof. It suffices to prove that assumptions in Theorem 7 are satified. We first vertify Assumptions (i)–(iii) in (12). We start from checking Assumption (i). The regret is bounded as

$$\mathsf{Reg}_T = \mathbb{E}\left[\sum_{t=1}^T \ell_t(A_t) - \sum_{t=1}^T \ell_t(a^*)\right] = \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, p_t - e_{a^*}\rangle\right] = \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, q_t - e_{a^*}\rangle + \sum_{t=1}^T \langle \ell_t, p_t - q_t\rangle\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, q_t - e_{a^*}\rangle + \sum_{t=1}^T \gamma_t \langle \ell_t, q_t - u\rangle\right] \leq \mathbb{E}\left[\sum_{t=1}^T \langle \widehat{\ell}_t, q_t - e_{a^*}\rangle + \sum_{t=1}^T \gamma_t\right]\,, \tag{76}$$

where the third equality follows from the defintion of γ_t. This implies that Assumption (i) is indeed satisfied.

We next check Assumption (ii) in (12). Now, recalling the defintion of the fractional domination number and the optimal value x^* of (19), and $u_i = x_i^* / \sum_{j \in V} x_j^*$, we have

$$\sum_{j \in N^{\mathsf{in}}(i)} u_j = \frac{\sum_{j \in N^{\mathsf{in}}(i)} x_j^*}{\sum_{i \in V} x_i^*} \geq \frac{1}{\sum_{i \in V} x_i^*} = \frac{1}{\delta^*}\,, \tag{77}$$

where the inequality follows from the first constraint in (19). Hence, combining this with the definition of $p_t = (1-\gamma_t) q_t + \gamma_t u$, we can lower-bound P_{ti} as

$$P_{ti} = \sum_{j \in N^{\mathsf{in}}(i)} p_{tj} \geq \gamma_t \sum_{j \in N^{\mathsf{in}}(i)} u_j \geq \frac{\gamma_t}{\delta^*} \quad \text{for all } i \in V\,. \tag{78}$$

This lower bound yields that for any $i \in V$

$$\left| \frac{\widehat{\ell}_{ti}}{\beta_t} \right| = \frac{\ell_{ti}}{\beta_t P_{ti}} \leq \frac{\delta^*}{\beta_t \gamma_t} = \frac{\delta^*}{u_t} = \frac{1-\alpha}{8} \frac{1}{\left(\min\{q_{t\tilde{I}_t}, 1 - q_{t\tilde{I}_t}\}\right)^{1-\alpha}}\,. \tag{79}$$

Hence, from Lemma 14 we obtain

$$\begin{aligned}
&\mathbb{E}_t\left[\left\langle \widehat{\ell}_t, q_t - q_{t+1} \right\rangle - \beta_t\, D_{(-H_\alpha)}(q_{t+1}, q_t)\right] = \beta_t \mathbb{E}_t\left[\left\langle \frac{\widehat{\ell}_t}{\beta_t}, q_t - q_{t+1} \right\rangle - D_{(-H_\alpha)}(q_{t+1}, q_t)\right] \\
&\leq \mathbb{E}_t\left[\frac{4}{\beta_t(1-\alpha)}\left(\sum_{i \in V \setminus \{\tilde{I}_t\}} q_{ti}^{2-\alpha} \widehat{\ell}_{ti}^2 + \left(\min\{q_{t\tilde{I}_t}, 1 - q_{t\tilde{I}_t}\}\right)^{2-\alpha} \widehat{\ell}_{t\tilde{I}_t}^2\right)\right] \\
&= \frac{4}{\beta_t(1-\alpha)}\left(\sum_{i \in V \setminus \{\tilde{I}_t\}} q_{ti}^{2-\alpha} \mathbb{E}_t\left[\widehat{\ell}_{ti}^2\right] + q_{t*}^{2-\alpha} \mathbb{E}_t\left[\widehat{\ell}_{t\tilde{I}_t}^2\right]\right)\,.
\end{aligned} \tag{80}$$

