Adaptive Learning Rates with Surrogate Probability for Follow-the-Perturbed-Leader

Jongyeong Lee JONGYEONG@KIST.RE.KR
Korea Institute of Science and Technology

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

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

Chansoo Kim EAU@UST.AC.KR
Korea Institute of Science and Technology, University of Science and Technology

Editors: Steve Hanneke and Tor Lattimore

Abstract

Follow-the-regularized-leader framework has shown effectiveness and flexibility in online learning problems, where the choice of learning rates are known to be crucial. Recently, adaptive learning rates defined in terms of the arm-selection probabilities, obtained by solving convex optimization, have achieved improved best-of-both-worlds (BOBW) guarantees in various bandit problems. In contrast, BOBW guarantees for its computationally efficient alternative, follow-the-perturbed-leader (FTPL), remain relatively limited since its optimization-free nature ironically makes the design of adaptive, probability-dependent learning rates non-trivial. To address this challenge, we propose an adaptive learning rate for FTPL by introducing surrogate probability functions that can be computed only from the available quantities, without requiring the exact probabilities. Based on these learning rates with surrogate functions, we provide the BOBW guarantee for FTPL with Pareto perturbations for any shape parameter $\alpha > 1$, generalizing prior results restricted to specific choices of $\alpha = 2$. We further show the BOBW guarantees for FTPL with adaptive learning rates in the bandit problem with expert advices. Our approach preserves the computational simplicity of FTPL while enabling probability-dependent adaptivity, and the surrogate-based methodology may be of independent interest in other algorithmic frameworks beyond FTPL and learning rate designs.

Keywords: follow-the-perturbed-leader, adaptive learning rate, best-of-both-worlds

1. Introduction

The multi-armed bandit (MAB) problem is a fundamental problem for sequential decision making under uncertainty. In this problem, at each round $t \in [T] := \{1, \dots, T\}$, an agent selects an arm i_t from a set of K arms and observes the corresponding loss ℓ_{t,i_t} from i_t, where the loss vectors $\ell_t = (\ell_{t,1}, \dots, \ell_{t,K}) \in [0,1]^K$ are determined by the environment. A central challenge in bandit problems is to minimize the cumulative loss by learning the environments only with partial feedback, where two canonical types of environments have been extensively studied.

In the stochastic regime, losses are identically independently distributed (i.i.d.) from an unknown but fixed distribution (Lai and Robbins, 1985; Katehakis and Robbins, 1995). In contrast, in the adversarial regime, losses may be chosen by an adversary, possibly adaptively based on the agent's past actions, so that no distributional assumption can be made (Auer et al., 2002). Since the true nature of the environment is usually unknown in practice, there has been considerable interest in

policies that achieve (near-)optimal performance guarantees in both regimes, which is referred to as Best-of-Both-Worlds (BOBW) guarantee (Bubeck and Slivkins, 2012; Seldin and Lugosi, 2017).

A dominant approach to achieving BOBW guarantees would be the Follow-the-Regularized-Leader (FTRL) framework, which has successfully obtained BOBW guarantees across a wide range of online learning problems including the standard multi-armed bandits (Zimmert and Seldin, 2021; Jin et al., 2023), decoupled bandits (Rouyer and Seldin, 2020), combinatorial-semi bandit (Zimmert et al., 2019; Ito, 2021), and partial monitoring (Tsuchiya and Ito, 2024), to name a few. This success can be attributed to its generality and flexibility: one can tailor the policy to the problem structure by carefully choosing the (convex) regularizers and learning rates. However, this generality comes at a cost, as it requires solving a convex optimization problem at each round to compute the arm-selection distribution, which can be computationally demanding in practice.

This limitation has motivated a line of work aimed at avoiding per-round optimization while preserving desirable regret guarantees. One approach is to approximate FTRL updates using simpler arithmetic operations, as in the Prod family of policies (Cesa-Bianchi et al., 2007; Gaillard et al., 2014), where Zimmert and Marinov (2024) obtained BOBW guarantee by leveraging a kind of first order approximation of FTRL with Tsallis entropy. Another prominent alternative is the Follow-the-Perturbed-Leader (FTPL) framework, which selects arms by adding random perturbations to cumulative losses (Poland, 2005; Kalai and Vempala, 2005). FTPL has gained attention due to its simplicity and (almost) optimization-free nature. Moreover, there is a deep theoretical correspondence between FTRL and FTPL, where specific FTRL is associated with particular perturbation distributions in FTPL (Abernethy et al., 2016; Li et al., 2024; Lee et al., 2025). Recent work has also established BOBW guarantees for FTPL in standard MABs (Honda et al., 2023; Lee et al., 2024), decoupled bandits (Kim et al., 2026), and combinatorial semi-bandits (Zhan et al., 2025; Chen et al., 2026).

Despite these recent advances, progress on BOBW guarantees for FTPL still remains limited compared with that for FTRL. In particular, recent advances in the FTRL framework exploit explicit arm-selection probabilities to design adaptive learning rates and employ hybrid regularizers, to obtain BOBW guarantees in various settings (Jin et al., 2023; Tsuchiya and Ito, 2024; Zhao et al., 2025). However, these techniques do not transfer straightforwardly to FTPL, where probabilities do not admit a closed form and are induced implicitly only through perturbations. In other words, while perturbations remove the need for optimization, they conversely appear to degrade the adaptivity of FTPL that has been crucial to recent advances in FTRL framework with adaptive learning rates.

Contribution. In this paper, we address this gap by developing adaptive learning rates for FTPL that preserve computational efficiency while achieving BOBW guarantees, *without requiring explicit arm-selection probabilities*. Our approach is inspired by the stability-penalty matching (SPM) methodology developed for FTRL frameworks (Ito et al., 2024; Nguyen et al., 2025) and by recent uses of surrogate probabilities in FTPL (Kim et al., 2026). The key technical idea is to introduce surrogate probability functions that replace true arm-selection probabilities and can be computed solely from the currently available quantities.

Based on SPM learning rates with surrogate probabilities, we first generalize existing BOBW guarantee of FTPL in the standard MAB. In particular, we show that FTPL with Pareto perturbations of any shape $\alpha > 1$ achieves BOBW guarantee, whereas previous results were restricted to the case $\alpha = 2$ (Lee et al., 2024). This result aligns with those obtained for FTRL with γ-Tsallis entropy for general $\gamma \in (0, 1)$ (Jin et al., 2023; Ito et al., 2024), extending beyond the case $\gamma = 1/2$ (Zimmert and Seldin, 2021). We further extend the BOBW guarantee for FTPL with SPM learning rates to

the bandit problems with expert advice (Auer et al., 2002), which is also referred to as contextual bandits (Dann et al., 2023).

One of the main advantage of our approaches over FTRL-based approaches lies in its simplicity and computational efficiency, especially when the hybrid regularizers are considered. While FTRL only with Tsallis entropy admits a solution that can be formulated as one-dimensional optimization problems, which can be efficiently computed via Newton's method or bisection method (Zimmert and Seldin, 2021), recent improvements in regret guarantees often rely on hybrid regularizers (Tsuchiya et al., 2023a; Jin et al., 2023; Ito et al., 2024; Nguyen et al., 2025). These regularizers may require solving a convex optimization problem at every round. In contrast, our policy follows the standard FTPL framework with Pareto perturbations and computes learning rates directly from currently available quantities, completely avoiding convex optimizations.

Besides its computational efficiency, the use of surrogate probability functions that do not belong to the probability simplex may be of independent interest. Although our analysis is grounded in FTPL, this surrogate-based approach could also be applicable to other algorithmic frameworks beyond FTPL. More broadly, it could extend to settings where explicit probability vectors are used, not limited to the design of adaptive learning rates. For example, in heavy-tailed bandits, Huang et al. (2022) addressed the effect of extreme observations by skipping large losses, where the skipping thresholds are chosen adaptively based on the arm-selection probability. We expect that our approach could be used to replace such explicit arm-selection probabilities with surrogate probability functions.

2. Preliminaries

In this section, we introduce notation and formulate the problem. Then, we introduce the intuition behind stability-penalty matching (SPM) methods in FTRL frameworks.

2.1. Problem formulation

In bandit settings, the environment determines the loss vector $\ell_t \in [0,1]^K$ and the agent selects an arm i_t at each round, where the performance of the agent's policy is measured by the pseudo-regret. When w_t denotes the arm-selection probabilities of policy at round t, the pseudo-regret is defined by

$$\mathrm{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \langle \ell_t, w_t - e_{i^*} \rangle\right], \; i^* \in \operatorname*{arg\,min}_{i \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right].$$

Here, i^* denotes the optimal arm in hindsight and is assumed to be unique following the prior studies (Lee et al., 2024; Ito et al., 2024). Since only partial feedback is observable, an estimator $\hat{\ell}_t$ of the loss vector ℓ_t can be used, which is specified later.

In this paper, we considers two possible environments, the adversarial regime (Auer et al., 2002) and adversarial regime with self-bounding constraints (Zimmert and Seldin, 2021). In the adversarial regime, ℓ_t is determined in an adversarial way possibly depending on the history, $\{(\ell_s, i_s)\}_{s=1}^{t-1}$. The adversarial regime with a (Δ, C, T) self-bounding constraint is an environment where the regret can be bounded from below as follows:

$$\mathrm{Reg}(T) \geq \mathrm{Reg}'(T) - C, \text{ where } \mathrm{Reg}'(T) = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{i_t}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{K} \Delta_i w_{t,i}\right].$$

This regime also includes the stochastic environments with adversarial corruption (Wei and Luo, 2018), where each $\Delta_i \geq 0$ is equivalent to the suboptimality gap of arm i and C is the magnitude of

corruption. Note that the unique optimal i^* assumption implies $\Delta_i > 0$ holds for all $i \neq i^*$, where we denote $\Delta_{\min} = \min_{i \neq i^*} \Delta_i$ in the adversarial regime with (Δ, C, T) self-bounding constraint.

2.2. Follow-the-Perturbed-Leader

Let $\hat{L}_t = \sum_{s=1}^{t-1} \hat{\ell}_s$ be the cumulative loss estimator up to round $t-1$. Then, FTPL is a policy that selects an arm i_t according to

$$i_t = \mathop{\arg\min}_{i \in [K]} \left\{ \hat{L}_{t,i} - \frac{r_{t,i}}{\eta_t} \right\} = \mathop{\arg\min}_{i \in [K]} \left\{ \eta_t \underline{\hat{L}}_{t,i} - r_{t,i} \right\}, \text{ where } r_{t,i} \overset{\text{i.i.d.}}{\sim} \mathcal{D}, \forall i \in [K]. \tag{1}$$

Here, the underline denotes the gap of a vector from its minimum, i.e., $\underline{\lambda} = \lambda - \mathbf{1} \min_{i \in [K]} \lambda_i$ for all-one vector $\mathbf{1}$, and the learning rate η_t will be defined later. In this paper, we consider perturbations generated from the shifted Pareto distribution with shape α, also known as the Lomax distribution with shape α and scale 1, whose density function f and distribution function F are defined by

$$f(z) = \frac{\alpha}{(z+1)^{\alpha+1}}, \text{ and } F(z) = 1 - \frac{1}{(1+z)^{\alpha}}, \forall z \geq 0. \tag{2}$$

Then, the arm-selection probability given $\hat{L}_t$ can be written as $w_{t,i} = \Pr[i_t = i | \hat{L}_t] = \phi_i(\eta_t \hat{L}_t)$, where for $\lambda \in [0, \infty)^K$

$$\phi_i(\lambda) := \Pr_{r_1, \ldots, r_K \sim \mathcal{D}} [i_t = i | \lambda] = \int_0^\infty f(z + \underline{\lambda}_i) \prod_{j \neq i} F(z + \underline{\lambda}_j) \mathrm{d}z. \tag{3}$$

While the importance-weighted (IW) estimator $\hat{\ell}_{t,i} = \mathbb{1}[i_t = i] \ell_{t,i} / w_{t,i}$ is commonly used, w_t of FTPL in (3) generally does not admit a closed form, which complicates its direct use. Therefore, it is standard to estimate $1/w_{t,i}$ in FTPL via resampling-based procedures (Abernethy et al., 2016; Honda et al., 2023). In particular, Neu and Bartók (2016) proposed the geometric resampling (GR) method, which produces an unbiased estimator of $1/w_{t,i_t}$ by repeatedly sampling perturbations r'_t from the perturbation distribution until the FTPL rule selects the same i_t arm with resampled r'_t.

2.3. Stability-penalty matching learning rates for FTRL

In the standard regret decomposition of FTRL, it is well known that the regret is bounded from above as (Lattimore and Szepesvári, 2020, Exercise 28.12)

$$\mathrm{Reg}(T) \lesssim \underbrace{\sum_{t=1}^{T} \eta_t z_t}_{\text{stability term}} + \underbrace{\sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) h_{t+1}}_{\text{penalty term}}, \tag{4}$$

where the formulations of z_t and h_t depend on the problem setting and the regularizer function of FTRL. While the learning rates η_t may be designed as a function of either z_t or h_t, recent advances in the analysis of FTRL suggest using η_t so that the contributions of the stability and penalty terms to be of the same order, which is referred to as stability–penalty matching (SPM) (Jin et al., 2023; Tsuchiya et al., 2023b). In particular, for $\beta_t := \eta_t^{-1}$, the update of SPM learning rates roughly takes the form of

$$\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_{t+1}} \implies \mathrm{Reg}(T) \lesssim \sum_{t=1}^{T} \frac{z_t}{\beta_t}, \tag{5}$$

where the appropriate choices of z_t and h_t have shown the BOBW guarantees in various bandit problems (Ito et al., 2024; Zhao et al., 2025). While the idea of SPM is intuitive, a technical obstacle arises from the appearance of h_{t+1} in the update rule since computing h_{t+1} usually requires information from the next round, which itself depends on β_{t+1}. To resolve this circular dependence, prior FTRL policies employed hybrid regularizers to ensure that $h_{t+1} = \mathcal{O}(h_t)$, thereby justifying the use of h_t in place of h_{t+1} in the update of β_{t+1} in (5) (Ito et al., 2024; Nguyen et al., 2025).

In the standard FTRL analysis, z_t usually depends on the arm-selection probabilities $w_{t,i}$ and h_t is determined by the value of the regularizer at $w_{t,i}$. For example, in the multi-armed bandits, FTRL with γ-Tsallis entropy satisfies $z_t \lesssim \sum_i (\mathbb{E}[w_{t,i}])^{1-\gamma}$ and $h_t \approx \sum_i (\mathbb{E}[w_{t,i}])^{\gamma}$ (Zimmert and Seldin, 2021). Hence, the update of β_{t+1} in Ito et al. (2024) explicitly relies on the values of $w_{t,i}$ at each round, which are obtained by solving the associated convex optimization in FTRL.

3. Adaptive learning rates for FTPL in multi-armed bandits

In this section, we propose an adaptive learning rate under SPM principle tailored to FTPL, which can be computed only from quantities available at the current round.

Given $\hat{L}_t$, we define $\sigma_{t,i}$ as the rank of $\hat{L}_{t,i}$ among $\{\hat{L}_{t,j}\}_{j\in[K]}$, where $\sigma_{t,i} = 1$ if $\hat{L}_{t,i}$ is the smallest and $\sigma_{t,i} = K$ to the largest with arbitrary tie-breaking rule. Let j_t be the arm satisfying $\sigma_{t,j_t} = 1$ after tie-breaking. Note that even if multiple arms satisfy $\underline{\hat{L}}_{t,i} = 0$, the tie-breaking rule ensures that there is a unique arm with $\sigma_{t,i} = 1$.

3.1. FTPL with conditional geometric resampling

Since w_t of FTPL does not admit the closed form in general, constructing the IW loss estimator $\hat{\ell}_t$ requires estimating $1/w_{t,i_t}$. In the standard multi-armed bandit setting, we adopt conditional geometric resampling (CGR) (Chen et al., 2025), which improves the computational efficiency of the original GR (Neu and Bartók, 2016). Beyond its computational advantages, CGR (precisely, CGR II-biased) provides a bounded loss estimator without sacrificing BOBW guarantees, which also simplifies the analysis, especially for $\alpha \in (1, 2)$. For $\alpha \geq 2$, the use of CGR II-biased is not essential as it is also possible to obtain BOBW guarantee with the original GR by applying the results in previous BOBW analysis (Honda et al., 2023; Lee et al., 2024).

The main idea of CGR is to avoid generating new perturbations that clearly violate the termination condition, under which the FTPL rule in (1) never selects i_t. This is achieved by restricting the perturbation distributions during the resampling step to satisfy certain necessary conditions. In particular, we employ the CGR II that generates the fresh perturbations from appropriately truncated conditional distributions so that resampled perturbation for the selected arm r'_{t,i_t} is larger than $\eta_t \underline{\hat{L}}_{t,i_t}$ as well as $r'_{t,j}$ for all arms with $\sigma_{t,j} < \sigma_{t,i_t}$. More details including explicit sampling distributions and explicit policy is provided in Appendix A.

Let M_t denote the number of resampling steps until the same arm i_t is selected again by FTPL rule with these resampled perturbations and G_t denote the maximum number of allowed resampling steps. Then, it was shown that $M_t \sigma_{t,i_t} / (1 - F^{\sigma_{t,i_t}}(\eta_t \underline{\hat{L}}_{t,i_t}))$ is an unbiased estimator of w_{t,i_t}^{-1} when $G_t = \infty$ (Chen et al., 2025). When G_t is finite, this early stopping introduces bias and thus incurs additional term in the regret. However, with an appropriate choice of G_t, such additional regret term is at most $\log T$. In Appendix D.3, we show that $G_t = K \log t$ is sufficient for Pareto perturbations. Moreover, our analysis can be extended to general Fréchet-type distributions, in contrast to the analysis in Chen et al. (2025, Lemma 8), which relies on properties of the Fréchet distribution.

Algorithm 1 FTPL with conditional geometric resampling II-biased and SPM learning rates

Input : $K \in \mathbb{N}$, $\alpha > 1$, $\beta_1 > 0$, $\hat{L}_1 = 0$.
for $t = 1, 2, \ldots$ **do**
  Sample $r_t = (r_{t,1}, \ldots, r_{t,K})$ i.i.d. from the Pareto distribution with shape α in (2).
  Select $i_t \in \arg\min_{i \in [K]} \{\hat{L}_{t,i} - \beta_t r_{t,i}\}$ and observe ℓ_{t,i_t}. `// FTPL`
  Find $j_t \in \arg\min_i \hat{L}_{t,i}$ and set $M_t := 0$ and $G_t := K \log t$.
  if $i_t = j_t$, $\alpha \in (1,2)$, and $\mathbb{1}[\mathcal{E}_{t,\alpha}] = 1$ in (6) **then** $G_t := 2 \log t$.
  repeat
    $M_t := M_t + 1$. `// CGR II-biased`
    Sample r'_t from the appropriately defined conditional distribution, details in Appendix A.
  until $i_t = \arg\min_{i \in [K]} \{\hat{L}_{t,i} - \beta_t r'_{t,i}\}$ or $M_t \geq G_t$.
  Set $\hat{\ell}_{t,i_t} = M_t \sigma_{t,i_t} \ell_{t,i_t} / (1 - F^{\sigma_{t,i_t}}(\eta_t \underline{\hat{L}}_{t,i_t}))$ and update $\hat{L}_{t+1} = \hat{L}_t + \hat{\ell}_{t,i_t} e_{i_t}$.
  Set β_{t+1} by the update rule of (11) based on z_t, h_t in (9) and q_t in (7).

For $\alpha \in (1,2)$, we additionally introduce an event $\mathcal{E}_{t,\alpha}$, where we further reduce the resampling budget G_t to avoid redundant resampling, defined by

$$\mathcal{E}_{t,\alpha} := \left\{ \sum_{i \neq j_t} \frac{1}{(1 + \eta_t \underline{\hat{L}}_{t,i})^\alpha} < \frac{1}{2} \right\}, \forall \alpha \in (1,2). \tag{6}$$

Simply speaking, this event roughly corresponds to the case where the probability of selecting the current best arm j_t is at least $1/2$. Hence, when both $i_t = j_t$ and $\mathcal{E}_{t,\alpha}$ occur, we reduce the resampling budget G_t from $K \log t$ to $2 \log t$ in order to prevent redundant resampling that can induce artificially large loss estimates. This event also plays an important role in the regret decomposition including $\alpha \geq 2$, where the definition of $\mathcal{E}_{t,\alpha}$ for $\alpha \geq 2$ involves an α-dependent threshold that is slightly larger than $1/2$. The pseudo-code of overall policy is given in Algorithm 1.

3.2. The surrogate probability function

As discussed in Section 2.3, SPM learning rates are designed to balance the stability and penalty terms, which are expressed explicitly as functions of the arm-selection probabilities w_t in the FTRL framework (Zimmert and Seldin, 2021). Accordingly, the update for SPM learning rates β_{t+1} in FTRL naturally relies on w_t (Jin et al., 2023; Ito et al., 2024). In contrast, since w_t of FTPL does not admit a closed form, existing BOBW analyses for FTPL instead express the stability and penalty terms by the cumulative loss estimators $\hat{L}_t$ and learning rates η_t (Honda et al., 2023; Lee et al., 2024). This motivates replacing w_t with suitable surrogate quantities that appear in the corresponding regret bounds. In this paper, we consider the following two surrogate functions, defined for any $i \in [K]$ and $t \in \mathbb{N}$ by

$$p_{t,i} = \min\left(\frac{1}{(1 + \eta_t \underline{\hat{L}}_{t,i})^\alpha}, \frac{1}{\sigma_{t,i}} \right), \text{ and } q_{t,i} = \min\left(\frac{1}{(1 + \eta_{t-1} \underline{\hat{L}}_{t,i})^\alpha}, \frac{1}{\sigma_{t,i}} \right). \tag{7}$$

When η_t is non-increasing, $q_{t,i} \leq p_{t,i}$ always hold by definition, and one can easily show $w_{t,i} \leq p_{t,i}$, whose proof is given in Appendix C.1 for completeness. Therefore, $p_{t,i}$ can be seen as a surrogate probability function. In particular, Lemma 7 in Appendix B shows that the stability and penalty

terms can be bounded from above in terms of $p_{t,i}$. This observation naturally motivates the use of $p_{t,i}$ instead of $w_{t,i}$, where a similar approach was explored in decoupled bandits (Kim et al., 2026).

While this approach is analytically reasonable, directly using p_t to design SPM learning rates as in (5) unfortunately leads to technical difficulties. As discussed in Section 2.3, updating β_{t+1} requires the value of the penalty term at round $t+1$, whose definition depends on $p_{t+1,i}$ and therefore implicitly on β_{t+1} itself. Although this circular dependency can, in principle, be resolved by iterative computation, doing so would introduce unpredictable computational overhead, which loses motivation to use FTPL. In the FTRL framework, circular dependency is resolved by employing hybrid regularizers, which leverages properties of the Bregman divergence and the exact probability w_t. In contrast, since the surrogate $p_{t,i}$ directly depends on the cumulative loss $\hat{L}_t$, it is not straightforward to derive a clean relation between $p_{t+1,i}$ and $p_{t,i}$. To circumvent these difficulties, we instead work with the surrogate q_t, which relies on the previous learning rate. While a detailed discussion of the role of p_t and its limitations is given in Appendix B, the following statement supports the use of q_t instead of p_t.

Lemma 1 (Informal) *For all $i \in [K]$, $q_{t,i} \le p_{t,i} \le 2q_{t,i}$ for certain learning rates η_t.*

Hence, in this paper, q_t will define the actual policy while p_t still serves as an analytical tool.

3.3. Learning rates with the surrogate probability for stability-penalty matching

The expected regret of FTPL with CGR II-biased can be decomposed into two terms, one for the regret by the policy itself and the other from the bias of the estimator (Chen et al., 2025, Lemma 7)

$$\mathrm{Reg}(T) \le \sum_{t=1}^{T} \mathbb{E}\left[\left\langle \hat{\ell}_t, w_t - e_{i^*} \right\rangle\right] + \mathrm{Reg}_{\mathrm{CGR}}(T) := \mathrm{Reg}_{\mathrm{FTPL}}(T) + \mathrm{Reg}_{\mathrm{CGR}}(T),$$

where Appendix D.3 provides the explicit form of $\mathrm{Reg}_{\mathrm{CGR}}(T)$ and shows that it is bounded by $\log T$. The first term is the main regret term, which can be decomposed as (Kim et al., 2026, Lemma 4):

$$\mathrm{Reg}_{\mathrm{FTPL}}(T) \lesssim \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\left\langle \hat{\ell}_t, \phi(\eta_t \hat{L}_t) - \phi(\eta_t \hat{L}_{t+1}) \right\rangle\right]}_{\text{stability term}} + \underbrace{\sum_{t=1}^{T} (\beta_{t+1} - \beta_t)\mathbb{E}\left[r_{t+1,i_{t+1}} - r_{t+1,i^*}\right]}_{\text{penalty term}}. \quad (8)$$

To design SPM learning rates following the idea as in (5), it is crucial to identify appropriate z_t and h_t, which are closely related to the bounds on the stability and penalty terms, respectively. To this end, we define z_t and h_t by

$$z_t = \alpha \sum_{i \neq j_{t+1}} q_{t+1,i}^{1/\alpha}, \text{ and } h_t = \frac{\alpha}{\alpha - 1} \sum_{i \neq j_{t+1}} q_{t+1,i}^{1-\frac{1}{\alpha}}. \quad (9)$$

Recall that, by construction, q_{t+1} depends only on the *previous* learning rate η_t and $\hat{L}_{t+1} = \hat{L}_t + \hat{\ell}_t$, both of which are available at the end of round t. Hence, we index z_t and h_t by t to align with the notation used for FTRL as in (5). To clarify the motivation behind these definitions, we compare them with the case of FTRL with γ-Tsallis entropy in multi-armed bandits. In this setting, Ito et al. (2024) defined

$$z_t^{\mathrm{FTRL}} = \frac{1}{1-\gamma} \sum_{i=1}^{K} \tilde{w}_{t,i}^{1-\gamma}, \text{ and } h_t^{\mathrm{FTRL}} = \frac{1}{\gamma}\left(\sum_{i=1}^{K} w_{t,i}^{\gamma} - 1\right), \quad (10)$$

where $\tilde{w}_{t,i} = \min(w_{t,i}, 1 - w_{t,j_t})$ and j_t denotes the current best arm at round t. These quantities are constructed so that the regret bound can be expressed as in (4), with z_t^{FTRL} and h_{t+1}^{FTRL}. Moreover, Ito et al. (2024) introduced hybrid regularizers to ensure that the arm-selection probability changes smoothly, i.e., $w_{t+1,i} = \mathcal{O}(w_{t,i})$. Therefore, the correspondence between γ-Tsallis entropy and Fréchet-type perturbation with shape $1/(1-\gamma)$ suggests that the analogous quantities would be of the formulation $\alpha w_{t,i}^{1/\alpha}$ or $\alpha w_{t+1,i}^{1/\alpha}$ when defining the FTPL counterpart of z_t^{FTRL}. Here, Lemma 1 implicitly reveals that our constructions of z_t and h_t in (9) satisfy the latter expression, $w_{t+1,i}$, through the dependence of q_{t+1} on $\hat{L}_{t+1}$ and its relation to p_{t+1}.

With these definitions in (7) and (9), the FTPL regret in (8) can be roughly rewritten for any learning rate β_t as

$$\mathrm{Reg}_{\mathrm{FTPL}}(T) \lesssim \mathcal{O}\left(\sum_{t=1}^{T} \frac{z_t}{\beta_t} + (\beta_{t+1} - \beta_t) h_t\right),$$

which is analogous to those of FTRL in (4), with the main difference on the appearance of h_t instead of h_{t+1}. This shift is due to the use of the surrogate $q_{t+1,i}$ in h_t, which depends on the previous learning rate η_t and ensures both z_t and h_t are measurable at the end of round t, thereby avoiding issues realted to h_{t+1}. Based on this decomposition, we design the following adaptive learning rates to equalize the order of the stability and penalty terms, consistent with the idea of SPM:

$$\beta_{t+1} = \begin{cases} \min(2^{\frac{1}{\alpha}}\beta_t, \beta_t + \frac{z_t}{\beta_t h_t}), & \text{if } \alpha \geq 2, \\ \beta_t + \max\left(\frac{z_t}{\beta_t h_t}, \frac{4}{2^{1/\alpha}-1}\frac{1}{t}\right), & \text{if } \alpha \in (1,2), \end{cases} \quad \text{and } \beta_1 \geq 2\alpha K^{\frac{1}{2}-\frac{1}{\alpha}}. \tag{11}$$

While FTRL approaches employed additional regularizer such as log-barrier (Jin et al., 2023) or complement Tsallis entropy (Ito et al., 2024) based on the parameter of Tsallis entropy to obtain generalized BOBW guarantees, we instead incorporate an additional term into the update of learning rates. For $\alpha \in (1,2)$, additional $\Theta(1/t)$ term is introduced to ensure $\beta_t = \Omega(\log t)$, which is required for our analysis to address extreme cases where the resampling procedure is repeated excessively even when w_{t,i_t} is sufficiently large. Combined with the budget of resampling steps $G_t = 2\log t$ in CGR II with $\mathcal{E}_{t,\alpha}$ in (6), this choice simplifies the analysis. For $\alpha \geq 2$, we restrict the learning rate updates so that β_t cannot increase too rapidly between rounds, which guarantees Lemma 1.

In particular, with these definitions, we obtain the following results, which provide the same structure to those for FTRL with SPM learning rates in (5).

Lemma 2 *With β_t in (11) and h_t, z_t defined in (9), Algorithm 1 with $\alpha > 1$ satisfies*

$$\mathrm{Reg}_{\mathrm{FTPL}}(T) \leq \begin{cases} \sum_{t=1}^{T-1} \mathcal{O}\left(\frac{z_t}{\beta_t}\right) + t_0(\alpha, K) + \mathcal{O}\left(\frac{\alpha^3}{(\alpha-1)^2}\sqrt{K}\right), & \text{if } \alpha \geq 2, \\ \sum_{t=1}^{T-1} \mathcal{O}\left(\frac{z_t}{\beta_t}\right) + \mathcal{O}\left(\frac{\alpha^2 K^{1/\alpha}}{(\alpha-1)}\log T\right) + \mathcal{O}\left(\frac{\alpha^3}{(\alpha-1)^2}\sqrt{K}\right) + 2, & \text{if } \alpha \in (1,2), \end{cases}$$

where $t_0(\alpha, K) = \mathcal{O}(\alpha^2 K^2 \log^2(\alpha K))$.

Although Lemma 2 includes $t_0(\alpha, K)$ term for $\alpha \geq 2$, which depends on α and K, the dependence on K can be eliminated. Specifically, when $i_t = j_t$ and w_{t,i_t} is sufficiently large, setting $G_t = \Theta(\log t)$ instead of $K \log t$ can eliminate such dependency. This condition can be verified in the same way as the case $\alpha \in (1,2)$ by introducing suitable events $\mathcal{E}_{t,\alpha}$ as in (6).

Therefore, the main leading term of the regret upper bound becomes the first term $\sum_t z_t/\beta_t$. By appropriately adapting the arguments in Lemmas 9 and 10 of Ito et al. (2024), we obtain the following lemma.

Lemma 3 *For Algorithm 1 with $\alpha > 1$ and β_t defined in (11), it holds that*

$$\sum_{t=1}^{T} \frac{z_t}{\beta_t} \leq \mathcal{O}\left(\min\left\{ \sqrt{\log T \sum_{t=1}^{T} h_t z_t} + \sqrt{h_{\max} z_{\max}}, \sqrt{h_{\max} \sum_{t=1}^{T} z_t} \right\} \right) + \mathcal{O}\left(\frac{\alpha z_{\max}}{\beta_1} \right),$$

where $z_{\max}$ and $h_{\max}$ satisfy $z_t \leq z_{\max}$ and $h_t \leq h_{\max}$ for all $t \in \mathbb{N}$.

Therefore, in the adversarial regime, it is sufficient to provide the bound of $h_{\max} \sum_t z_t$. For the adversarial regime with (Δ, C, T) self-bounding constraint, the key step is to show $h_t z_t \leq \omega(\Delta) \cdot \langle w_{t+1}, \Delta \rangle$ for some constants $\omega(\Delta)$, although we need to control additional terms due to the use of surrogate probabilities.

Theorem 4 *For the K-armed bandit problem, Algorithm 1 with β_t in (11), and any $\alpha > 1$ achieves the following bounds simultaneously. In the adversarial regime, we have*

$$\mathrm{Reg}(T) \leq \mathcal{O}\left(\frac{\alpha^2}{\alpha - 1} \sqrt{KT} \right).$$

In the adversarial regime with (Δ, C, T) self-bounding constraint, we have

$$\mathrm{Reg}(T) \leq \mathcal{O}\left(\omega(\Delta) \log T + \sqrt{C \omega(\Delta) \log T} + c(\alpha, K) \right),$$

where $\omega(\Delta) = \mathcal{O}\left(\frac{\alpha^4}{(\alpha-1)^2} \frac{K}{\Delta_{\min}} \right)$ and $c(\alpha, K)$ denotes the constant depending on α, K.

To the best of our knowledge, this is the first result establishing BOBW guarantees for FTPL with Fréchet-type perturbations for general $\alpha > 1$. While our analysis recovers the adversarial results of Lee et al. (2024), one limitation is that it does not recover the optimal gap-dependent bound $\omega(\Delta) = \mathcal{O}(\sum_{i \neq i^*} 1/\Delta_i)$ when $\alpha = 2$, in contrast to the results of Ito et al. (2024).

This gap is due to the use of surrogate probabilities p_t, q_t, which, unlike w_t, do not necessarily lie in the probability simplex. The suboptimal dependence on Δ comes from the worst case where we cannot provide a lower bound on w_t in terms of p_t without the dependence on K. A representative example is when $\underline{\hat{L}}_{t,i} = 0$ for all i, where $w_{t,i} = 1/K$ but $p_{t,i} = 1/\sigma_{t,i}$ for all i. Although it is possible to recover the optimal Δ dependence by allowing such K-dependent bounds, doing so would introduce additional K term in the current bounds. Therefore, there is room for improvement both in the choice of surrogate probabilities and in the current BOBW analysis of FTPL. In particular, it may be possible to avoid the use of overly conservative surrogates.

To compare our results with those obtained under explicit probability computation in the FTRL framework, we recall the known relationship between γ-Tsallis entropy and Fréchet-type perturbations with shape α, where the correspondence $\alpha \approx 1/(1-\gamma)$ has been observed (Kim and Tewari, 2019; Lee et al., 2025). Under this correspondence, the quantity $\omega(\Delta)$ in Ito et al. (2024) can be seen as a bound of order $\mathcal{O}(\frac{\alpha^2 K}{(\alpha-1)\Delta_{\min}})$, which shows a more favorable dependence on α than our result. Such factor is again due to the use of surrogate probabilities in (7), where the worst case summations such as $\sum_{n=2}^{K} n^{-1/\alpha}$ introduce additional α-dependent terms that do not appear when working directly with the probability vectors on the simplex.

From a computational perspective, our approach can be more efficient than FTRL methods, especially those with hybrid regularizers (Jin et al., 2023; Ito et al., 2024) that cannot be reduced

Algorithm 2 FTPL with geometric resampling and SPM learning rates for contextual bandits

Input : $K, N \in \mathbb{N}$, $\alpha \geq 2$, $\beta_1 > 0$, $\hat{L}_1 = 0$.
for $t = 1, 2, \dots$ **do**
  Sample $r_t = (r_{t,1}, \dots, r_{t,K})$ i.i.d. from the Pareto distribution with shape α in (2).
  Select $i_t \in \arg\min_{i\in[K]}\{\hat{L}_{t,i} - \beta_t r_{t,i}\}$ and observe advices $\{\pi_{t,i}\}_{i\in[K]}$
  Select $a_t \sim \pi_{t,i_t}$, observe ℓ_{t,i_t}, and set $M_t := 0$.
  repeat
    $M_t := M_t + 1$. `// Geometric resampling`
    Sample $r'_t = (r'_{t,1}, \dots, r'_{t,K})$ i.i.d. from the Pareto distribution with shape α in (2).
    Select $i' \in \arg\min_{i\in[K]}\{\hat{L}_{t,i} - \beta_t r'_{t,i}\}$ and $a' \sim \pi_{t,i'}$.
  until $a_t = a'$.
  Set $\hat{\ell}_{t,i} = M_t \pi_{t,i,a_t} \ell_{t,a_t}$ for all $i \in [K]$ and update $\hat{L}_{t+1} = \hat{L}_t + \hat{\ell}_t$.
  Set β_{t+1} by the update rule of (11) based on z_t, h_t in (12) and q_t in (7).

to one-dimensional optimizations. In particular, Algorithm 1 follows the standard FTPL with i.i.d. Pareto perturbations, where the only difference is in the design of learning rates. The dominant computational complexity is from CGR II and sorting $\hat{L}_t$, leading to an average complexity of $\mathcal{O}(K \log K)$. Indeed, Chen et al. (2025) showed that FTPL with CGR II runs faster than FTRL with Tsallis entropy, even though it can be computed efficiently. By avoiding both hybrid regularization and convex optimization, and using CGR II-biased algorithm, our policy remains computationally efficient while preserving the desired regret guarantees in terms of K and T for general $\alpha > 1$.

4. Application of SPM learning rates for FTPL in bandit problems with expert advices

In this section, we extend the SPM learning rates for FTPL to the bandits with expert advice, also known as contextual bandit settings, where there are K experts and N arms (Auer et al., 2002; Dann et al., 2023). At each round, expert $i \in [K]$ provides an advice distribution $\pi_{t,i} \in \mathcal{P}_N$ over N arms, where $\mathcal{P}_N$ denotes the $(N-1)$-dimensional probability simplex. The agent selects an expert $i_t \in [K]$ according to FTPL rule in (1), observes the advice from all the expert $\{\pi_{t,i}\}_i$, and then selects $a_t \in [N]$ following distribution π_{t,i_t}. Let $\tilde{\ell}_{t,i} = \mathbb{E}[\sum_{a=1}^K \pi_{t,i,a}\ell_{t,a}]$ denote the expected loss of the expert i at round t. Then, the regret with respect to the best experts i^* is given by

$$\mathrm{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^T\sum_{a=1}^N \pi_{t,i_t,a}\ell_{t,a}\right] - \min_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^T\sum_{a=1}^N \pi_{t,i,a}\ell_{t,a}\right] = \mathbb{E}\left[\sum_{t=1}^T \tilde{\ell}_{t,i_t}\right] - \mathbb{E}\left[\sum_{t=1}^T \tilde{\ell}_{t,i^*}\right].$$

Let $w_{t,i}$ be the probability that FTPL rule in (1) selects an expert i for given $\hat{L}_t$ in contextual bandits and $P_t \in \mathcal{P}_N$ denote the marginal distribution over arms given the expert-selection probability w_t and the advice $\{\pi_{t,i}\}_i$, i.e., $P_{t,a} = \sum_{i=1}^K w_{t,i}\pi_{t,i,a}$. Since we observe only the loss from the selected arm a_t, one can consider the IW estimator of $\tilde{\ell}_{t,i}$, given by $\ell_{t,a_t}\pi_{t,i,a_t}/P_{t,a_t}$. Note that, although we choose only one expert i_t, the loss estimators can be constructed for all experts since we observe π_{t,i,a_t} from all experts i, in contrast to multi-armed bandits.

Since P_{t,a_t} cannot be computed in the general FTPL framework, we consider the original geometric resampling, which samples i'_t with resampled perturbations and $a'_t \sim \pi_{t,i,i'_t}$ until the same

a_t is selected. Then, we construct an unbiased estimator $\hat{\ell}_{t,i} = M_t \ell_{t,a_t} \pi_{t,i,a_t}$, where M_t denotes the number of resampling steps at round t. The policy for contextual bandit is given in Algorithm 2.

In general, the same regret decomposition can be obtained as the standard multi-armed bandits given in (8). The main difference arises in the analysis of the stability term due to the difference in estimator $\hat{\ell}_t$. Nevertheless, most of the techniques developed in Section 3, as well as those from previous BOBW analysis can extend to contextual bandits with minor modifications. For the sake of analysis, we impose a mild assumption on the advice distributions $\pi_{t,i,a}$, as follows.