Then, by using the lower bound of P_t in (78), for any $i \in V$ the variance of the loss estimator $\widehat{\ell}_{ti}$ is bounded as

$$\mathbb{E}_t\left[\widehat{\ell}_{ti}^2\right] = \sum_{j=1}^k p_{tj} \frac{\ell_{ti}^2}{P_{ti}^2} \mathbb{1}\left[i \in N^{\mathsf{out}}(j)\right] = \frac{\ell_{ti}^2}{P_{ti}^2} \sum_{j \in V \colon i \in N^{\mathsf{out}}(j)} p_{tj} = \frac{\ell_{ti}^2}{P_{ti}} \leq \frac{\delta^*}{\gamma_t}\,. \tag{81}$$

Hence, combining (80) with (81), we obtain

$$\mathbb{E}_t[\langle \widehat{y}_t, q_t - q_{t+1} \rangle - \beta_t D_{\psi_t}(q_{t+1}, q_t)] \leq \frac{4\delta^*}{\beta_t \gamma_t (1-\alpha)}\left(\sum_{i \in V \setminus \{\tilde{I}_t\}} q_{ti}^{2-\alpha} + q_{t*}^{2-\alpha}\right) = \frac{z_t}{\beta_t \gamma_t} \leq \frac{z_t}{\beta_t \gamma_t'}\,, \tag{82}$$

which implies that Assumption (ii) in (12) is satisfied.

Next, we will prove $h_{t+1} \lesssim h_t$ of Assumption (iii) in (12). To prove this, we will check the condition in Lemma 15. For any $i \in V$,

$$|\widehat{\ell}_{ti}| \leq \frac{1}{P_{ti}} \leq \frac{\delta^*}{\gamma_t} \leq \frac{\delta^* \beta_t}{u_t} = \frac{1-\alpha}{8} \frac{\beta_t}{q_{t*}^{1-\alpha}} \leq \frac{1 - (\sqrt{2})^{\alpha-1}}{2} \frac{\beta_t}{q_{t*}^{1-\alpha}}\,, \tag{83}$$

where the second inequality follows from (78), the third inequality from $\gamma_t \geq u_t/\beta_t$, and the last inequality from the fact that $(1-x)/4 \leq 1 - (\sqrt{2})^{x-1}$ for $x \in [0,1]$. Thus, the condition (58) is satisfied.

We next check the condition (59). Recalling $q_{t*} = \min\{q_{t\tilde{I}_t}, 1 - q_{t\tilde{I}_t}\}$, we observe that the parameters z_t and u_t satisfy

$$\sqrt{z_t} = \sqrt{\frac{4\delta^*}{1-\alpha}\left(\sum_{i \in V \setminus \{\tilde{I}_t\}} q_{ti}^{2-\alpha} + q_{t*}^{2-\alpha}\right)} \leq \frac{2\sqrt{k\delta^*}}{\sqrt{1-\alpha}} q_{t*}^{1-\frac{1}{2}\alpha}\,, \quad u_t = \frac{8\delta^*}{1-\alpha} q_{t*}^{1-\alpha}\,, \tag{84}$$

where the last inequality follows from $q_{ti} \leq q_{t*}$ for $i \neq \tilde{I}_t$. We can also lower-bound h_t as

$$h_t = H_\alpha(q_t) = \frac{1}{\alpha}\sum_{i=1}^{k}(q_{ti}^\alpha - q_{ti}) \geq \frac{1-(1/2)^{1-\alpha}}{\alpha} q_{t*}^\alpha \geq \frac{1-\alpha}{4\alpha} q_{t*}^\alpha \,, \tag{85}$$

which can be proven by the same manner as in (67). Hence, using the upper bounds on z_t, u_t, and h_t in (84) and (85), we have