Assumption 1 *For all $t \in \mathbb{N}$, $i \in [K]$, and $a \in [N]$, $\pi_{t,i,a} \in [\nu, 1] \cup \{0\}$, i.e., the advice probability is uniformly bounded from below by some constants $\nu \in (0, 1/N)$ if it is not zero.*

Note that Algorithm 2 does not require the information of ν, which implies that Assumption 1 is used only in the analysis. For the stability term, we have the following results, which corresponds to Lemma 7 for the standard multi-armed bandits in Appendix B.

Lemma 5 *For any $t \in \mathbb{N}$, Algorithm 2 with $\alpha \geq 2$ satisfies that*

$$\mathbb{E}\left[\left\langle \hat{\ell}_t, \phi(\eta_t \hat{L}_t) - \phi(\eta_t \hat{L}_{t+1}) \right\rangle \middle| \hat{L}_t\right] \leq \mathcal{O}\left(\frac{\alpha N}{\beta_t} \max_{i \neq j_t} p_{t,i}^{1/\alpha}\right) + g_t(\alpha; \nu),$$

where p_t is in (7) and $g_t(\alpha)$ is a function satisfying $\sum_t g_t(\alpha; \nu) = \mathcal{O}(\alpha^2/\nu)$ if Assumption 1 holds.

Based on this observation, we adopt the learning rate β_t defined in (11) with following z_t and h_t:

$$z_t = \alpha N \max_{i \neq j_{t+1}} q_{t+1,i}^{1/\alpha}, \text{ and } h_t = \frac{\alpha}{\alpha - 1} \sum_{i \neq j_{t+1}} q_{t+1,i}^{1-1/\alpha}, \tag{12}$$

where h_t coincides with that of multi-armed bandits in (9). This choice is natural, as the penalty term remains unchanged. With these choices, Algorithm 2 obtains the following BOBW guarantee.

Theorem 6 *For contextual bandits of N arms with K experts under Assumption 1, Algorithm 2 with β_t in (11) with h_t, z_t in (12) and $\beta_1 = \mathcal{O}(\alpha N)$, and any $\alpha \geq 2$ achieves the following bounds simultaneously. In the adversarial regime, we have*

$$\mathrm{Reg}(T) \leq \mathcal{O}\left(\sqrt{\frac{\alpha^3}{\alpha - 1} N K^{1/\alpha} T}\right).$$

In the adversarial regime with a (Δ, C, T) self-bounding constraint, we have

$$\mathrm{Reg}(T) \leq \mathcal{O}\left(\omega(\Delta) \log T + \sqrt{C \omega(\Delta) \log T} + \sqrt{\frac{\alpha^3 N K^{1/\alpha}}{\alpha - 1}} + \alpha^2/\nu\right),$$

where $\omega(\Delta) = \mathcal{O}\left(\frac{\alpha^3}{\alpha - 1} \frac{N K^{1/\alpha}}{\Delta_{\min}}\right)$.

While our results match those of Ito et al. (2024) in terms of N, K and T, the additional $\sqrt{\alpha}$ factor in our bounds introduces an extra $\log K$ dependence when α is set to minimize $\alpha^2 K^{1/\alpha}$ term. Specifically, setting $\alpha = \Theta(\log K)$ provides an adversarial regret of $\mathcal{O}(\sqrt{NT \log^2 K})$ and a regret of $\mathcal{O}(N \log^2 K \log T / \Delta_{\min})$ under self-bounding constraints. In contrast, the corresponding bounds

in Dann et al. (2023) and Ito et al. (2024), which use explicit probabilities, are $\mathcal{O}(\sqrt{NT \log K})$ in the adversarial regime and $\mathcal{O}(N \log K \log T / \Delta_{\min})$ for adversarial regime with self-bounding constraints. These results show one limitation of surrogate-based approaches, especially when the shape parameter, the parameter of the distribution, is determined by the problem-dependent constants.

In the contextual bandit setting, our bounds do not improve upon the best previously known BOBW guarantees. Nevertheless, a key strength of our approach lies in its generality through the use of SPM learning rates: both Algorithm 1 for multi-armed bandits and Algorithm 2 for contextual bandits share the same selection rule including perturbation distributions, differing only in the loss estimation and update of learning rate, which naturally reflects the differences between the problem settings.[1] Moreover, our analysis expresses regret in terms of surrogate quantities p_t, q_t for both settings, making the framework largely agnostic to the specific bandit model, as achieved in FTRL framework (Ito et al., 2024; Nguyen et al., 2025). Therefore, we expect the similar analytical approach can be extended beyond the classical bandit setting, where similar regret decompositions and bounds can be obtained with minor modifications.

5. Conclusion and future work

We proposed adaptive learning rates for FTPL based on the SPM principle, which was originally developed for the FTRL framework with explicit arm-selection probabilities. We showed that the SPM methodology can be extended to FTPL by appropriately choosing surrogate probabilities, and established the BOBW guarantees for FTPL with general perturbation parameters both in standard multi-armed bandits and in bandit problems with expert advice. Although our analysis is grounded in the FTPL framework, we expect that our approach, the use of surrogate probabilities, could be applicable to other algorithmic frameworks as well. In particular, surrogate probabilities may offer a way to replace explicit probability computations in settings where FTRL-based methods are applicable, potentially leading to more computationally efficient alternatives.

These observations open several directions for future work on efficient BOBW policies in settings not covered in this paper, such as graph bandits, linear bandits, partial monitoring, and bandits with heavy-tailed losses. In the heavy-tailed setting, for example, surrogate probabilities may be used in place of arm-selection probabilities to define adaptive threshold values for loss estimators (Huang et al., 2022). Recently, Zhao et al. (2025) further employed both the adaptive thresholding and the SPM learning rate in the FTRL framework to obtain the BOBW guarantee in heavy-tailed linear bandits. However, handling negative losses introduces additional challenges in the current BOBW analysis of FTPL, since large negative loss can make a previously suboptimal arm to become the best arm after being selected, a case that does not occur in settings with nonnegative losses.

In addition, while some FTRL policies employ arm-dependent learning rates (Jin et al., 2023; Nguyen et al., 2025), it remains unclear whether such designs can be extended to FTPL. From the FTPL perspective, arm-dependent learning rates correspond to FTPL with arm-dependent perturbation distributions and arm-independent learning rates, that is, FTPL with non-i.i.d. perturbations. Clarifying whether comparable guarantees, especially in the analysis of the stability term, can be obtained in this setting is an interesting open question.

1. While Algorithm 1 employs CGR II-biased algorithm, the similar regret bounds can be obtained by using the original GR when $\alpha \geq 2$. This observation indicates that Algorithms 1 and 2 indeed share the same algorithmic structure.

Acknowledgments

JL was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00395303). JH was supported by JSPS KAKENHI Grant Number JP25K02232. SI was supported by JSPS KAKENHI Grant Number JP25K03184 and by JST PRESTO, Japan, Grant Number JPMJPR2511. CK was supported by the grant Nos. 2024-00460980; and 2025-02304717 (IITP) funded by the Korea government (the Ministry of Science and ICT).

References

Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. Perturbation techniques in online learning and optimization. *Perturbations, Optimization, and Statistics*, 233, 2016.

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

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42.1–42.23. PMLR, 2012.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007.

Botao Chen, Jongyeong Lee, and Junya Honda. Geometric resampling in nearly linear time for follow-the-perturbed-leader with best-of-both-worlds guarantee in bandit problems. In *International Conference on Machine Learning*, PMLR, pages 8403–8426, 2025.

Botao Chen, Jongyeong Lee, Chansoo Kim, and Junya Honda. A further efficient algorithm with best-of-both-worlds guarantees for m-set semi-bandit problem. *arXiv preprint arXiv:2603.11764*, 2026.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *Conference on Learning Theory*, pages 5503–5570. PMLR, 2023.

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

Junya Honda, Shinji Ito, and Taira Tsuchiya. Follow-the-Perturbed-Leader achieves best-of-both-worlds for bandit problems. In *International Conference on Algorithmic Learning Theory*, volume 201, pages 726–754. PMLR, 2023.

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

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

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adaptive learning rate for follow-the-regularized-leader: Competitive analysis and best-of-both-worlds. In *Conference on Learning Theory*. PMLR, 2024.

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*, 2023.

Adam Kalai and Santosh S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71:291–307, 2005.

Michael N Katehakis and Herbert Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences*, 92(19):8584–8585, 1995.

Baekjin Kim and Ambuj Tewari. On the optimality of perturbations in stochastic and adversarial multi-armed bandit problems. In *Advances in Neural Information Processing Systems*, pages 2695–2704, 2019.

Chaiwon Kim, Jongyeong Lee, and Min-hwan Oh. Follow-the-perturbed-leader for decoupled bandits: Best-of-both-worlds and practicality. *International Conference on Machine Learning*, 2026.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

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

Jongyeong Lee, Junya Honda, Shinji Ito, and Min-hwan Oh. Follow-the-Perturbed-Leader with Fréchet-type tail distributions: Optimality in adversarial bandits and best-of-both-worlds. In *Conference on Learning Theory*, pages 3375–3430. PMLR, 2024.

Jongyeong Lee, Junya Honda, Shinji Ito, and Min-hwan Oh. Revisiting follow-the-perturbed-leader with unbounded perturbations in bandit problems. In *Advances in Neural Information Processing Systems*, volume 38, pages 83623–83675. Curran Associates, Inc., 2025.

Mengmeng Li, Daniel Kuhn, and Bahar Taşkesen. Optimism in the face of ambiguity principle for multi-armed bandits. *arXiv preprint arXiv:2409.20440*, 2024.

Gergely Neu and Gábor Bartók. Importance weighting without importance weights: An efficient algorithm for combinatorial semi-bandits. *Journal of Machine Learning Research*, 17(154):1–21, 2016.

Quan Nguyen, Shinji Ito, Junpei Komiyama, and Mehta Nishant. Data-dependent bounds with T-optimal best-of-both-worlds guarantees in multi-armed bandits using stability-penalty matching. In *Conference on Learning Theory*, pages 4386–4451. PMLR, 2025.

Frank WJ Olver, Daniel W Lozier, Ronald F Boisvert, and Charles W Clark. *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.

Jan Poland. FPL analysis for adaptive bandits. In *International Symposium on Stochastic Algorithms*, pages 58–69. Springer, 2005.

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

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 1743–1759. PMLR, 2017.

Taira Tsuchiya and Shinji Ito. 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. *Advances in Neural Information Processing Systems*, 37:8477–8514, 2024.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Further adaptive best-of-both-worlds algorithm for combinatorial semi-bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 8117–8144. PMLR, 2023a.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds. *Advances in Neural Information Processing Systems*, pages 47406–47437, 2023b.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference on Learning Theory*, pages 1–29. PMLR, 2018.

Jingxin Zhan, Yuchen Xin, Chenjie Sun, and Zhihua Zhang. Follow-the-perturbed-leader nearly achieves best-of-both-worlds for the m-set semi-bandit problems. In *Advances in Neural Information Processing Systems*, 2025.

Canzhe Zhao, Shinji Ito, and Shuai Li. Heavy-tailed linear bandits: Adversarial robustness, best-of-both-worlds, and beyond. *arXiv preprint arXiv:2508.13679*, 2025.

Julian Zimmert and Teodor Vanislavov Marinov. Productive bandits: Importance weighting no more. *Advances in Neural Information Processing Systems*, 37:85360–85388, 2024.

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

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

Appendix A. Details on conditional geometric resampling

Here, we provide more details on CGR II. The main idea of CGR is not to generate new perturbations that clearly violates the termination condition. To do this, it restricts the distributions of perturbations in the resampling steps by considering certain necessary condition $\mathcal{A}_t$. In CGR II, we used the following condition

$$\mathcal{A}_t = \left\{ r'_{t,i_t} = \max_{i:\sigma_{t,i} \leq \sigma_{t,i_t}} r'_{t,i},\, r'_{t,i_t} \geq \eta_t \underline{\hat{L}}_{t,i_t} \right\}, \tag{13}$$

where we force the new perturbation from the actually selected arm at round t to be larger than those of arms with smaller cumulative loss estimates and also to be larger than $\eta_t \underline{\hat{L}}_{t,i_t}$. This is equivalent

Algorithm 3 FTPL with conditional geometric resampling II-biased and SPM learning rates

Input : $K \in \mathbb{N}$, $\alpha > 1$, $\beta_1 > 0$, $\hat{L}_1 = 0$.
for $t = 1, 2, \dots$ **do**
  Sample $r_t = (r_{t,1}, \dots, r_{t,K})$ i.i.d. from the Pareto distribution with shape α in (2).
  Select $i_t \in \arg\min_{i \in [K]} \{\hat{L}_{t,i} - \beta_t r_{t,i}\}$ and observe ℓ_{t,i_t}.
  Find $j_t = \arg\min_i \hat{L}_{t,i}$ and set $M_t := 0$, and $G_t = K \log t$.
  if $i_t = j_t$, $\alpha \in (1,2)$, and $\mathbb{1}[\mathcal{E}_{t,\alpha}] = 1$ in (6) **then** $G_t = 2 \log t$.
  repeat
    $M_t := M_t + 1$. `// CGR II-biased`
    Sample r'_{t,i_t} from truncated distribution over $[\eta_t \underline{\hat{L}}_{t,i_t}, \infty)$ as in (14).
    Sample $\{r_{t,i} : \sigma_{t,i} < \sigma_{t,i_t}\}$ i.i.d. from truncated distribution over $[0, r'_{t,i_t}]$ as in (15).
    Sample $\{r_{t,i} : \sigma_{t,i} > \sigma_{t,i_t}\}$ i.i.d. from Pareto with shape α in (2).
  until $i_t = \arg\min_{i \in [K]} \{\hat{L}_{t,i} - \beta_t r'_i\}$ or $M_t \geq G_t$
  Set $\hat{\ell}_{t,i_t} = M_t \sigma_{t,i_t} / (1 - F^{\sigma_{t,i_t}}(\eta_t \underline{\hat{L}}_{t,i_t}))$ and update $\hat{L}_{t+1} = \hat{L}_t + \hat{\ell}_t$.
  Set β_{t+1} by the update rule of (11) based on z_t, h_t in (9) and q_t in (7).
end

to sample r'_{t,i_t} with the distribution whose distribution function is

$$F_{i_t}(x; \eta_t \underline{\hat{L}}_{t,i_t}) = \frac{F^{\sigma_{t,i_t}}(x) - F^{\sigma_{t,i_t}}(\eta_t \underline{\hat{L}}_{t,i_t})}{1 - F^{\sigma_{t,i_t}}(\eta_t \underline{\hat{L}}_{t,i_t})}, \; x \geq \eta_t \underline{\hat{L}}_{t,i_t}. \tag{14}$$

After sampling r'_{t,i_t}, CGR II samples the perturbations for $i \in \{j : \sigma_{t,j} < \sigma_{t,i_t}\}$ i.i.d. from the truncated distribution over $[0, r'_{t,i_t})$ whose distribution function is given by

$$F_i(x; r'_{t,i_t}) = F(x)/F(r'_{t,i_t}), \; x \in [0, r'_{t,i_t}]. \tag{15}$$

For the remaining arms, $\{j : \sigma_{t,j} > \sigma_{t,j}\}$, the perturbations are sampled independently from the original perturbation distribution. Then, Algorithm 1 can be explicitly written as Algorithm 3.

Appendix B. Intuition and difficulties behind the surrogate-based learning rates

Here, we discuss the intuition behind on the choice of $q_{t,i}$, and definitions of z_t and h_t. In the standard multi-armed bandits, we can obtain the following bounds by applying the results of Lee et al. (2024).

Lemma 7 *For any $t \in \mathbb{N}$, FTPL with Pareto distribution with $\alpha > 1$ and monotonically decreasing learning rates $\eta_t = 1/\beta_t$ satisfies that for any $i \in [K]$*

$$\mathbb{E}\left[\hat{\ell}_{t,i}(\phi_i(\eta_t \hat{L}_t) - \phi_i(\eta_t(\hat{L}_t + \hat{\ell}_{t,i} e_i)) \middle| \hat{L}_t\right] \leq \frac{2e\alpha}{\beta_t} p_{t,i}^{\frac{1}{\alpha}},$$
$$\mathbb{E}[\mathbb{1}[i_t = i] r_{t,i}] \leq \frac{2\alpha}{\alpha - 1} p_{t,i}^{1 - \frac{1}{\alpha}}.$$

This lemma implies that we can rewrite the regret of FTPL in (8) in terms of $p_{t,i}$ as follows:

$$\mathrm{Reg}_{\mathrm{FTPL}}(T) \lesssim \mathcal{O}\left(\sum_{t=1}^{T} \frac{\sum_{i=1}^{K} p_{t,i}^{\frac{1}{\alpha}}}{\beta_t} + \sum_{t=1}^{T} (\beta_{t+1} - \beta_t) \sum_{i \neq i^*} p_{t+1,i}^{1 - \frac{1}{\alpha}} \right). \tag{16}$$

Since i^* is unknown, one may recover a formulation analogous to (4) by setting z_t and h_t by

$$z'_t = \alpha \sum_{i=1}^{K} p_{t,i}^{1/\alpha} \text{ and } h'_t = \frac{\alpha}{\alpha - 1} \sum_{i=1}^{K} p_{t,i}^{1-1/\alpha}, \tag{17}$$

which have a structure very similar to those of FTRL in (10). However, directly using these quantities to design SPM learning rates as in (5) incurs two technical difficulties. Firstly, as briefly discussed in Section 2.3, updating β_{t+1} requires the value of h'_{t+1}, whose definition depends on $p_{t+1,i}$ and hence on β_{t+1}. In principle, one can compute such β_{t+1} even with this $p_{t+1,i}$ by iteratively solving this circular dependence. However, doing so would introduce undesirable computational cost, which loses the motivation to use FTPL. To justify the use of h'_{t+1}, one therefore need to show $p_{t+1,i} = \mathcal{O}(p_{t,i})$ and then replace h'_{t+1} with h'_t as done in FTRL (Ito et al., 2024; Nguyen et al., 2025). Secondly, the definitions of z'_t and h'_t involve summation over all arms, which introduces an additional obstacle when relating these quantities to w_t in adversarial regime with self-bounding constraints. In particular, p_{t,j_t} is always 1 by the construction, whereas, in the FTRL framework, the corresponding term can be replaced with $\min(1 - w_{t,j_t}, w_{t,j_t})$ term (Ito et al., 2024).

To address the difficulties above, our definitions of z_t and h_t in (9) allocate the contribution of the current best arm j_{t+1} into the summation over the other arms as well as it uses q_t that are computable without access to β_{t+1}. This allow us to construct an update rule for β_{t+1} that depends only on quantities available at the round t. Nevertheless, to justify this construction, we require the following relationship between $q_{t,i}$ and $p_{t,i}$.

Lemma 8 (Formal version of Lemma 1) *It holds that $q_{t,i} \leq p_{t,i} \leq 2q_{t,i}$ for $t \geq 3$ when $\alpha \in (1, 2)$, and for $t \in \mathbb{N}$ when $\alpha \geq 2$.*

Lemma 8 implies that we can still utilize the results of Lemma 7 in terms of $q_{t,i}$. Moreover, it clearly shows a correspondence between z'_{t+1}, h'_{t+1} in (17), which are not efficiently computable at round t, and z_t, h_t in (9), which can be easily computed at the end of round t. Therefore, the regret upper bound in (16) can be roughly expressed as follows, where we ignore the term related to j_t in h'_t, z'_t and some constants for the purpose of illustration.

$$\begin{aligned}
\sum_{t=1}^{T} \frac{z'_t}{\beta_t} + (\beta_{t+1} - \beta_t) h'_{t+1} &= \sum_{t=1}^{T-1} \left(\frac{z'_{t+1}}{\beta_{t+1}} + (\beta_{t+1} - \beta_t) h'_{t+1} \right) + \frac{z'_1}{\beta_1} + (\beta_{T+1} - \beta_T) h'_{T+1} \\
&\lesssim \sum_{t=1}^{T-1} \left(\frac{2^{\frac{1}{\alpha}} z_t}{\beta_{t+1}} + 2^{1-\frac{1}{\alpha}} (\beta_{t+1} - \beta_t) h_t \right) + \frac{z'_1}{\beta_1} + \frac{z_T}{\beta_T h_T} h'_{T+1} \\
&\lesssim \sum_{t=1}^{T-1} \left(\frac{2^{\frac{1}{\alpha}} z_t}{\beta_t} + 2^{1-\frac{1}{\alpha}} (\beta_{t+1} - \beta_t) h_t \right) \approx \sum_{t=1}^{T-1} 2 \frac{z_t}{\beta_t},
\end{aligned}$$

which roughly recovers the formulation in Lemma 2. The second line follows from the definition of β_t in (11). In third line, we ignore constant terms since the last term becomes negligible for large T due to $\beta_T = \Omega(\log T)$. While it may be possible to avoid the above difficulties directly using p_t, we find it technically more complicated due to the use of surrogate, whereas our construction leads to a considerably simpler and more convenient analysis.

Appendix C. Proofs of Lemmas in multi-armed bandits

In this section, we provide the proofs omitted in Section 3.

C.1. Proof on the relationship between surrogate probability and arm-selection probability

Here, we show that

$$w_{t,i} \leq p_{t,i},\ \forall i \in [K],\ t \in \mathbb{N}.$$

Proof Since FTPL plays an arm according to (1), where all the perturbations are generated independently from the identical distribution, it is clear that the arm with the smallest cumulative loss $\hat{L}_{t,\cdot}$ will be of the highest probability to be selected. When $\hat{L}_{t,i}$ is $\sigma_{t,i}$-th smallest, then its arm-selection probability should be smaller than $1/\sigma_{t,i}$ since there exist $\sigma_{t,i}-1$ arms with smaller cumulative loss.

The left $\hat{L}_t$ dependent bounds can be directly obtained by the definition of $w_{t,i}$ in (3) as follows.

$$\begin{aligned} w_{t,i} &= \int_0^\infty \frac{\alpha}{(z+\eta_t \underline{\hat{L}}_{t,i}+1)^{\alpha+1}} \prod_{j\neq i}\left(1-\frac{1}{(z+\eta_t \underline{\hat{L}}_{t,i}+1)^{\alpha}}\right)\mathrm{d}z \\ &\leq \int_0^\infty \frac{\alpha}{(z+\eta_t \underline{\hat{L}}_{t,i}+1)^{\alpha+1}}\mathrm{d}z = \frac{1}{(1+\eta_t \underline{\hat{L}}_{t,i})^{\alpha}}, \end{aligned}$$

which concludes the proof. ■

C.2. Proof of Lemma 7

While the overall proofs follow the results in previous FTPL analysis with Pareto perturbation (Honda et al., 2023; Lee et al., 2024; Kim et al., 2026), we provide the detailed proofs since we use the slightly tighter results than their presented results.

Proof Let us start from proving the first results, which is the bound for the stability term.

Stability analysis. For any $\lambda \in \mathbb{R}^K$, define

$$\begin{aligned} \phi_i'(\lambda) := \frac{\partial \phi_i(\lambda)}{\partial \lambda_i} &= \int_0^\infty f'(z+\underline{\lambda}_i)\prod_{j\neq i}F(z+\underline{\lambda}_j)\mathrm{d}z \\ &= \int_0^\infty \frac{-\alpha(\alpha+1)}{(z+\underline{\lambda}_i+1)^{\alpha+2}}\prod_{j\neq i}\left(1-\frac{1}{(z+\underline{\lambda}_i+1)^{\alpha}}\right)\mathrm{d}z. \end{aligned} \tag{18}$$

Then, by definition, we can obtain

$$\begin{aligned} \phi_i(\eta_t\hat{L}_t) - \phi_i(\eta_t(\hat{L}_t+\hat{\ell}_{t,i}e_i) &\leq \int_0^{\eta_t\hat{\ell}_{t,i}} -\phi_i'(\eta_t\hat{L}_t+xe_i)\mathrm{d}x \\ &\leq \int_0^{\eta_t\hat{\ell}_{t,i}} -\phi_i'(\eta_t\hat{L}_t)\mathrm{d}x \qquad (\because \text{decreasing w.r.t. } \lambda_i \text{ by (18)}) \\ &\leq -\eta_t\hat{\ell}_{t,i}\phi_i'(\eta_t\hat{L}_t). \end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{\ell}_{t,i}(\phi_i(\eta_t\hat{L}_t) - \phi_i(\eta_t(\hat{L}_t + \hat{\ell}_{t,i}e_i))\middle|\hat{L}_t\right] &\le \mathbb{E}\left[-\eta_t\hat{\ell}_{t,i}^2\phi_i'(\eta_t\hat{L}_t)\middle|\hat{L}_t\right] && (19)\\
&= \mathbb{E}\left[-\mathbb{1}[i_t = i]\eta_t\ell_{t,i}^2 M_t^2\phi_i'(\eta_t\hat{L}_t)\middle|\hat{L}_t\right]\\
&\le \mathbb{E}\left[-\mathbb{1}[i_t = i]\eta_t\ell_{t,i}^2\frac{2}{w_{t,i}^2}\phi_i'(\eta_t\hat{L}_t)\middle|\hat{L}_t\right] && (20)\\
&\le \mathbb{E}\left[\eta_t\frac{-2\phi_i'(\eta_t\hat{L}_t)}{w_{t,i}}\middle|\hat{L}_t\right], && (\because \ell_t \in [0,1]^K)
\end{aligned}$$

where (20) follows from $\mathbb{E}[M_t^2|\hat{L}_t, i_t] \le 2/w_{t,i_t}^2$. Here, Lemma 9 in Lee et al. (2024) shows that $-\phi_i'(\lambda)/\phi_i(\lambda)$ is monotonically increasing with respect to λ_j for any $j \neq i$. Therefore, we have

$$\frac{-\phi_i'(\eta_t\hat{L}_t)}{\phi_i(\eta_t\hat{L}_t)} \le \frac{-\phi_i'(\eta_t L^*)}{\phi_i(\eta_t L^*)}, \text{ where } L_j^* = \begin{cases} L_i, & \text{if } \sigma_{t,j} \le \sigma_{t,i},\\ \infty, & \text{if } \sigma_{t,j} > \sigma_{t,i}.\end{cases}$$

By definition of L^*, we have

$$\begin{aligned}
\frac{-\phi_i'(\eta_t L^*)}{\phi_i(\eta_t L^*)} &= \frac{\int_0^\infty \frac{\alpha(\alpha+1)}{(z+\eta_t\underline{\hat{L}}_{t,i}+1)^{\alpha+2}}\left(1 - \frac{1}{(z+\eta_t\underline{\hat{L}}_{t,i}+1)^\alpha}\right)^{\sigma_{t,i}-1}\mathrm{d}z}{\int_0^\infty \frac{\alpha}{(z+\eta_t\underline{\hat{L}}_{t,i}+1)^{\alpha+1}}\left(1 - \frac{1}{(z+\eta_t\underline{\hat{L}}_{t,i}+1)^\alpha}\right)^{\sigma_{t,i}-1}\mathrm{d}z}\\
&= \frac{(\alpha+1)\int_0^{\frac{1}{(1+\eta_t\underline{\hat{L}}_{t,i})^\alpha}} y^{1/\alpha}(1-y)^{\sigma_{t,i}-1}\mathrm{d}y}{\int_0^{\frac{1}{(1+\eta_t\underline{\hat{L}}_{t,i})^\alpha}}(1-y)^{\sigma_{t,i}-1}\mathrm{d}y} && (y = 1/(z+\eta_t\underline{\hat{L}}_{t,i}+1)^\alpha)\\
&= (\alpha+1)\frac{B(1/(1+\eta_t\underline{\hat{L}}_{t,i})^\alpha; 1+1/\alpha, \sigma_{t,i})}{B(1/(1+\eta_t\underline{\hat{L}}_{t,i})^\alpha; 1, \sigma_{t,i})} && (\text{incomplete Beta function } B(x;a,b))\\
&\le (\alpha+1)\frac{e\alpha}{(\alpha+1)}\frac{1}{(1+\eta_t\underline{\hat{L}}_{t,i})} && (\text{by (36) of Lee et al. (2024)})
\end{aligned}$$

which concludes the proof for the first term in the stability term. For the second term, Lee et al. (2024) showed that

$$\begin{aligned}
\frac{B(x; 1+1/\alpha, \sigma_{t,i})}{B(x; 1, \sigma_{t,i})} &\le \frac{B(1+1/\alpha, \sigma_{t,i})}{B(1, \sigma_{t,i})} && (\text{Beta function } B(a,b))\\
&\le \frac{2\alpha}{\alpha+1}\Gamma\left(1+\frac{1}{\alpha}\right)\frac{1}{\sigma_{t,i}^{1/\alpha}} && (\text{Gamma function } \Gamma(a))
\end{aligned}$$

Therefore, we obtain that

$$\mathbb{E}\left[\hat{\ell}_{t,i}(\phi_i(\eta_t\hat{L}_t) - \phi_i(\eta_t(\hat{L}_t + \hat{\ell}_{t,i}e_i))\middle|\hat{L}_t\right] \le 2\eta_t \min\left(\frac{e\alpha}{1+\eta_t\underline{\hat{L}}_{t,i}}, \frac{2\alpha\Gamma(1+1/\alpha)}{\sigma_{t,i}^{1/\alpha}}\right).$$

Since $\Gamma(1+1/\alpha) < \Gamma(2) = 1$ for $\alpha > 1$, it concludes the proof for the stability term.

Penalty analysis. By definition, we have for any $i \in [K]$ that

$$\begin{aligned}\mathbb{E}\Big[\mathbb{1}[i_t = i] r_{t,i} \Big| \hat{L}_t\Big] &= \int_0^\infty \frac{\alpha}{(z + \eta_t \underline{\hat{L}}_{t,i} + 1)^\alpha} \prod_{j \neq i} \left(1 - \frac{1}{(z + \eta_t \underline{\hat{L}}_{t,j} + 1)^\alpha}\right) \mathrm{d}z \\ &\leq \int_0^\infty \frac{\alpha}{(z + \eta_t \underline{\hat{L}}_{t,i} + 1)^\alpha} \mathrm{d}z \\ &\leq \frac{\alpha}{\alpha - 1} \frac{1}{(1 + \eta_t \underline{\hat{L}}_{t,i})^{\alpha - 1}}.\end{aligned}$$

For the second part of penalty term, since $F(z + \lambda_j)$ is increasing with respect to λ_j, we have

$$\begin{aligned}\mathbb{E}\Big[\mathbb{1}[i_t = i] r_{t,i} \Big| \hat{L}_t\Big] &= \int_0^\infty \frac{\alpha}{(z + \eta_t \underline{\hat{L}}_{t,i} + 1)^\alpha} \prod_{j \neq i} \left(1 - \frac{1}{(z + \eta_t \underline{\hat{L}}_{t,j} + 1)^\alpha}\right) \mathrm{d}z \\ &\leq \int_0^\infty \frac{\alpha}{(z + \eta_t \underline{\hat{L}}_{t,i} + 1)^\alpha} \left(1 - \frac{1}{(z + \eta_t \underline{\hat{L}}_{t,i} + 1)^\alpha}\right)^{\sigma_{t,i} - 1} \mathrm{d}z \\ &= \int_0^{\frac{1}{(1 + \eta_t \underline{\hat{L}}_{t,i})^\alpha}} y^{-\frac{1}{\alpha}} (1 - y)^{\sigma_{t,i} - 1} \mathrm{d}y \qquad (y = 1/(z + \eta_t \underline{\hat{L}}_{t,i} + 1)^\alpha) \\ &\leq \int_0^1 y^{-\frac{1}{\alpha}} (1 - y)^{\sigma_{t,i} - 1} \mathrm{d}y \\ &= B\left(1 - \frac{1}{\alpha}, \sigma_{t,i}\right) = \frac{\Gamma(1 - 1/\alpha)\Gamma(\sigma_{t,i})}{\Gamma(\sigma_{t,i} + 1 - 1/\alpha)}.\end{aligned}$$

By Gautschi's inequality, we have for any $x > 0$ and $s \in (0, 1)$ that

$$\frac{\Gamma(x)}{\Gamma(x + s)} < \frac{(x + 1)^{1 - s}}{x}.$$

Since $i \in \mathbb{N}$ and $\alpha > 1$, we have

$$\begin{aligned}\frac{\Gamma(\sigma_{t,i})}{\Gamma(\sigma_{t,i} + 1 - 1/\alpha)} \leq \frac{(\sigma_{t,i} + 1)^{\frac{1}{\alpha}}}{\sigma_{t,i}} &= \frac{\sigma_{t,i}^{1/\alpha}}{\sigma_{t,i}} \left(\frac{\sigma_{t,i} + 1}{\sigma_{t,i}}\right)^{\frac{1}{\alpha}} \\ &\leq 2^{1/\alpha} \frac{1}{\sigma_{t,i}^{1 - 1/\alpha}}.\end{aligned}$$

While we can show that $\frac{\alpha}{\alpha - 1} \leq 2^{\frac{1}{\alpha}} \Gamma(1 - 1/\alpha)$ for $\alpha > 1$, for clear α dependency, we show that $2^{\frac{1}{\alpha}} \Gamma(1 - 1/\alpha) \leq \frac{2\alpha}{\alpha - 1}$ for $\alpha > 1$. This is equivalent to show that

$$2^{\frac{1}{\alpha}} \left(1 - \frac{1}{\alpha}\right) \Gamma\left(1 - \frac{1}{\alpha}\right) \leq 2.$$

By the property of the Gamma function, $\Gamma(x + 1) = x\Gamma(x)$, this is also equivalent to show

$$2^{\frac{1}{\alpha}} \Gamma\left(2 - \frac{1}{\alpha}\right) \leq 2.$$

Since $\Gamma(x) \leq 1$ for $x \in [1, 2]$ and $2^{1/\alpha} < 2$ for $\alpha > 1$, the above inequality is valid. ■

Remark 9 *In (36) of Lee et al. (2024), the original inequality includes an additional factor $1/\sigma_{t,i}$ in the upper bound, which does not hold in general. However, the authors also used a version of the bound without this factor, which is valid. Therefore, this issue does not affect the correctness of their results; we note it here for completeness.*

C.3. Proof of Lemma 1 (Lemma 8)

Proof For any $i \in [K]$ and $t \in \mathbb{N}$, by definition of β_t in (11), we have

$$\frac{\frac{1}{(1+\eta_{t+1}\underline{\hat{L}}_{t+1,i})^\alpha}}{\frac{1}{(1+\eta_t\underline{\hat{L}}_{t+1,i})^\alpha}} = \left(\frac{1+\eta_t\underline{\hat{L}}_{t+1,i}}{1+\eta_{t+1}\underline{\hat{L}}_{t+1,i}}\right)^\alpha = \left(\frac{\beta_{t+1}}{\beta_t}\frac{\beta_t+\underline{\hat{L}}_{t+1,i}}{\beta_{t+1}+\underline{\hat{L}}_{t+1,i}}\right)^\alpha \le \left(\frac{\beta_{t+1}}{\beta_t}\right)^\alpha.$$

When $\alpha \ge 2$. In this case, by definition of β_t in (11), the result directly follows.

When $\alpha \in (1,2)$. By the update rule of the learning rates, it holds that

$$\frac{\beta_{t+1}}{\beta_t} = 1 + \frac{1}{\beta_t}\max\left(\frac{z_t}{\beta_t h_t}, \frac{4}{(2^{1/\alpha}-1)t}\right),$$

where

$$\frac{z_t}{h_t} = \frac{\sum_{i\neq j_{t+1}} \alpha q_{t+1,i}^{1/\alpha}}{\sum_{i\neq j_{t+1}} \frac{\alpha}{\alpha-1} q_{t+1,i}^{1-1/\alpha}} = (\alpha-1)\frac{\sum_{i\neq j_{t+1}} q_{t+1,i}^{1/\alpha}}{\sum_{i\neq j_{t+1}} q_{t+1,i}^{1-1/\alpha}}.$$

When $\alpha \in (1,2)$, $\frac{1}{\alpha} \ge 1-\frac{1}{\alpha}$ holds. Since $q_{t,i} \in (0,1/2]$ by its definition for any $i \neq j_{t+1}$ and $t \in \mathbb{N}$, we have $q_{t+1,i}^{\frac{1}{\alpha}} \le q_{t+1,i}^{1-\frac{1}{\alpha}}$. Therefore, $z_t/h_t \le 1$ for any t, which implies

$$\frac{\beta_{t+1}}{\beta_t} \le 1 + \frac{1}{\beta_t}\max\left(\frac{1}{\beta_t}, \frac{4}{(2^{1/\alpha}-1)t}\right).$$

Here, Lemma 11 shows that $\beta_t \ge \frac{4\log t}{(2^{1/\alpha}-1)}$ for $\alpha \in (1,2)$. This implies that for $t \ge 3$

$$\begin{aligned}\frac{\beta_{t+1}}{\beta_t} \le 1 + \frac{2^{1/\alpha}-1}{4\log t}\max\left(\frac{2^{1/\alpha}-1}{4\log t}, \frac{4}{(2^{1/\alpha}-1)t}\right) &\le 1 + \frac{(2^{1/\alpha}-1)^2}{16\log^2 t} + \frac{1}{t\log t} \\ &\le 1 + \frac{1}{16\log^2 3} + \frac{1}{3\log 3} \\ &\le 1.36.\end{aligned}$$

Since $(1.36)^2 < 2$, we obtain the desired results. ■

C.4. Proof of Lemma 2

To decompose the regret in the desired formulation, we need to relocate the contribution from j_t in the stability to those from $i \neq j_t$. For this purpose, we need the following lemma, whose proof is given in Appendix C.5.

Lemma 10 *Algorithm 1 with shape $\alpha > 1$ and β_t defined in (11) satisfies that*

$$\mathbb{E}\left[\hat{\ell}_{t,j_t}(\phi_{j_t}(\eta_t\hat{L}_t) - \phi_{j_t}(\eta_t(\hat{L}_t + \hat{\ell}_{t,j_t}e_{j_t}))\middle|\hat{L}_t\right] \le \mathcal{O}\left(\frac{\sum_{i\neq j_t}\alpha p_{t,i}^{1/\alpha}}{\beta_t}\right).$$

The above inequality holds for all t in the case of $\alpha \in (1,2)$ and $t \ge t_0(\alpha, K)$ when $\alpha \ge 2$.

Proof of Lemma 2 By using Lemma 4 of Kim et al. (2026) with Lemma 18 of Lee et al. (2024), it holds that

$$\mathrm{Reg}_{\mathrm{FTPL}}(T) \le \sum_{t=1}^{T}\mathbb{E}\left[\left\langle\hat{\ell}_t, \phi(\eta_t\hat{L}_t) - \phi(\eta_t\hat{L}_{t+1})\right\rangle\right] + (\beta_{t+1}-\beta_t)\mathbb{E}\left[r_{t+1,i_{t+1}} - r_{t+1,i^*}\right] + \beta_1\left(\frac{\alpha}{\alpha-1}\right)^2 K^{\frac{1}{\alpha}}. \quad (21)$$

For the penalty term, which is the second term of (21), we have

$$\mathbb{E}\left[r_{t+1,i_{t+1}} - r_{t+1,i^*}\right] = \mathbb{E}\left[\mathbb{E}\left[r_{t+1,i_{t+1}} - r_{t+1,i^*}\middle|\hat{L}_{t+1}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[r_{t+1,i_{t+1}} - r_{t+1,j_{t+1}}\middle|\hat{L}_{t+1}\right]\right]$$

since j_{t+1} is fixed given $\hat{L}_{t+1}$ and $r_{t,i}$s are independently distributed from the identical distribution.