$$\begin{aligned}
\beta_{t+1} - \beta_t &= \frac{1}{\widehat{h}_{t+1}}\left(2\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t}\right) = \frac{2}{h_t}\sqrt{\frac{z_t}{\beta_t}} + \frac{1}{h_t}\frac{u_t}{\beta_t} \\
&\leq \frac{16\alpha\sqrt{k\delta^*}}{\sqrt{\beta_1}(1-\alpha)^{3/2}} q_{t*}^{1-\frac{3}{2}\alpha} + \frac{32\alpha\delta^*}{\sqrt{\beta_1}(1-\alpha)^2} q_{t*}^{1-2\alpha} \\
&\leq \alpha\bar{\beta} q_{t*}^{1-\frac{3}{2}\alpha} + \alpha\bar{\beta} q_{t*}^{1-2\alpha} \\
&\leq 2(1-\bar{\alpha})\bar{\beta} q_{t*}^{\bar{\alpha}-\alpha} \leq 2\frac{1-(\sqrt{2})^{\bar{\alpha}-1}}{\sqrt{2}}\bar{\beta} q_{t*}^{\bar{\alpha}-\alpha} \,,
\end{aligned} \tag{86}$$

where the first inequality follows from (84), (85), and $\beta_t \geq \beta_1 \geq 1$, the second inequality from the definition of $\bar{\beta}$, the third inequality from $\min\{1-\frac{3}{2}\alpha, 1-2\alpha\} \geq \bar{\alpha} - \alpha$ since $\bar{\alpha} = 1-\alpha$, and the last inequality from $1 - x \leq (1-(\sqrt{2})^{x-1})/\sqrt{2}$ for $x \leq 1$. Thus the condition (59) is satisfied. Therefore, from Lemma 15, we have $h_{t+1} = H_\alpha(q_{t+1}) \leq 2H_\alpha(q_t) = 2h_t$, which implies that Assumption (iii) in (12) is satisfied.

Finally, we check the assumption (14) in Theorem 7. We first consider the first inequality in (14). From the definition of z_t and the fact that $q_{ti} \leq q_{t\tilde{I}_t}$ for $i \neq \tilde{I}_t$, we get

$$\begin{aligned}
z_t &= \frac{4\delta^*}{1-\alpha}\left\{\sum_{i\in V\setminus\{\tilde{I}_t\}} q_{ti}^{2-\alpha} + \left(\min\{q_{t\tilde{I}_t}, 1-q_{t\tilde{I}_t}\}\right)^{2-\alpha}\right\} \\
&\leq \frac{4\delta^*}{1-\alpha}\left\{\sum_{i\in V\setminus\{\tilde{I}_t\}} q_{ti}^{2-\alpha} + \left(\sum_{i\neq\tilde{I}_t} q_{ti}\right)^{2-\alpha}\right\} \\
&\leq \frac{8\delta^*}{1-\alpha}\left(\sum_{i\in V\setminus\{\tilde{I}_t\}} q_{ti}\right)^{2-\alpha} \leq \frac{8\delta^*}{1-\alpha}\left(\sum_{i\neq a^*} q_{ti}\right)^{2-\alpha} = \frac{8\delta^*}{1-\alpha}(1-q_{ta^*})^{2-\alpha} \,,
\end{aligned} \tag{87}$$

where the second inequality holds from the inequality $x^a + y^a \leq (x+y)^a$ for $x, y \geq 0$ and $a \in [0,1]$, and the third inequality from $q_{ti} \leq q_{t\tilde{I}_t}$. Hence, combining this with (87) with the upper bound on h_t in (70), we obtain

$$z_t h_t \leq \frac{8\delta^*}{1-\alpha}(1-q_{ta^*})^{2-\alpha} \cdot \frac{1}{\alpha}(k-1)^{1-\alpha}(1-q_{ta^*})^\alpha = \underbrace{\frac{8\delta^*(k-1)^{1-\alpha}}{\alpha(1-\alpha)}}_{=\rho_1}(1-q_{ta^*})^2 \,. \tag{88}$$