For the stability term, the first term of (21), Lemmas 7 and 10 imply that

$$\begin{aligned}\mathbb{E}\left[\left\langle\hat{\ell}_t, \phi(\eta_t\hat{L}_t) - \phi(\eta_t\hat{L}_{t+1})\right\rangle\middle|\hat{L}_t\right] &\le \frac{\sum_{i\neq j_t}2e\alpha p_{t,i}^{\frac{1}{\alpha}}}{\beta_t} + \mathbb{E}\left[\hat{\ell}_{t,j_t}(\phi_{j_t}(\eta_t\hat{L}_t) - \phi_{j_t}(\eta_t(\hat{L}_t+\hat{\ell}_t))\middle|\hat{L}_t\right] \\ &\le \mathcal{O}\left(\frac{\alpha\sum_{i\neq j_t}p_{t,i}^{1/\alpha}}{\beta_t}\right),\end{aligned}$$

for all $t \in \mathbb{N}$ if $\alpha \in (1,2)$ and $t \ge t_0(\alpha, K)$ for $\alpha \ge 2$. Therefore with Lemma 7, (21) is written

$$\mathrm{Reg}_{\mathrm{FTPL}}(T) \le \sum_{t=1}^{T}\mathcal{O}\left(\frac{\alpha\sum_{i\neq j_t}p_{t,i}^{\frac{1}{\alpha}}}{\beta_t}\right) + (\beta_{t+1}-\beta_t)\sum_{i\neq j_{t+1}}\frac{2\alpha}{\alpha-1}p_{t+1,i}^{1-\frac{1}{\alpha}} + \beta_1\left(\frac{\alpha}{\alpha-1}\right)^2 K^{\frac{1}{\alpha}}.$$

Here, we have

$$\begin{aligned}\sum_{t=1}^{T}\frac{\alpha\sum_{i\neq j_t}p_{t,i}^{\frac{1}{\alpha}}}{\beta_t} &= \frac{\alpha\sum_{i\neq j_1}p_{1,i}^{1/\alpha}}{\beta_1} + \sum_{t=1}^{T-1}\frac{\alpha\sum_{i\neq j_{t+1}}p_{t+1,i}^{\frac{1}{\alpha}}}{\beta_{t+1}} \\ &\le \frac{\alpha\sum_{i\neq j_1}p_{1,i}^{1/\alpha}}{\beta_1} + \sum_{t=1}^{T-1}\frac{\alpha\sum_{i\neq j_{t+1}}p_{t+1,i}^{\frac{1}{\alpha}}}{\beta_t} && (\because \beta_{t+1}\ge\beta_t) \\ &= \frac{\alpha\sum_{n=2}^{K}n^{-1/\alpha}}{\beta_1} + \sum_{t=1}^{T-1}\frac{\alpha\sum_{i\neq j_{t+1}}p_{t+1,i}^{\frac{1}{\alpha}}}{\beta_t} && (\because \hat{L}_1 = 0) \\ &\le \frac{\alpha^2}{\alpha-1}\frac{K^{1-\frac{1}{\alpha}}}{\beta_1} + \sum_{t=1}^{T-1}\frac{\alpha\sum_{i\neq j_{t+1}}p_{t+1,i}^{\frac{1}{\alpha}}}{\beta_t},\end{aligned}$$

which implies that

$$\mathrm{Reg}_{\mathrm{FTPL}}(T) \le \sum_{t=1}^{T-1} \mathcal{O}\left(\frac{\alpha \sum_{i \neq j_{t+1}} p_{t+1,i}^{1/\alpha}}{\beta_t}\right) + (\beta_{t+1} - \beta_t)\mathcal{O}\left(\frac{\alpha}{\alpha - 1} \sum_{i \neq j_{t+1}} p_{t+1,i}^{1-1/\alpha}\right) + \frac{1}{\beta_1}\frac{\alpha^2 K^{1-1/\alpha}}{\alpha - 1} + \beta_1 \left(\frac{\alpha}{\alpha - 1}\right)^2 K^{\frac{1}{\alpha}} + \frac{z_T h'_{T+1}}{\beta_T h_T} + \mathbb{1}[\alpha \ge 2] t_0(\alpha, K)$$

Since $\beta_{T+1} - \beta_T = \frac{z_T}{\beta_T h_T} \le \mathcal{O}(\alpha K / \beta_T)$, whenever T is sufficiently large, the last term becomes negligible since $\beta_T \ge \log T$ for any $\alpha > 1$. Then, Lemma 8 (Lemma 1) implies that

$$\mathrm{Reg}_{\mathrm{FTPL}}(T) \le \sum_{t=1}^{T-1} \mathcal{O}\left(\frac{\alpha \sum_{i \neq j_{t+1}} q_{t+1,i}^{1/\alpha}}{\beta_t}\right) + (\beta_{t+1} - \beta_t)\mathcal{O}\left(\frac{\alpha}{\alpha - 1} \sum_{i \neq j_{t+1}} q_{t+1,i}^{1-1/\alpha}\right) + \frac{1}{\beta_1}\frac{\alpha^2 K^{1-1/\alpha}}{\alpha - 1} + \beta_1 \left(\frac{\alpha}{\alpha - 1}\right)^2 K^{\frac{1}{\alpha}} + \mathbb{1}[\alpha \ge 2] t_0(\alpha, K) + \mathbb{1}[\alpha \in (1,2)] 2,$$

which concludes the proof. Note that the additional $\log T$ term for $\alpha \in (1,2)$ comes from the summation of $\frac{4}{(2^{1/\alpha}-1)t} \cdot \frac{\alpha}{\alpha - 1} \sum_{i \neq j_{t+1}} q_{t+1,i}^{1-1/\alpha}$. ■

C.5. Proof of Lemma 10

Before the proof of Lemma 10, we provide a lower bound on the learning rates β_t defined in (11). Although this bound follows directly from its definition, we include the proof for completeness in Appendix C.7.

Lemma 11 *For β_t defined in (11) satisfies that*

$$\beta_t \ge \begin{cases} \sqrt{\beta_1^2 + 2(t-1)}, & \text{if } \alpha \ge 2, \\ \beta_1 + \frac{4 \log t}{(2^{1/\alpha} - 1)}, & \text{if } \alpha \in (1,2). \end{cases}$$

Proof of Lemma 10 Define a variable ξ_α as

$$\xi_\alpha = \begin{cases} \frac{1}{2}, & \text{if } \alpha \in (1,2), \\ \frac{1}{(2 - 2^{-1/(\alpha+1)})^\alpha}, & \text{if } \alpha \ge 2. \end{cases}$$

Note that $\xi_\alpha \ge 1/2$ for all $\alpha > 1$ by definition. Then, we consider the event $\mathcal{E}_{t,\alpha}$, which is

$$\mathcal{E}_{t,\alpha} = \left\{ \sum_{i \neq j_t} \frac{1}{(1 + \eta_t \underline{\hat{L}}_{t,i})^\alpha} < \xi_\alpha \right\}.$$

Note that by the relationship $w_{t,i} \le p_{t,i} \le 1/(1 + \eta_t \underline{\hat{L}}_{t,i})^\alpha$, (see Section C.1 for the first inequality), that

$$\sum_{i \neq j_t} w_{t,i} \le \xi_\alpha, \text{ and } w_{t,j_t} \ge 1 - \xi_\alpha. \tag{22}$$

Then, we consider the case $\mathcal{E}_{t,\alpha}$ and $\mathcal{E}_{t,\alpha}^c$ separately.

C.5.1. THE CASE OF $\mathcal{E}_{t,\alpha}^c$

Even on $\mathcal{E}_{t,\alpha}^c$, the results in Lemma 7 are still valid for j_t, which implies that

$$\mathbb{E}\Big[\hat{\ell}_{t,j_t}\Big(\phi_{j_t}(\eta_t\hat{L}_t)-\phi_{j_t}(\eta_t\hat{L}_{t+1})\Big)\Big|\hat{L}_t\Big]\leq\frac{e\alpha}{\beta_t}p_{t,j_t}^{1/\alpha}=\frac{e\alpha}{\beta_t},$$

where the last equality holds since $\sigma_{t,j_t}=1$ and $\hat{\underline{L}}_{t,j_t}=0$ must hold by definition. Therefore, it suffices to show that $1\leq a\sum_{i\neq j_t}p_{t,i}^{1/\alpha}$ for some K-independent constant a, where we show the results for the case of $a=2$. Note that $\sum_{i\neq j_t}1/(1+\eta_t\hat{\underline{L}}_{t,i})^\alpha\geq\xi_\alpha\geq 1/2$ by definition of $\mathcal{E}_{t,\alpha}^c$.

Firstly, assume for all $i\neq j_t$ that

$$\frac{1}{(1+\eta_t\hat{\underline{L}}_{t,i})^\alpha}\leq\frac{1}{\sigma_{t,i}}\iff p_{t,i}=\frac{1}{(1+\eta_t\hat{\underline{L}}_{t,i})^\alpha}.$$

Then by definition of $\mathcal{E}_{t,\alpha}^c$, we obtain that

$$\frac{1}{2}\leq\sum_{i\neq j_t}\frac{1}{(1+\eta_t\hat{\underline{L}}_{t,i})^\alpha}\leq\sum_{i\neq j_t}\frac{1}{1+\eta_t\hat{\underline{L}}_{t,i}}=\sum_{i\neq j_t}p_{t,i}^{\frac{1}{\alpha}}.$$

Next, let us consider the case where there exists an arm $i\neq j_t$ such that $1/(1+\eta_t\hat{\underline{L}}_{t,i})^\alpha\geq 1/\sigma_{t,i}$. In this case, since $1/(1+z)^\alpha$ is decreasing with respect to z, arms j with $\sigma_{t,j}\leq\sigma_{t,i}$ should satisfy that

$$\frac{1}{(1+\eta_t\hat{\underline{L}}_{t,j})^\alpha}\geq\frac{1}{(1+\eta_t\hat{\underline{L}}_{t,i})^\alpha}\geq\frac{1}{\sigma_{t,i}},\ \forall j\text{ s.t. }\sigma_{t,j}\leq\sigma_{t,i}.$$

This implies that

$$\sum_{i\neq j_t}p_{t,i}^{\frac{1}{\alpha}}\geq\sum_{i\neq j_t}p_{t,i}\geq\sum_{j:\sigma_{t,j}\leq\sigma_{t,i},j\neq j_t}\frac{1}{\sigma_{t,i}}=1-\frac{1}{\sigma_{t,i}}\geq\frac{1}{2}.$$

Therefore, for any cases, we obtain that

$$\mathbb{E}\Big[\mathbb{1}[\mathcal{E}_{t,\alpha}]\hat{\ell}_{t,j_t}\Big(\phi_{j_t}(\eta_t\hat{L}_t)-\phi_{j_t}(\eta_t\hat{L}_{t+1})\Big)\Big|\hat{L}_t\Big]\leq\mathbb{1}[\mathcal{E}_{t,\alpha}]\frac{e\alpha}{\beta_t}\leq\mathbb{1}[\mathcal{E}_{t,\alpha}]\frac{2e\alpha}{\beta_t}\sum_{i\neq j_t}p_{t,i}^{\frac{1}{\alpha}},$$

which concludes the proof for the case on $\mathcal{E}_{t,\alpha}^c$.

C.5.2. THE CASE OF $\mathcal{E}_{t,\alpha}$

On $\mathcal{E}_{t,\alpha}$, it is clear that $\eta_t\hat{\underline{L}}_{t,i}\geq\xi_\alpha^{-1/\alpha}-1>0$ holds. Let ζ_α be an α-dependent dependent constant in $(0,\xi_\alpha^{-1/\alpha}-1)$, specified later. Then, whenever $\hat{\ell}_{t,j_t}\leq\zeta_\alpha/\eta_t$, we can apply the same techniques

in Lemma 25 of Lee et al. (2024), which shows that

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{1}[\mathcal{E}_{t,\alpha}, \hat{\ell}_{t,j_t} \le \zeta_\alpha \beta_t] \hat{\ell}_{t,j_t} \Big(\phi_{j_t}(\eta_t \hat{L}_t) - \phi_{j_t}(\eta_t(\hat{L}_t + \hat{\ell}_{t,j_t} e_{j_t}))\Big) \Big| \hat{L}_t\Big] && (\beta_t = 1/\eta_t) \\
&\le \mathbb{E}\left[\mathbb{1}[\mathcal{E}_{t,\alpha}, \hat{\ell}_{t,j_t} \le \zeta_\alpha \beta_t] e^2(1-e^{-1}) \hat{\ell}_{t,j_t}^2 \sum_{i \neq j_t} \frac{\eta_t \alpha}{(1+\eta_t(\underline{\hat{L}}_{t,i} - \zeta_\alpha))^{\alpha+1}} \middle| \hat{L}_t\right] \\
&\le \mathbb{E}\left[\mathbb{1}[\mathcal{E}_{t,\alpha}, \hat{\ell}_{t,j_t} \le \zeta_\alpha \beta_t] e^2(1-e^{-1}) \frac{2\ell_{t,j_t}^2 \mathbb{1}[i_t = j_t]}{w_{t,j_t}^2} \sum_{i \neq j_t} \frac{\eta_t \alpha}{(1+\eta_t(\underline{\hat{L}}_{t,i} - \zeta_\alpha))^{\alpha+1}} \middle| \hat{L}_t\right] \\
&= \mathbb{E}\left[\mathbb{1}[\mathcal{E}_{t,\alpha}, \hat{\ell}_{t,j_t} \le \zeta_\alpha \beta_t] \frac{2e^2(1-e^{-1})\ell_{t,j_t}^2}{w_{t,j_t}} \sum_{i \neq j_t} \frac{\eta_t \alpha}{(1+\eta_t(\underline{\hat{L}}_{t,i} - \zeta_\alpha))^{\alpha+1}} \middle| \hat{L}_t\right] \\
&\le \frac{2e^2(1-e^{-1})}{1-\xi_\alpha} \sum_{i \neq j_t} \frac{\eta_t \alpha}{(1+\eta_t(\underline{\hat{L}}_{t,i} - \zeta_\alpha))^{\alpha+1}}, && (23)
\end{aligned}$$

where the last inequality follows from (22) and $\ell_t \in [0,1]^K$.

When $\alpha \in (1,2)$. In this case, we set $\xi_\alpha = 1/2$, where $\xi_\alpha^{-1/\alpha} - 1 \in (\sqrt{2}-1, 1)$. Here, we take $\zeta_\alpha = (\xi_\alpha^{-1/\alpha} - 1)/2$ for analytical simplicity, which implies $\hat{\ell}_{t,j_t} \le \underline{\hat{L}}_{t,i}/2$. Then, (23) satisfies that

$$\begin{aligned}
\frac{2e^2(1-e^{-1})}{1-\xi_\alpha} \sum_{i \neq j_t} \frac{\eta_t \alpha}{(1+\eta_t(\underline{\hat{L}}_{t,i} - \zeta_\alpha))^{\alpha+1}} &\le \sum_{i \neq j_t} \eta_t \frac{2^{\alpha+2} e^2 (1-e^{-1})\alpha}{(2+\eta_t \underline{\hat{L}}_{t,i})^{\alpha+1}} \\
&\le \sum_{i \neq j_t} \eta_t \frac{2^{\alpha+2} e^2 (1-e^{-1})\alpha}{(1+\eta_t \underline{\hat{L}}_{t,i})^{\alpha+1}}. && (24)
\end{aligned}$$

Since $\alpha \in (1,2)$, the multiplicative constant is at most $16e^2(1-e^{-1}) \lesssim 75$. Then, on $\mathcal{E}_{t,\alpha}$ and $\hat{\ell}_{t,j_t} \le \zeta_\alpha \beta_t$, it remains to show how (24) provides the desired results, i.e., the upper bounds in terms of $\sum_{i \neq j_t} p_{t,i}^{1/\alpha}$.

By definition of $\mathcal{E}_{t,\alpha}$, as mentioned above, $\underline{\hat{L}}_{t,i} > 0$ holds for any $i \neq j_t$ on $\mathcal{E}_{t,\alpha}$. Since $1/(1+x)^\alpha$ is decreasing with respect to x, for any $i \neq j_t$, it holds that

$$\sum_{i \neq j_t} \frac{1}{(1+\eta_t \underline{\hat{L}}_{t,i})^\alpha} \ge \frac{\sigma_{t,i} - 1}{(1+\eta_t \underline{\hat{L}}_{t,i})^\alpha},$$

which implies that on $\mathcal{E}_{t,\alpha}$ for $\alpha \in (0,1)$

$$\frac{1}{(1+\eta_t \underline{\hat{L}}_{t,i})^\alpha} \le \frac{1}{2} \frac{1}{\sigma_{t,i} - 1} \le \frac{1}{\sigma_{t,i}} \implies p_{t,i} = \frac{1}{(1+\eta_t \underline{\hat{L}}_{t,i})^\alpha}, \qquad (25)$$

where the last inequality holds since $\sigma_{t,i} \geq 2$ for $i \neq j_t$ by definition. Therefore, we obtain that

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{1}[\mathcal{E}_{t,\alpha}, \hat{\ell}_{t,j_t} \leq \zeta_\alpha \beta_t] \hat{\ell}_{t,j_t} \Big(\phi_{j_t}(\eta_t \hat{L}_t) - \phi_{j_t}(\eta_t(\hat{L}_t + \hat{\ell}_{t,j_t} e_{j_t}))\Big) \Big| \hat{L}_t\Big] \\
&\qquad \leq \sum_{i \neq j_t} \eta_t \frac{2^{\alpha+2} e^2 (1 - e^{-1}) \alpha}{\sigma_{t,i}} \frac{1}{1 + \eta_t \hat{\underline{L}}_{t,i}} \\
&\qquad \leq \sum_{i \neq j_t} 2^{\alpha+1} e^2 (1 - e^{-1}) \alpha \frac{\eta_t}{1 + \eta_t \hat{\underline{L}}_{t,i}} \\
&\qquad = \sum_{i \neq j_t} 2^{\alpha+1} e^2 (1 - e^{-1}) \alpha \frac{p_{t,i}^{1/\alpha}}{\beta_t}, \qquad (\beta_t = 1/\eta_t)
\end{aligned}$$

as desired. Finally, it remains to consider the case $\hat{\ell}_{t,j_t} > \zeta_\alpha \beta_t$ on $\mathcal{E}_{t,\alpha}$. We show that this event cannot occur under the design of Algorithm 1, which employs CGR II-biased with the number of maximum resampling steps G_t. Therefore, it suffices to show that when $\mathbf{1}[\mathcal{E}_{t,\alpha}, i_t = j_t] = 1$,

$$\hat{\ell}_{t,j_t} \leq \ell_{t,j_t} G_t \leq G_t \leq \zeta_\alpha \beta_t.$$

For $\alpha \in (1, 2)$, we set $G_t = 2 \log t$ when both $\mathcal{E}_{t,\alpha}$ and $i_t = j_t$ occur. Therefore, Lemma 11 concludes the proof.

When $\alpha \geq 2$. Let $\zeta_\alpha = 1 - 2^{-1/(\alpha+1)}$, such that $\frac{1}{(1-\zeta_\alpha)^{\alpha+1}} = 2$. Then, on $\mathcal{E}_{t,\alpha}$, it is clear that $\eta_t \hat{\underline{L}}_{t,i} \geq \zeta_\alpha$ for all $i \neq j_t$ by the choice of ξ_α and ζ_α. Whenever $\hat{\ell}_{t,j_t} \leq \zeta_\alpha / \eta_t$, we can apply the same techniques as the case of $\alpha \in (1, 2)$, which shows that

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{1}[\mathcal{E}_{t,\alpha}, \hat{\ell}_{t,j_t} \leq \zeta_\alpha \beta_t] \hat{\ell}_{t,j_t} \Big(\phi_{j_t}(\eta_t \hat{L}_t) - \phi_{j_t}(\eta_t(\hat{L}_t + \hat{\ell}_{t,j_t} e_{j_t}))\Big) \Big| \hat{L}_t\Big] \\
&\leq \mathbb{E}\left[\mathbb{1}[\mathcal{E}_{t,\alpha}, \hat{\ell}_{t,j_t} \leq \zeta_\alpha \beta_t] \frac{e^2(1 - e^{-1})}{(1 - \zeta_\alpha)^{\alpha+1}} \hat{\ell}_{t,j_t}^2 \sum_{i \neq j_t} \frac{\eta_t \alpha}{(1 + \eta_t \hat{\underline{L}}_{t,i})^{\alpha+1}} \middle| \hat{L}_t\right] \\
&\leq \mathbb{E}\left[\mathbb{1}[\mathcal{E}_{t,\alpha}, \hat{\ell}_{t,j_t} \leq \zeta_\alpha \beta_t] \frac{e^2(1 - e^{-1})}{(1 - \zeta_\alpha)^{\alpha+1}} \frac{2\ell_{t,j_t}^2 \mathbb{1}[i_t = j_t]}{w_{t,j_t}^2} \sum_{i \neq j_t} \frac{\eta_t \alpha}{(1 + \eta_t \hat{\underline{L}}_{t,i})^{\alpha+1}} \middle| \hat{L}_t\right] \\
&= \mathbb{E}\left[\mathbb{1}[\mathcal{E}_{t,\alpha}, \hat{\ell}_{t,j_t} \leq \zeta_\alpha \beta_t] \frac{4e^2(1 - e^{-1}) \ell_{t,j_t}^2}{w_{t,j_t}} \sum_{i \neq j_t} \frac{\eta_t \alpha}{(1 + \eta_t \hat{\underline{L}}_{t,i})^{\alpha+1}} \middle| \hat{L}_t\right] \\
&\leq \frac{4e^2(1 - e^{-1})}{1 - \xi_\alpha} \sum_{i \neq j_t} \frac{\eta_t \alpha}{(1 + \eta_t \hat{\underline{L}}_{t,i})^{\alpha+1}} \qquad \text{(by (22) and } \ell_t \in [0,1]^K) \\
&\leq 13 e^2 (1 - e^{-1}) \alpha \sum_{i \neq j_t} \frac{\eta_t}{(1 + \eta_t \hat{\underline{L}}_{t,i})^{\alpha+1}},
\end{aligned}$$

where the last inequality follows from that $\frac{1}{\xi_\alpha - 1} = \frac{(1+\zeta_\alpha)^\alpha}{(1+\zeta_\alpha)^\alpha - 1}$ is decreasing with respect to $\alpha > 1$ and its value at $\alpha = 2$ is less than 3.2.

Similarly, it remains to consider the case $\hat{\ell}_{t,j_t} > \zeta_\alpha \beta_t$ on $\mathcal{E}_{t,\alpha}$ for $\alpha \geq 2$ case. As one can easily expect, since $\beta_t = \Omega(\sqrt{t})$, it is possible to directly apply Lemma 20, which provides an additional

term whose summation over t is at most $\mathcal{O}(K^{2/\alpha-1}) \le \mathcal{O}(1)$ for $\alpha \ge 2$. Since this direct application requires to modify the constant term appear on $\mathcal{E}_{t,\alpha}^c$ case, for the coherence with $\alpha \in (1,2)$, we show that

$$\hat{\ell}_{t,j_t} \le \ell_{t,j_t} G_t \le K \log t \le \zeta_\alpha \sqrt{\beta_1^2 + 2(t-1)}.$$

for $t \ge t_0(\alpha, K)$. Since $\beta_1^2 = \mathcal{O}(\alpha^2 K^{1-2/\alpha})$ by our choice, it suffices to find the solution of $\log t \le a\sqrt{t}$ for $a = \zeta_\alpha \sqrt{2}/K$. Let $t = y^2$. Then,

$$\log y = ay/2 \iff y = e^{\frac{a}{2}y} \iff ze^z = -\frac{a}{2}. \qquad (z = -\tfrac{a}{2}y)$$

Since $-\frac{a}{2} = -\frac{\zeta_\alpha\sqrt{2}}{K} \in (-1/e, 0)$, the above equality admits two real solution $z = W_n(-a/2)$, where $W_n(\cdot)$ denotes the Lambert W function with branch n and $n \in \{-1, 0\}$ (i.e., only the principal branch since we consider the real value) (Olver et al., 2010, Section 4). Therefore, we obtain the desired results for any $t \ge t_0 := \frac{K^2}{2\zeta_\alpha^2}\left(W_{-1}\left(-\frac{\zeta_\alpha\sqrt{2}}{K}\right)\right)^2$. For sufficiently small a, we can approximate the Lambert W function as (Olver et al., 2010, 4.13.11)

$$-W_{-1}(-a) \approx \log\frac{1}{a} + \log\log\frac{1}{a} \implies t_0(\alpha, K) \approx \left(\frac{K}{\sqrt{2}\zeta_\alpha}\right)^2 \left(\log\frac{K}{\sqrt{2}\zeta_\alpha} + \log\log\frac{K}{\sqrt{2}\zeta_\alpha}\right)^2,$$

which implies that $t_0(\alpha, K) \le \mathcal{O}(\alpha^2 K^2 \log^2(\alpha K))$ with the current choice of ζ_α.

In sum, for $t \ge t_0(\alpha, K)$, we obtain

$$\mathbb{E}\left[\mathbb{1}[\mathcal{E}_{t,\alpha}]\hat{\ell}_{t,j_t}\left(\phi_{j_t}(\eta_t \hat{L}_t) - \phi_{j_t}(\eta_t \hat{L}_{t+1})\right)\middle|\hat{L}_t\right] \le \sum_{i \ne j_t} \mathcal{O}\left(\frac{\eta_t \mathbb{1}[\mathcal{E}_{t,\alpha}]}{(1 + \eta_t \underline{\hat{L}}_{t,i})^{\alpha+1}}\right).$$

Similarly to the case of $\alpha \in (1,2)$ in (25), by definition, we have for $i \ne j_t$

$$\frac{1}{(1+\eta_t \underline{\hat{L}}_{t,i})^\alpha} \le \frac{\xi_\alpha}{\sigma_{t,i} - 1} \le 2\xi_\alpha \frac{1}{\sigma_{t,i}}, \qquad (\because \sigma_{t,i} \ge 2,\ \forall i \ne j_t)$$

which concludes the proof. ■

Remark 12 *In the current analysis, we obtain a loose bound in constant, especially the bound on $\mathcal{E}_{t,\alpha}$, e.g. 75 in $\alpha \in (1,2)$ case. However, we can tune both ξ_α and ζ_α to reduce the constant term. In $\alpha \in (1,2)$ example, we can choose $\zeta_\alpha = 4(\xi_\alpha^{-1/\alpha} - 1)/5$, we change the term $2^{\alpha+1}$ to $(5/4)^{\alpha+1}$, which results in around 18 (and thus 9 in the final bound) instead of 75 in the multiplicative constant. Note that when we modify the choice of ζ_α, this will affect the bound on $\mathcal{E}_{t,\alpha}^c$, where multiplicative constant 2 becomes $1/\xi_\alpha$ and we need to modify the constant in $1/t$ term in β_t in (11) since $4/(2^{1/\alpha} - 1)$ is chosen to satisfy $G_t \approx \log t/(1-\xi_\alpha) \le \zeta_\alpha \beta_t$. Therefore, the choice of ξ_α and ζ_α should take multiple factors into account simultaneously.*

C.6. Proof of Lemma 3

Although the proof of Lemma 3 can be directly obtained by Lemmas 9 and 10 in Ito et al. (2024), we provide the results for the completeness since our β_t in (11) for $\alpha \in (1,2)$ includes additional term and for $\alpha \geq 2$ includes the restriction on the exponential growth. Before the proof, we introduce the following results.

Lemma 13 (Lemma 10 of Ito et al. (2024)) *It holds that*

$$\sum_{t=1}^{T} \frac{z_t}{\sqrt{\sum_{s=1}^{t} \frac{z_s}{h_s}}} \leq \mathcal{O}\left(\min\left\{ \sqrt{\log T \sum_{t=1}^{T} h_t z_t} + \sqrt{h_{\max} z_{\max}}, \sqrt{h_{\max} \sum_{t=1}^{T} z_t} \right\} \right).$$

Proof of Lemma 3 Define an auxiliary sequence $\beta'_t = \sqrt{\beta_1^2 + 2\sum_{s=1}^{t-1} z_s/h_s}$ so that $\beta'_t \leq \beta_t$ holds by their definitions for $\alpha \in (1,2)$. Note that for the case of $\alpha \in (1,2)$, since we take the maximum of two values, $\beta'_t \leq \beta_t$ always holds. Then, it remains to follow the proofs in previous results (Ito et al., 2024; Nguyen et al., 2025).

Define a set of rounds $\mathcal{T} := \left\{ t \in [T] : \beta'_{t+1} \geq \sqrt{2}\beta'_t \right\}$, where we can rewrite

$$\sum_{t=1}^{T} \frac{z_t}{\beta_t} = \sum_{t \in \mathcal{T}} \frac{z_t}{\beta_t} + \sum_{t \in \mathcal{T}^c} \frac{z_t}{\beta_t}.$$

By definition of $\mathcal{T}$ and $\beta_1 = \mathcal{O}(\alpha K^{\frac{1}{2}-\frac{1}{\alpha}})$, we have

$$\begin{aligned}
\sum_{t \in \mathcal{T}} \frac{z_t}{\beta_t} \leq \sum_{t \in \mathcal{T}} \frac{z_{\max}}{\beta'_t} &\leq \sum_{s=0}^{\infty} \left(\frac{1}{\sqrt{2}} \right)^s \frac{z_{\max}}{\beta_1} \\
&\leq (2+\sqrt{2}) \frac{z_{\max}}{\beta_1} \\
&\leq (2+\sqrt{2}) \frac{\alpha \sum_{n=2}^{K} n^{-1/\alpha}}{\beta_1} \\
&\leq \frac{\alpha^2 (2+\sqrt{2})}{\alpha - 1} \frac{(K+1)^{1-1/\alpha} - 1}{\beta_1} \\
&\leq \frac{\alpha^2 (2+\sqrt{2})}{\alpha - 1} \frac{K^{1-1/\alpha}}{\beta_1} = \mathcal{O}\left(\frac{\alpha}{\alpha - 1} \sqrt{K} \right).
\end{aligned} \tag{26}$$

On the other hand, we have

$$\sum_{t \in \mathcal{T}^c} \frac{z_t}{\beta_t} \leq \sum_{t \in \mathcal{T}^c} \frac{z_t}{\beta'_t} \leq \sqrt{2} \sum_{t \in \mathcal{T}^c} \frac{z_t}{\beta'_{t+1}} = \sqrt{2} \sum_{t \in \mathcal{T}^c} \frac{z_t}{\sqrt{\beta_1^2 + 2\sum_{s=1}^{t} z_s/h_s}} \leq \sum_{t=1}^{T} \frac{z_t}{\sqrt{\sum_{s=1}^{t} \frac{z_s}{h_s}}},$$

which concludes the proof for $\alpha \in (1,2)$ by applying Lemma 13.

For $\alpha \geq 2$, define a set of rounds $\mathcal{T} := \{ t \in [T] : \beta_{t+1} \geq 2^{1/\alpha} \beta_t \}$, where $\beta_{t+1} = 2^{1/\alpha} \beta_t$ by definition of β_t in (11) for $\alpha \geq 2$. Note that for $t \in \mathcal{T}$, it holds that $\frac{z_t}{\beta_t h_t} \geq \beta_t (2^{1/\alpha} - 1)$. Then, by following the same steps in (26) with $2^{1/\alpha}$ instead of $\sqrt{2}$, we obtain

$$\sum_{t \in \mathcal{T}} \frac{z_t}{\beta_t} \leq \frac{\alpha}{(\alpha - 1)(2^{1/\alpha} - 1)} K^{\frac{1}{2}-\frac{1}{\alpha}} \leq \frac{\alpha^2}{(\alpha - 1)\log 2} K^{\frac{1}{2}-\frac{1}{\alpha}}.$$

On the other hand, since $\beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}$ holds on $\mathcal{T}^c$, we have

$$\begin{aligned}\sum_{t\in\mathcal{T}^c} \frac{z_t}{\beta_t} &\le \sum_{t\in\mathcal{T}^c} \frac{2^{1/\alpha} z_t}{\beta_{t+1}} \\ &\le \sum_{t\in\mathcal{T}^c} \frac{2^{1/\alpha} z_t}{\sqrt{\beta_1^2 + 2\sum_{s\in\mathcal{T}^c\cap[t]} \frac{z_s}{h_s}}} \\ &\le \sum_{t\in\mathcal{T}^c} \frac{z_t}{\sqrt{\sum_{s\in\mathcal{T}^c\cap[t]} \frac{z_s}{h_s}}},\end{aligned}$$

where we only consider the effect of updates in β_t on $\mathcal{T}^c$. By applying Lemma 13, we obtain

$$\begin{aligned}\sum_{t\in\mathcal{T}^c} \frac{z_t}{\beta_t} &\le \mathcal{O}\left(\min\left\{\sqrt{\log|\mathcal{T}^c| \sum_{t\in\mathcal{T}^c} h_t z_t} + \sqrt{h_{\max} z_{\max}}, \sqrt{h_{\max} \sum_{t\in\mathcal{T}^c} z_t}\right\}\right) \\ &\le \mathcal{O}\left(\min\left\{\sqrt{\log T \sum_{t=1}^{T} h_t z_t} + \sqrt{h_{\max} z_{\max}}, \sqrt{h_{\max} \sum_{t=1}^{T} z_t}\right\}\right),\end{aligned}$$

which concludes the proof. ■

C.7. Proof of Lemma 11

Proof For $\alpha \in (1,2)$, it is clear that

$$\beta_t \ge \beta_1 + \frac{4}{2^{1/\alpha}-1} \sum_{s=1}^{t-1} \frac{1}{s} \ge \beta_1 + \frac{4}{2^{1/\alpha}-1} \log t.$$

For $\alpha \ge 2$, by definition of z_s and h_s in (9), we have for $\alpha \ge 2$

$$\frac{z_s}{h_s} = \frac{\alpha \sum_{i\neq j_s} q_{s,i}^{1/\alpha}}{\frac{\alpha}{\alpha-1} \sum_{i\neq j_s} q_{s,i}^{1-1/\alpha}} = (\alpha-1) \frac{\sum_{i\neq j_s} q_{s,i}^{1/\alpha}}{\sum_{i\neq j_s} q_{s,i}^{1-1/\alpha}} \ge 1.$$

Since $q_{t,i} \in (0,1)$ for $i \neq j_t$ by the choice of $q_{t,i}$ in (7) and $\frac{1}{\alpha} \le 1 - \frac{1}{\alpha}$, i.e., $q_{t,i}^{1/\alpha} \ge q_{t,i}^{1-1/\alpha}$ always holds. Note that whenever $\beta_{t+1} = 2^{1/\alpha}\beta_t$ occurs, this means $\beta_{t+1} = \beta_t + (2^{1/\alpha}-1)\beta_t$. Since $\beta_1 = 2\alpha K^{\frac{1}{2}-\frac{1}{\alpha}}$ and $2^{1/\alpha} - 1 \ge \frac{\log 2}{\alpha}$, the increment is always larger than $2\log 2$. Therefore, it holds that $\beta_{t+1} \ge \beta_t + \frac{1}{\beta_t}$ for any $t \in \mathbb{N}$. This implies that $\beta_t \ge \sqrt{\beta_1^2 + 2\sum_{s=1}^{t-1} 1} \ge \sqrt{2t}$. ■

Appendix D. Proof of Theorem 4

In this section, we prove the BOBW guarantee of Algorithm 1.

D.1. Adversarial regime

In this regime, it suffices to show that $h_{\max}\sum_{t=1}^{T} z_t$ is at most KT from the second term in (13). By definition, for any t, it holds that

$$z_t = \alpha \sum_{i \neq j_{t+1}} q_{t+1,i}^{1/\alpha} \leq \sum_{i \neq j_{t+1}} \frac{\alpha}{\sigma_{t+1,i}^{1/\alpha}}$$
$$= \sum_{n=2}^{K} \frac{\alpha}{n^{1/\alpha}} \leq \frac{\alpha^2}{\alpha - 1}((K+1)^{1-1/\alpha} - 1) \leq \frac{\alpha^2}{\alpha - 1} K^{1-1/\alpha}, \tag{27}$$

and

$$h_t = \frac{\alpha}{\alpha - 1} \sum_{i \neq j_{t+1}} q_{t+1,i}^{1-1/\alpha} \leq \frac{\alpha}{\alpha - 1} \sum_{i \neq j_{t+1}} \frac{1}{\sigma_{t+1,i}^{1-1/\alpha}}$$
$$= \frac{\alpha}{\alpha - 1} \sum_{n=2}^{K} \frac{1}{n^{1-1/\alpha}} \leq \frac{\alpha^2((K+1)^{1/\alpha} - 1)}{\alpha - 1} \leq \frac{\alpha^2}{\alpha - 1} K^{1/\alpha}. \tag{28}$$

Therefore,

$$h_{\max} \sum_{t=1}^{T} z_t \leq \sum_{t=1}^{T} \left(\frac{\alpha^2}{\alpha - 1}\right)^2 K = \left(\frac{\alpha^2}{\alpha - 1}\right)^2 KT,$$

which concludes the proof for the adversarial regime.

D.2. Adversarial regime with self-bounding constraint

To analyze the regret in this regime, we introduce the event $\mathcal{D}_{t,\alpha}$ defined by

$$\mathcal{D}_{t,\alpha} := \left\{ \sum_{i \neq j_t} \frac{1}{(2^{1/\alpha} + \eta_t \underline{\hat{L}}_{t,i})^\alpha} \leq \frac{1}{2} \right\},$$

which is a slightly modified version of $\mathcal{E}_{t,\alpha}$. This definition is to utilize the previous results in Kim et al. (2026), while there is a subtle difference due to j_t parts instead of i^*. On this event, we have the following lemma, which slightly improves Lemma 10 in Kim et al. (2026) in the constant factor.

Lemma 14 *For Algorithm 1 with $\alpha > 1$, it holds that on $\mathcal{D}_{t,\alpha}$*

$$w_{t,i} \geq \frac{1}{8} \frac{1}{(1 + \eta_t \underline{\hat{L}}_{t,i})^\alpha}, \forall i \neq j_t.$$

In addition, the current best arm satisfies $\frac{1}{4} \leq w_{t,j_t}$ on $\mathcal{D}_{t,\alpha}$.

While Kim et al. (2026) considered the case $j_t = i^*$, the following results can be directly obtained.

Lemma 15 (Lemma 11 in Kim et al. (2026)) *On $\mathcal{D}_{t,\alpha}^c$, $w_{t,j_t} \leq \frac{1+e^{-1/2}}{2}$.*

The implication of $\mathcal{D}_{t,\alpha}$ is that we can link the $q_{t,i}$ and $w_{t,i}$ on this event, which will be not rare when the self-bounding constraint is small, i.e., as close as to stochastic settings. Specifically, we have

$$w_{t,i} \geq \frac{1}{8}\frac{1}{(1+\eta_t \underline{\hat{L}}_{t,i})^\alpha} \geq \frac{p_{t,i}}{8} \implies w_{t,i} \in \left[\frac{p_{t,i}}{8}, p_{t,i}\right], \text{ and } p_{t,i} \leq 8 w_{t,i}.$$

Then, by Lemma 8, we have

$$q_{t,i} \leq 16 w_{t,i}. \tag{29}$$

Then, similarly to the recent BOBW analysis of FTPL (Honda et al., 2023; Lee et al., 2024), we consider the analysis on $\mathcal{D}_{t,\alpha}$ and $\mathcal{D}^c_{t,\alpha}$ separately. In terms of Lemma 13, what we will show is that

$$\sum_{t=1}^{T-1} \frac{z_t}{\beta_t} \lesssim \sqrt{\log T \left(\sum_{t=1}^{T-1} \mathbb{1}[\mathcal{D}_{t+1,\alpha}] h_t z_t + \sum_{t=1}^{T-1} \mathbb{1}[\mathcal{D}^c_{t+1,\alpha}] h_t z_t \right)} + \mathcal{O}\left(\frac{\alpha^2}{\alpha - 1} \sqrt{K} \right).$$