We next consider the second inequality in (14). We can bound u_t from above as

$$\begin{aligned}
u_t &= \frac{8\delta^*}{1-\alpha}\left(\min\{q_{t\tilde{I}_t}, 1-q_{t\tilde{I}_t}\}\right)^{1-\alpha} \leq \frac{8\delta^*}{1-\alpha}\left(\sum_{i\neq\tilde{I}_t} q_{ti}\right)^{1-\alpha} \\
&\leq \frac{8\delta^*}{1-\alpha}\left(\sum_{i\neq a^*} q_{ti}\right)^{1-\alpha} = \frac{8\delta^*}{1-\alpha}(1-q_{ta^*})^{1-\alpha} \,,
\end{aligned} \tag{89}$$

where the second inequality follows from $q_{t\tilde{I}_t} \geq q_{ti}$ for all $i \neq \tilde{I}_t$. Hence, combining the last inequality with (70),

$$u_t h_t \leq \underbrace{\frac{4\delta^*(k-1)^{1-\alpha}}{\alpha(1-\alpha)}}_{=\rho_2}(1-q_{ta^*}) \,. \tag{90}$$

Hence, the assumption (14) is satified with above ρ_1 and ρ_2, and thus we have completed the proof. $\square$

F.3 Technical challenges to derive best-of-both-worlds bounds depending on (fractional) weak domination number

Here, we discuss the technical challenges of making our upper bound in Theorem 10 depend on the weak domination number δ instead of the fracional domination number δ^* or the weak fractional domination number $\tilde{\delta}^* \leq \delta$.

First, we need to use Tsallis entropy to derive a regret upper bound with a stochastic bound of $\log T$. While we can prove a BOBW bound if we use the Shannon entropy regularizer [25], the bound in the stochastic regime is $O((\log T)^2)$, which is not desirable. which is not desirable. Hence, a possible approach is to use the log-barrier regularizer or the Tsallis entropy. The log-barrier regularizer has a penalty term of $\Omega(k)$ due to the strength of its regularization, and the regret upper bound in the final adversarial regime is $\Omega(k^{1/3})$, which can be much larger than $\delta^{1/3}$. Therefore, the most hopeful solution would be to use Tsallis entropy with an appropriate exponent $\alpha \simeq 1$, where we note that the Tsallis entropy with $\alpha \to 1$ corresponds to the Shanon entropy.

Recalling the definition of the weak domination number in Section 6, we can see that the weak domination set dominates only vertices without self-loop $U = \{i \in V : i \notin N^{\mathsf{out}}(i)\}$. Thus, to achieve a BOBW bound that depends on the weak domination number, vertices with self-loop and those without self-loop should be treated separately by decomposing the stability term as follows:

$$
\begin{aligned}
&\langle \widehat{\ell}_t, q_t - q_{t+1} \rangle - \beta_t \, D_{(-H_\alpha)}(q_{t+1}, q_t) \\
&= \sum_{i \in U} \Big(\widehat{\ell}_{ti}(q_{ti} - q_{t+1,i}) - \beta_t \, d(q_{t+1,i}, q_{t,i}) \Big) + \sum_{i \in V \setminus U} \Big(\widehat{\ell}_{ti}(q_{ti} - q_{t+1,i}) - \beta_t \, d(q_{t+1,i}, q_{t,i}) \Big) \,,
\end{aligned}
$$

where $d(p, q)$ is the Bregman divergence induced by the real-valued convex function $x \mapsto -\frac{1}{\alpha}(x^\alpha - x)$. However, if we use this approach, we cannot use Lemma 14, which is useful to prove an upper bound with $(1 - q_{ta^*})$ (see (14)). This is because this lemma exploits the fact that q and r are probability vectors. This prevents us from deriving an upper bound with an $O(\log T)$ stochastic bound depending on the weak domination number.