Note that the last term is directly obtained by the results (27) and (28) in adversarial regime, where we showed that

$$\sqrt{h_{\max} z_{\max}} \leq \frac{\alpha^2}{\alpha - 1} \sqrt{K}.$$

On $\mathcal{D}_{t,\alpha}$. Note that $w_{t,j_t} \geq w_{t,i}$ for any $i \in [K]$ and $t \in \mathbb{N}$. By Hölder's inequality, we have

$$\begin{aligned}
\mathbb{1}[\mathcal{D}_{t,\alpha}] z_{t-1} = \mathbb{1}[\mathcal{D}_{t,\alpha}] \alpha \sum_{i \neq j_t} q_{t,i}^{1/\alpha} &\leq \mathbb{1}[\mathcal{D}_{t,\alpha}] \alpha 16^{1/\alpha} \sum_{i \neq j_t} w_{t,i}^{1/\alpha} && \text{(by (29))} \\
&\leq \mathbb{1}[\mathcal{D}_{t,\alpha}] \alpha 16^{1/\alpha} \sum_{i \neq i^*} w_{t,i}^{1/\alpha} && (w_{t,j_t} \geq w_{t,i}, \forall i) \\
&= \mathbb{1}[\mathcal{D}_{t,\alpha}] \alpha 16^{1/\alpha} \sum_{i \neq i^*} \frac{1}{\Delta_i^{1/\alpha}} (\Delta_i w_{t,i})^{1/\alpha} \\
&\leq \mathbb{1}[\mathcal{D}_{t,\alpha}] \alpha 16^{1/\alpha} \left(\sum_{i \neq i^*} \frac{1}{\Delta_i^{1/(\alpha-1)}} \right)^{1-\frac{1}{\alpha}} \left(\sum_{i \neq i^*} \Delta_i w_{t,i} \right)^{1/\alpha}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{1}[\mathcal{D}_{t,\alpha}] h_{t-1} &= \mathbb{1}[\mathcal{D}_{t,\alpha}] \frac{\alpha}{\alpha - 1} \sum_{i \neq j_t} q_{t,i}^{1-1/\alpha} \\
&\leq \mathbb{1}[\mathcal{D}_{t,\alpha}] \frac{\alpha}{\alpha - 1} 16^{1-1/\alpha} \sum_{i \neq j_t} w_{t,i}^{1-1/\alpha} \\
&\leq \mathbb{1}[\mathcal{D}_{t,\alpha}] \frac{\alpha}{\alpha - 1} 16^{1-1/\alpha} \sum_{i \neq i^*} w_{t,i}^{1-1/\alpha} \\
&= \mathbb{1}[\mathcal{D}_{t,\alpha}] \frac{\alpha}{\alpha - 1} 16^{1-1/\alpha} \sum_{i \neq i^*} \frac{1}{\Delta_i^{1-1/\alpha}} (\Delta_i w_{t,i})^{1-1/\alpha} \\
&\leq \mathbb{1}[\mathcal{D}_{t,\alpha}] \frac{\alpha}{\alpha - 1} 16^{1-1/\alpha} \left(\sum_{i \neq i^*} \frac{1}{\Delta_i^{\alpha-1}} \right)^{\frac{1}{\alpha}} \left(\sum_{i \neq i^*} \Delta_i w_{t,i} \right)^{1-1/\alpha}.
\end{aligned} \tag{30}$$

Therefore,

$$\mathbb{1}[\mathcal{D}_{t+1,\alpha}]h_t z_t \le \mathbb{1}[\mathcal{D}_{t+1,\alpha}]\omega'(\Delta)\left\langle \Delta, w_{t+1}\right\rangle,$$

where

$$\omega'(\Delta) = \frac{16\alpha^2}{\alpha-1}\left(\sum_{i\neq i^*}\Delta_i^{-\frac{1}{\alpha-1}}\right)^{1-\frac{1}{\alpha}}\left(\sum_{i\neq i^*}\Delta_i^{1-\alpha}\right)^{\frac{1}{\alpha}}.$$

On $\mathcal{D}_{t,\alpha}^c$. In this case, we have

$$\begin{aligned}\mathbb{1}[\mathcal{D}_{t,\alpha}^c]h_{t-1}z_{t-1} \le \mathbb{1}[\mathcal{D}_{t,\alpha}^c]\frac{\alpha^2}{\alpha-1}\sum_{i\neq j_t}q_{t,i}^{1/\alpha}\cdot\sum_{i\neq j_t}q_{t,i}^{1-1/\alpha} &\le \mathbb{1}[\mathcal{D}_{t,\alpha}^c]\frac{\alpha^2}{\alpha-1}\frac{\alpha}{\alpha-1}K^{1-1/\alpha}\cdot\alpha K^{1/\alpha}\\ &= \mathbb{1}[\mathcal{D}_{t,\alpha}^c]\left(\frac{\alpha^2}{\alpha-1}\right)^2 K.\end{aligned}$$

Derivation of the desired results. Therefore, we obtain that

$$\begin{aligned}\sum_{t=1}^{T-1}\frac{z_t}{\beta_t} \lesssim \sqrt{\log T\left(\sum_{t=1}^{T}\mathbb{1}[\mathcal{D}_{t,\alpha}]\omega'(\Delta)\left\langle\Delta,w_t\right\rangle + \sum_{t=1}^{T}\mathbb{1}[\mathcal{D}_{t,\alpha}^c]\left(\frac{\alpha^2}{\alpha-1}\right)^2 K\right)}\\ + \mathcal{O}\left(\frac{\alpha^2}{\alpha-1}\sqrt{K}\right).\end{aligned}$$

Combined with the regret upper bounds in Lemmas 2 and 3, we obtain for $\alpha\in(1,2)$ that

$$\begin{aligned}\mathrm{Reg}_{\mathrm{FTPL}}(T) \lesssim \mathcal{O}\left(\sqrt{\log T\left(\sum_{t=1}^{T}\omega'(\Delta)\left\langle\Delta,w_t\right\rangle + \sum_{t=1}^{T}\mathbb{1}[\mathcal{D}_{t,\alpha}^c]\left(\frac{\alpha^2}{\alpha-1}\right)^2 K\right)}\right)\\ + \mathcal{O}\left(\frac{\alpha^2}{\alpha-1}K^{1/\alpha}\log T\right) + \mathcal{O}\left(\frac{\alpha^3\sqrt{K}}{(\alpha-1)^2}\right). \quad (31)\end{aligned}$$

and for $\alpha\ge 2$ that

$$\begin{aligned}\mathrm{Reg}_{\mathrm{FTPL}}(T) \lesssim \mathcal{O}\left(\sqrt{\log T\left(\sum_{t=1}^{T}\omega'(\Delta)\left\langle\Delta,w_t\right\rangle + \sum_{t=1}^{T}\mathbb{1}[\mathcal{D}_{t,\alpha}^c]\left(\frac{\alpha^2}{\alpha-1}\right)^2 K\right)}\right)\\ + \mathcal{O}\left(\frac{\alpha^3}{(\alpha-1)^2}\sqrt{K}\right) + t_0(\alpha,K). \quad (32)\end{aligned}$$

Therefore, it remains to control the term related to $\sum_t \mathbb{1}[\mathcal{D}_{t,\alpha}^c]$. Recall that, in the adversarial regime with self-bounding constraint (Δ, C, T), the regret satisfies that

$$\mathrm{Reg}(T) \ge \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\Delta,w_t\right\rangle\right] - C.$$

On $\mathcal{D}_{t,\alpha}^c$, Lemma 15 shows that

$$\begin{aligned}
\mathbb{1}[\mathcal{D}_{t,\alpha}^c] \sum_{i=1}^K \Delta_i w_{t,i} = \mathbb{1}[\mathcal{D}_{t,\alpha}^c] \sum_{i \neq i^*} \Delta_i w_{t,i} &\geq \mathbb{1}[\mathcal{D}_{t,\alpha}^c] \Delta_{\min} \sum_{i \neq i^*} w_{t,i} \\
&\geq \mathbb{1}[\mathcal{D}_{t,\alpha}^c] \Delta_{\min} \sum_{i \neq j_t} w_{t,i} \qquad (w_{t,j_t} \geq w_{t,i}, \forall i) \\
&= \mathbb{1}[\mathcal{D}_{t,\alpha}^c] \Delta_{\min} (1 - w_{t,j_t}) \\
&\geq \mathbb{1}[\mathcal{D}_{t,\alpha}^c] \Delta_{\min} \frac{1 - e^{-1/2}}{2}.
\end{aligned}$$

This implies that

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[\mathcal{D}_{t,\alpha}^c] 0.31 \Delta_{\min}\right] \leq \mathrm{Reg}(T) + C \implies \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}[\mathcal{D}_{t,\alpha}^c]\right] \leq \frac{\mathrm{Reg}(T) + C}{0.31 \Delta_{\min}}.$$

By applying this result into (31) and (32), the regret can be upper bound in the form of

$$\mathrm{Reg}_{\mathrm{FTPL}}(T) \leq \mathcal{O}\left(\sqrt{\log T \left(\omega'(\Delta) + \frac{K}{\Delta_{\min}}\right)(\mathrm{Reg}(T) + C)}\right) + \mathcal{O}(\log T). \tag{33}$$

Therefore, we have

$$\begin{aligned}
\mathrm{Reg}_{\mathrm{FTPL}}(T) \leq \mathcal{O}\left(\omega(\Delta) \log T + \sqrt{C \omega(\Delta) \log T}\right) + \mathcal{O}\left(\frac{\alpha^3 \sqrt{K}}{(\alpha - 1)^2}\right) \\
+ \mathbb{1}[\alpha \geq 2] t_0(\alpha, K) + \mathbb{1}[\alpha \in (1, 2)] 2,
\end{aligned}$$

where

$$\begin{aligned}
\omega(\Delta) &= \omega'(\Delta) + \frac{\alpha^2 K}{(\alpha - 1) 0.31 \Delta \min} + \mathbb{1}[\alpha \in (1, 2)] \frac{\alpha^2}{\alpha - 1} K^{1/\alpha} \\
&= \mathcal{O}\left(\frac{8\alpha^2}{\alpha - 1} \left(\sum_{i \neq i^*} \Delta_i^{-\frac{1}{\alpha - 1}}\right)^{1 - \frac{1}{\alpha}} \left(\sum_{i \neq i^*} \Delta_i^{1 - \alpha}\right)^{\frac{1}{\alpha}} + \frac{\alpha^4 K}{(\alpha - 1)^2 0.31 \Delta_{\min}}\right).
\end{aligned}$$

Note that $\omega(\Delta) \leq \mathcal{O}\left(\frac{\alpha^4 K}{(\alpha - 1)^2 \Delta_{\min}}\right)$, whose α-dependency is square of Ito et al. (2024), which can be seen as a drawback of using surrogate instead of explicit probability. For the comparison with Ito et al. (2024), one can see that the relationship between γ-Tsallis entropy and Fréchet-type perturbation with shape α, where $\alpha \approx \frac{1}{1 - \gamma}$ correspondence observed. Therefore, $\omega(\Delta)$ can be seen as the results with $\mathcal{O}\left(\frac{1}{(\gamma(1 - \gamma))^2 \Delta_{\min}}\right)$.

D.3. Bias term by CGR II

So far, we derived the upper bounds of $\mathrm{Reg}_{\mathrm{FTPL}}$, which is the main leading term of the regret. In this section, we show that $\mathrm{Reg}_{\mathrm{CGR}}$ is at most $\log T$, which does not affect the overall regret.

We start from Lemma 7 of Chen et al. (2025) that showed that the expected regret of FTPL with CGR II satisfies

$$\mathrm{Reg}(T) \leq \sum_{t=1}^{T} \mathbb{E}\left[\left\langle \hat{\ell}_t, w_t - e_{i^*} \right\rangle\right] + \sum_{t=1}^{T}\sum_{i=1}^{K} \mathbb{E}\left[w_{t,i} \left(1 - \frac{w_{t,i}}{\Pr\left[\mathcal{A}_t \middle| \hat{L}_t, i_t = i\right]} \right)^{G_t} \right],$$

where the second term was denoted by $\mathrm{Reg}_{\mathrm{CGR}}(T)$. Here, recall the definition of $\mathcal{A}_t$ in (13), which is

$$\mathcal{A}_t = \left\{ r'_{t,i_t} = \max_{i:\sigma_{t,i} \leq \sigma_{t,i_t}} r'_{t,i},\ r'_{t,i_t} \geq \eta_t \underline{\hat{L}}_{t,i_t} \right\}.$$

Lemma 16 *Algorithm 1 satisfies that*

$$\sum_{t=1}^{T}\sum_{i=1}^{K} \mathbb{E}\left[w_{t,i} \left(1 - \frac{w_{t,i}}{\Pr\left[\mathcal{A}_t \middle| \hat{L}_t, i_t = i\right]} \right)^{G_t} \right] \leq \log T.$$

Proof Similarly to the proof in Chen et al. (2025), where they consider the Fréchet distribution, we denote $\Pr\left[\mathcal{A}_t \middle| \hat{L}_t, i_t = i\right]$ by $\Pr[\mathcal{A}_{t,i}]$ in this proof. Then, it is sufficient to prove

$$\exp\left(-\frac{w_{t,i}}{\Pr[\mathcal{A}_{t,i}]} G_t \right) \leq \frac{1}{t}.$$

Let $\underline{\lambda}_t = \eta_t \underline{\hat{L}}_t$. By definition of $\mathcal{A}_{t,i}$, it holds that

$$\Pr[\mathcal{A}_{t,i}] = \int_{\underline{\lambda}_{t,i}}^{\infty} f(z) F^{\sigma_{t,i}-1}(z) \mathrm{d}z = \frac{1 - F^{\sigma_{t,i}}(\underline{\lambda}_{t,i})}{\sigma_{t,i}} \leq 1 - F(\underline{\lambda}_{t,i}).$$

Then, we consider the lower bound of $w_{t,i}$. Let $r^{\max}_{t,-i} := \max_{j \neq i} r_{t,j}$ and define an event,

$$\mathcal{B}_{t,i} := \left\{ r_{t,i} \geq \underline{\lambda}_{t,i} + r^{\max}_{t,-i} \right\},$$

which is the sufficient condition for $\{i_t = i\}$. This is because

$$\begin{aligned}
\{i_t = i\} &= \left\{ \arg\min_{j \in [K]} \left\{ \underline{\lambda}_{t,j} - r_{t,j} \right\} = i \right\} \\
&= \left\{ \arg\max_{j \in [K]} \left\{ r_{t,j} - \underline{\lambda}_{t,j} \right\} = i \right\} \\
&= \left\{ \forall j \neq i : r_{t,i} \geq r_{t,j} + \underline{\lambda}_{t,i} - \underline{\lambda}_{t,j} \right\} \\
&\supseteq \left\{ \forall j \neq i : r_{t,i} \geq r_{t,j} + \underline{\lambda}_{t,i} \right\} = \mathcal{B}_{t,i}.
\end{aligned}$$

Therefore, $w_{t,i} \geq \Pr[\mathcal{B}_{t,i} | \lambda_t]$. Let $\bar{\mathcal{P}}^K_\alpha$ denotes the distribution of $K-1$ block maximum for Pareto distributed random variables, i.e., the distribution of $r^{\max}_{t,-i}$. Here, by definition of $\mathcal{B}_{t,i}$ and definition

of $\bar{\mathcal{P}}_{\alpha,K}$, we have

$$\begin{aligned}\Pr[\mathcal{B}_{t,i}|\lambda_t] = \mathbb{E}_{r^{\max}_{t,-i}\sim\bar{\mathcal{P}}^K_\alpha}\left[\Pr\left(r_{t,i} \geq \underline{\lambda}_{t,i} + r^{\max}_{t,-i} | r^{\max}_{t,-i}\right)\right] &= \mathbb{E}_{r^{\max}_{t,-i}}[1 - F(\underline{\lambda}_{t,i} + r^{\max}_{t,-i})] \\ &= \mathbb{E}_{r^{\max}_{t,-i}}\left[\frac{1}{(1+\underline{\lambda}_{t,i} + r^{\max}_{t,-i})^\alpha}\right] \\ &\geq \mathbb{E}_{r^{\max}_{t,-i}}\left[\frac{1}{(1+\underline{\lambda}_{t,i})^\alpha(1 + r^{\max}_{t,-i})^\alpha}\right] \\ &= (1 - F(\underline{\lambda}_{t,i})\mathbb{E}_{r^{\max}_{t,-i}}[1 - F(r^{\max}_{t,-i})].\end{aligned} \tag{34}$$

Here,

$$\mathbb{E}_{r^{\max}_{t,-i}}[1 - F(r^{\max}_{t,-i})] = \Pr\left[r_{t,i} \text{ is the maximum among } \{r_{t,j}\}_{j\in[K]}\right] = \frac{1}{K}.$$

Since $r_{t,i}$s are i.i.d. from the same perturbations, the probability that $r_{t,i}$ is not the maximum over K samples becomes $1 - 1/K$. Therefore,

$$w_{t,i} \geq \frac{1 - F(\underline{\lambda}_{t,i})}{K},$$

which implies that

$$\frac{\Pr[\mathcal{A}_{t,i}]}{w_{t,i}} \leq K \implies \exp\left(-\frac{w_{t,i}}{\Pr[\mathcal{A}_{t,i}]}G_t\right) \leq \exp\left(-\frac{G_t}{K}\right) = \frac{1}{t}.$$

When $\alpha \in (1,2)$. In this case, we sometimes set $G_t = 2\log t$ instead of $K\log t$. The precise condition is when $i_t = j_t$ and $\mathcal{E}_{t,\alpha}$ in (6) occurs. As we mentioned in Section 3.1, the definition of $\mathcal{E}_{t,\alpha}$ denotes the case when $w_{t,j_t} \geq 1/2$. To be precise, $\mathcal{E}_{t,\alpha}$ means the case when $\sum_{i\neq j_t}\frac{1}{(1+\eta_t\underline{\hat{L}}_{t,i})^\alpha} \leq \frac{1}{2}$. Since $w_{t,i} \leq \frac{1}{(1+\eta_t\underline{\hat{L}}_{t,i})^\alpha}$ for any i and t (see, Appendix C.1), this implies that

$$\left\{\sum_{i\neq j_t} w_{t,i} \leq \frac{1}{2}\right\} = \left\{w_{t,j_4 t} > \frac{1}{2}\right\} \supset \mathcal{E}_{t,\alpha}.$$

Since $\Pr[\mathcal{A}_{t,j_t}] \leq 1$ and $w_{t,j_t} \geq 1/2$, $G_t = 2\log t$ is still valid to obtain the desired result. ■

Remark 17 *In the current analysis, the only part that depends on the specific form of the Pareto distribution is in (34). We expect that this argument can be extended to more general Fréchet-type distributions. Indeed, the tail function of a Fréchet-type distribution can be expressed as $x^{-\alpha}S_F(x)$ for some slowly varying function S_F, which means that the tail function can be written as $S_F(x)(x+1)^{-\alpha}$ for the shifted Fréchet-type distribution considered in Lee et al. (2024, Eq. (7)). Therefore, as long as $S_F(x)$ admits a uniform lower bound by a positive constant, incorporating this constant into the choice of G_t will suffice to obtain the same results.*

D.4. Proof of Lemma 14

Proof By definition of $w_{t,i}$, it holds that for $i \neq j_t$

$$
\begin{aligned}
w_{t,i} &= \int_0^\infty f(z+\eta_t \underline{\hat{L}}_{t,i}) \prod_{j\neq i} F(z+\eta_t \underline{\hat{L}}_{t,j}) \mathrm{d}z \\
&= \int_0^\infty f(z+\eta_t \underline{\hat{L}}_{t,i}) F(z) \prod_{j\neq i,j_t} F(z+\eta_t \underline{\hat{L}}_{t,j}) \mathrm{d}z && (\because \underline{\hat{L}}_{t,j_t} = 0 \text{ on } \mathcal{D}_{t,\alpha},\ i \neq j_t) \\
&\geq \int_{2^{1/\alpha}-1}^\infty f(z+\eta_t \underline{\hat{L}}_{t,i}) F(z) \prod_{j\neq i,j_t} F(z+\eta_t \underline{\hat{L}}_{t,j}) \mathrm{d}z \\
&\geq \frac{1}{2} \int_{2^{1/\alpha}-1}^\infty f(z+\eta_t \underline{\hat{L}}_{t,i}) \prod_{j\neq i,j_t} F(z+\eta_t \underline{\hat{L}}_{t,j}) \mathrm{d}z \\
&= \frac{1}{2} \int_{2^{1/\alpha}-1}^\infty f(z+\eta_t \underline{\hat{L}}_{t,i}) \prod_{j\neq i,j_t} \left(1 - \frac{1}{(z+\eta_t \underline{\hat{L}}_j + 1)^\alpha}\right) \mathrm{d}z \\
&\geq \frac{1}{2} \int_{2^{1/\alpha}-1}^\infty f(z+\eta_t \underline{\hat{L}}_{t,i}) \left(1 - \sum_{j\neq i,j_t} \frac{1}{(z+\eta_t \underline{\hat{L}}_j + 1)^\alpha}\right) \mathrm{d}z && (\textstyle\prod_i (1-x_i) \geq 1 - \sum_i x_i) \\
&\geq \frac{1}{2} \int_{2^{1/\alpha}-1}^\infty f(z+\eta_t \underline{\hat{L}}_{t,i}) \left(1 - \sum_{j\neq i,j_t} \frac{1}{(2^{1/\alpha} + \eta_t \underline{\hat{L}}_j)^\alpha}\right) \mathrm{d}z \\
&\geq \frac{1}{4} \int_{2^{1/\alpha}-1}^\infty f(z+\eta_t \underline{\hat{L}}_{t,i}) \mathrm{d}z = \frac{1}{4} \frac{1}{(2^{1/\alpha} + \eta_t \underline{\hat{L}}_{t,j})^\alpha}.
\end{aligned}
$$

Since $\frac{(x+1)^\alpha}{(x+2^{1/\alpha})^\alpha}$ is increasing with respect to $x \geq 0$ for any $\alpha > 1$, we have

$$
\frac{(1+\eta_t \underline{\hat{L}}_{t,i})^\alpha}{(2^{1/\alpha} + \eta_t \underline{\hat{L}}_{t,j})^\alpha} \geq \frac{1}{2},
$$

which concludes the proof for $i \neq j_t$. For $i = j_t$, we have

$$
\begin{aligned}
w_{t,j_t} &= \int_0^\infty \frac{\alpha}{(z+1)^{\alpha+1}} \prod_{i\neq j_t} \left(1 - \frac{1}{(z+\eta_t \underline{\hat{L}}_{t,j} + 1)^\alpha}\right) \mathrm{d}z \\
&\geq \int_{2^{1/\alpha}-1}^\infty \frac{\alpha}{(z+1)^{\alpha+1}} \prod_{i\neq j_t} \left(1 - \frac{1}{(z+\eta_t \underline{\hat{L}}_{t,j} + 1)^\alpha}\right) \mathrm{d}z \\
&\geq \int_{2^{1/\alpha}-1}^\infty \frac{\alpha}{(z+1)^{\alpha+1}} \left(1 - \sum_{i\neq j_t} \frac{1}{(2^{1/\alpha} + \eta_t \underline{\hat{L}}_{t,j})^\alpha}\right) \mathrm{d}z \\
&\geq \frac{1}{2} \int_{2^{1/\alpha}-1}^\infty \frac{\alpha}{(z+1)^{\alpha+1}} \mathrm{d}z = \frac{1}{4},
\end{aligned}
$$

which concludes the proof. ■

Appendix E. Proofs of Lemmas in bandit problems with expert advices

In this section, we provide the proofs for Lemma 5, which is required to prove BOBW guarantee in the contextual bandit settings. To prove this lemma, we need the following results.

Lemma 18 *For any $t \in \mathbb{N}$, Algorithm 2 with $\alpha > 1$ satisfies that for any $i \in [K]$*

$$\mathbb{E}\left[\hat{\ell}_{t,i}(\phi_i(\eta_t \hat{L}_t) - \phi_i(\eta_t(\hat{L}_t + \hat{\ell}_t))\middle| \hat{L}_t\right] \leq \sum_{a=1}^{N} \mathbb{E}\left[\frac{2e\alpha}{\beta_t} p_{t,i}^{1/\alpha} \frac{w_{t,i}\pi_{t,i,a}}{P_{t,a}}\middle| \hat{L}_t\right].$$

Note that $\hat{\ell}_{t,i} \neq 0$ is possible even when $i_t \neq i$ in contextual setting since $\hat{\ell}_{t,i}$ can be updated by using ℓ_{t,a_t} and π_{t,i,a_t}.

Lemma 19 *For any $t \in \mathbb{N}$, Algorithm 2 with $\alpha \geq 2$ satisfies that*

$$\mathbb{E}\left[\hat{\ell}_{t,j_t}(\phi_{j_t}(\eta_t \hat{L}_t) - \phi_{j_t}(\eta_t(\hat{L}_t + \hat{\ell}_t))\middle| \hat{L}_t\right] \leq \sum_{a=1}^{N} \mathbb{E}\left[\sum_{i \neq j_t} \mathcal{O}\left(\frac{\alpha}{\beta_t} p_{t,i}^{1/\alpha}\right) \frac{w_{t,j_t}\pi_{t,j_t,a}}{P_{t,a}}\middle| \hat{L}_t\right] + g_t(\alpha;\nu),$$

where $g_t(\alpha)$ is a function such that $\sum_t g_t(\alpha;\nu) = \mathcal{O}(\alpha^2/\nu)$.

Proof of Lemma 5 From Lemmas 18 and 19, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \hat{\ell}_t, \phi(\eta_t \hat{L}_t) - \phi(\eta_t \hat{L}_{t+1})\right\rangle \middle| \hat{L}_t\right] \\
&\leq \sum_{i=1}^{K} \mathbb{E}\left[\hat{\ell}_{t,i}(\phi_i(\eta_t \hat{L}_t) - \phi_i(\eta_t \hat{L}_t + \hat{\ell}_{t,i} e_i))\middle| \hat{L}_t\right] \\
&\leq \sum_{a=1}^{N} \sum_{i \neq j_t} \mathbb{E}\left[\frac{2e\alpha}{\beta_t} p_{t,i}^{1/\alpha} \frac{w_{t,i}\pi_{t,i,a}}{P_{t,a}}\middle| \hat{L}_t\right] + \sum_{a=1}^{N} \sum_{i \neq j_t} \mathbb{E}\left[\mathcal{O}\left(\frac{\alpha}{\beta_t}\right) p_{t,i}^{1/\alpha} \frac{w_{t,j_t}\pi_{t,j_t,a}}{P_{t,a}}\middle| \hat{L}_t\right] + g_t(\alpha;\nu) \\
&\leq \sum_{a=1}^{N} \mathbb{E}\left[\mathcal{O}\left(\frac{\alpha}{\beta_t}\right) \max_{j \neq j_t} p_{t,j}^{1/\alpha} \cdot \frac{\sum_{i=1}^{K} w_{t,i}\pi_{t,i,a}}{P_{t,a}}\middle| \hat{L}_t\right] + g_t(\alpha;\nu) \\
&\leq \mathcal{O}\left(\frac{\alpha N}{\beta_t}\right) \max_{i \neq j_t} p_{t,i}^{1/\alpha} + g_t(\alpha;\nu),
\end{aligned}$$

which concludes the proof. ■

E.1. Proof of Lemma 18

Proof While the loss estimators can be updated for all experts $i \in \{j \in [K] : \pi_{t,j,a_t} > 0\}$ in the contextual setting, we can apply the intermediate results in Lemma 7. The main observation is the increasing property of $\phi_i(\lambda)$ with respect to λ_j for $i \neq j$, which is obvious from (1) with

i.i.d. perturbations. Therefore, for any $i \in [K]$, we have

$$\begin{aligned}
\mathbb{E}\Big[\hat{\ell}_{t,i}\Big(\phi_i(\eta_t \hat{L}_t) - \phi_i(\eta_t \hat{L}_{t+1})\Big)\Big|\hat{L}_t\Big] &\le \mathbb{E}\Big[\hat{\ell}_{t,i}\Big(\phi_i(\eta_t \hat{L}_t) - \phi_i(\eta_t(\hat{L}_t + \hat{\ell}_{t,i}e_i))\Big)\Big|\hat{L}_t\Big] \\
&\le \mathbb{E}\Big[-\eta_t \hat{\ell}_{t,i}^2 \phi_i'(\eta_t \hat{L}_t)\Big|\hat{L}_t\Big] \qquad \text{(by (19))} \\
&\le \mathbb{E}\Big[-\eta_t \ell_{t,a_t}^2 M_t^2 \pi_{t,i,a_t}^2 \phi_i'(\eta_t \hat{L}_t)\Big|\hat{L}_t\Big] \\
&\le \mathbb{E}\left[-2\eta_t \pi_{t,i,a_t}^2 \frac{\phi_i'(\eta_t \hat{L}_t)}{P_{t,a_t}^2}\middle|\hat{L}_t\right] \qquad \text{(GR and } \ell_t \in [0,1]^N\text{)} \\
&= \mathbb{E}\left[-2\eta_t \phi_i'(\eta_t \hat{L}_t)\sum_{a=1}^N \frac{\pi_{t,i,a}^2}{P_{t,a}}\middle|\hat{L}_t\right] \\
&\le \mathbb{E}\left[-2\eta_t \phi_i'(\eta_t \hat{L}_t)\sum_{a=1}^N \frac{\pi_{t,i,a}}{P_{t,a}}\middle|\hat{L}_t\right] \\
&= \mathbb{E}\left[-2\eta_t \frac{\phi_i'(\eta_t \hat{L}_t)}{w_{t,i}}\sum_{a=1}^N \frac{w_{t,i}\pi_{t,i,a}}{P_{t,a}}\middle|\hat{L}_t\right] \\
&\le \sum_{a=1}^N \mathbb{E}\left[2e\alpha\eta_t p_{t,i}^{1/\alpha}\frac{w_{t,i}\pi_{t,i,a}}{P_{t,a}}\middle|\hat{L}_t\right],
\end{aligned}$$

where the last inequality follows from the results on the ratio $-\phi_i/\phi_i$ in Appendix C.2. ■

E.2. Proof of Lemma 19

Since Algorithm 2 adopts naive GR, $\beta_t \ge \sqrt{t}$, and $\alpha \ge 2$, we can utilize Lemma 11 in Honda et al. (2023), given as follows.

Lemma 20 (Partial results of Lemma 11 in Honda et al. (2023)) *For any $\hat{L}_t \in \mathbb{R}^K$, $\zeta \in (0,1)$ and $i \in [K]$, if $w_{t,i} \ge w'$ for some fixed constant w', then it holds that*

$$\left[\mathbb{1}[\hat{\ell}_{t,i} > \zeta\beta_t]\hat{\ell}_{t,i}\middle|\hat{L}_t\right] \le \frac{1}{(1-w')}(1-w')^{\zeta\beta_t}(\zeta\beta_t + 1/w').$$

Moreover, when $\beta_t \ge a\sqrt{t}$ for some $a > 0$, it holds that

$$\sum_{t=1}^{\infty}\frac{1}{(1-w')}(1-w')^{\zeta\beta_t}(\zeta\beta_t + 1/w') \le \mathcal{O}(1/a^2).$$

Proof of Lemma 19 Similarly to the proof of Lemma 10, we used the events for $\alpha \ge 2$ defined by

$$\bar{\mathcal{E}}_{t,\alpha} := \left\{\sum_{i\neq j_t}\frac{1}{(1+\eta_t \underline{\hat{L}}_{t,i})^\alpha} \le \xi_\alpha\right\}.$$

for some $\xi_\alpha \in (0,1)$ specified later.

Then, following the same steps in Appendix C.5.1, we can obtain that

$$\mathbb{1}[\bar{\mathcal{E}}_{t,\alpha}^c] p_{t,j_t}^{1/\alpha} = \mathbb{1}[\bar{\mathcal{E}}_{t,\alpha}^c] 1 \leq \frac{2}{\xi_\alpha} \sum_{i \neq j_t} p_{t,i}^{\frac{1}{\alpha}}.$$

On $\bar{\mathcal{E}}_{t,\alpha}$, we have $\eta_t \underline{\hat{L}}_{t,i} \geq \xi_\alpha^{-1/\alpha} - 1 > 0$. Therefore, by the same trick in Lemma 18, where we bound the stability term of j_t in contextual bandits by that in multi-armed bandits, we obtain that

$$\mathbb{E}\Big[\mathbb{1}[\bar{\mathcal{E}}_{t,\alpha}^c \cup \{\bar{\mathcal{E}}_{t,\alpha}, \hat{\ell}_{t,j_t} \leq \zeta_\alpha \beta_t\}] \hat{\ell}_{t,j_t} (\phi_{j_t}(\eta_t \hat{L}_t) - \phi_{j_t}(\eta_t(\hat{L}_t + \hat{\ell}_t)) \Big| \hat{L}_t \Big] \\ \leq \sum_{a=1}^N \mathbb{E}\left[\sum_{i \neq j_t} \mathcal{O}\left(\frac{\alpha}{\beta_t} p_{t,i}^{1/\alpha} \right) \frac{w_{t,j_t} \pi_{t,j_t,a}}{P_{t,a}} \middle| \hat{L}_t \right],$$

where ζ_α also can be tuned as in MABs.

Therefore, it remains to consider the case $\bar{\mathcal{E}}_{t,\alpha} \cup \{\hat{\ell}_{t,j_t} > \zeta_\alpha \beta_t\}$ for some $\zeta_\alpha < \xi_\alpha^{-1/\alpha} - 1$. Since Algorithm 2 utilize simple GR and $\beta_t \geq \sqrt{t}$ for $\alpha \geq 2$ (which is the current interest), we can utilize Lemma 20. Here, note that the results in Lemma 20 considers the IW estimator for MABs, i.e., $\hat{\ell}_{t,i} = \mathbb{1}[i_t = i] M_t \ell_{t,i}$, while our setting is $\hat{\ell}_{t,i} = M_t \ell_{t,i} \pi_{t,i,a_t}$. Therefore, the condition $\hat{\ell}_{t,j_t} \geq \zeta_\alpha \beta_t$ is related to the condition of $M_t \pi_{t,j_t,a_t} \geq \zeta_\alpha \beta_t$.

Specifically, on $\bar{\mathcal{E}}_{t,\alpha}$, as shown in MABs, we have $w_{t,j_t} \geq 1 - \xi_\alpha$ on $\bar{\mathcal{E}}_{t,\alpha}$, which implies that $P_{t,a_t} \geq \pi_{t,j_t,a_t}(1 - \xi_\alpha) \geq \nu(1 - \xi_\alpha)$ by Assumption 1. Therefore, $\nu(1 - \xi_\alpha)$ plays the same role in w' in Lemma 20. Here, note that we do not need to consider the case $\pi_{t,j_t,a_t} = 0$ since this case is included in the case of $\hat{\ell}_{t,j_t} \leq \zeta_\alpha \beta_t$.

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{1}[\bar{\mathcal{E}}_{t,\alpha}, \hat{\ell}_{t,j_t} > \zeta_\alpha \beta_t] \hat{\ell}_{t,j_t} (\phi_{j_t}(\eta_t \hat{L}_t) - \phi_{j_t}(\eta_t(\hat{L}_t + \hat{\ell}_t)) \Big| \hat{L}_t \Big] \\
&\leq \mathbb{E}\Big[\mathbb{1}[\bar{\mathcal{E}}_{t,\alpha}, \hat{\ell}_{t,j_t} > \zeta_\alpha \beta_t] \hat{\ell}_{t,j_t} \Big| \hat{L}_t \Big] \\
&\leq \frac{1}{1 - (1 - \xi_\alpha)\nu} (1 - (1 - \xi_\alpha)\nu)^{\zeta_\alpha \beta_t} (\zeta_\alpha \beta_t + 1/(1 - (1 - \xi_\alpha)\nu)) \qquad \text{(by Lemma 20)} \\
&=: g_t(\alpha; \nu).
\end{aligned}$$

Then, Lemma 20 shows that $\sum_t g_t(\alpha) \leq \mathcal{O}(1)$. More precisely, $g_t(\alpha; \nu)$ is the upper bounds of

$$P_{t,a_t} \sum_{m = \lfloor \frac{\zeta_\alpha \beta_t}{\pi_{t,j_t,a_t}} \rfloor + 1}^{\infty} m (1 - P_{t,a_t})^{m-1} \leq (1 - P_{t,a_t})^{\lfloor \frac{\zeta_\alpha \beta_t}{\pi_{t,j_t,a_t}} \rfloor} \left(\lfloor \frac{\zeta_\alpha \beta_t}{\pi_{t,j_t,a_t}} \rfloor + \frac{1}{P_{t,a_t}} \right), \tag{35}$$

where $P_{t,a_t} = \sum_{i=1}^K w_{t,i} \pi_{t,i,a_t} \geq w_{t,j_t} \pi_{t,j_t,a_t}$. The introduction of ν is to obtain a general upper bound by removing the dependency of π_{t,j_t,a_t} in $g_t(\alpha)$, which we cannot control.

Finally, we show the order of $\sum_t g_t(\alpha)$ in terms of ν. From (35), we consider the order of

$$(1 - P_{t,a_t})^{\lfloor \frac{\zeta_\alpha \beta_t}{\pi_{t,j_t,a_t}} \rfloor} \left(\lfloor \frac{\zeta_\alpha \beta_t}{\pi_{t,j_t,a_t}} \rfloor + \frac{1}{P_{t,a_t}} \right) \leq (1 - \pi_{t,j_t,a_t})^{\lfloor \frac{\zeta_\alpha \beta_t}{\pi_{t,j_t,a_t}} \rfloor} \left(\lfloor \frac{\zeta_\alpha \beta_t}{\pi_{t,j_t,a_t}} \rfloor + \pi_{t,j_t,a_t} \right).$$

Therefore, it is sufficient to consider the order of

$$\sum_{t=1}^{\infty} (1 - a)^{\frac{b\sqrt{t}}{a}} \left(\frac{b\sqrt{t}}{a} + a \right), \text{ where } a, b \in (0, 1).$$

It is easy to see that it is decreasing with respect to $a \in (0,1)$. Since we have

$$(1-a)^{\frac{b\sqrt{t}}{a}}\left(\frac{b\sqrt{t}}{a}+a\right) \leq e^{-b\sqrt{t}}\left(\frac{b\sqrt{t}}{a}+a\right) = \frac{b\sqrt{t}e^{-b\sqrt{t}}}{a} + ae^{-b\sqrt{t}},$$

this implies that

$$\sum_{t=1}^{\infty}(1-a)^{\frac{b\sqrt{t}}{a}}\left(\frac{b\sqrt{t}}{a}+a\right) \leq \frac{2e^{-b}(b^2+2b+2)}{ab^2} + \frac{2ae^{-b}(b+1)}{b^2} \leq \mathcal{O}\left(\frac{1}{ab^2}+\frac{a}{b^2}\right).$$

Therefore, $\sum_t g_t(\alpha;\nu) = \mathcal{O}(\alpha^2/\nu)$ if we choose $\zeta_\alpha \approx 1-2^{-1/\alpha}$ as in the multi-armed bandit. ■

Appendix F. Proof of Theorem 6

In this section, we show the BOBW guarantee of Algorithm 2. The most of the proofs are essentially the same to that of Theorem 4, where the only difference is related to the change of z_t.

F.1. Adversarial regime

In this regime, it suffices to show that $h_{\max}z_t$ is at most $NK^{1/\alpha}$ from the second term in (13). By definition of h_t and z_t in (12), we obtain

$$h_t = \sum_{i\neq j_{t+1}} \frac{\alpha}{\alpha-1} q_{t+1,i}^{1-1/\alpha} \leq \sum_{i\neq j_{t+1}} \frac{\alpha}{\alpha-1}\frac{1}{\sigma_{t+1,i}^{1-1/\alpha}} = \sum_{n=2}^{K}\frac{\alpha}{\alpha-1}\frac{1}{n^{1-1/\alpha}} \leq \frac{\alpha^2}{\alpha-1}((K+1)^{1/\alpha}-1) \leq \frac{\alpha^2}{\alpha-1}K^{1/\alpha}$$

and

$$z_t = N\alpha \max_{j\neq j_{t+1}} q_{t+1,j}^{1/\alpha} \leq N\alpha 2^{-1/\alpha}.$$

Therefore,

$$h_{\max}\sum_{t=1}^{T} z_t \leq \frac{\alpha^3}{\alpha-1}(K/2)^{1/\alpha}NT.$$

which concludes the proof for the adversarial regime.

F.2. Adversarial regime with self-bounding constraint

Basically, we can directly utilize the techniques used to prove BOBW guarantee for MABs given in Appendix D.2. Therefore, it suffices to show the upper bounds of $h_t z_t$ on $\mathcal{D}_{t+1,\alpha}$ since we can just use the $\max_t h_t z_t \leq \alpha^3 K^{1/\alpha}N/(\alpha-1)$ in the case of $\mathcal{D}_{t+1,\alpha}^c$.

Note that $w_{t,j_t} \geq w_{t,i}$ for any $i \in [K]$. From (30), we already obtain that

$$\mathbb{1}[\mathcal{D}_{t+1,\alpha}]h_t \leq \mathbb{1}[\mathcal{D}_{t+1,\alpha}]\frac{\alpha}{\alpha-1}16^{1-1/\alpha}\left(\sum_{i\neq i^*}\frac{1}{\Delta_i^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(\sum_{i\neq i^*}\Delta_i w_{t+1,i}\right)^{1-1/\alpha}.$$

For z_t, by Hölder's inequality, we have

$$\begin{aligned}
\mathbb{1}[\mathcal{D}_{t+1,\alpha}]z_t &= \mathbb{1}[\mathcal{D}_{t+1,\alpha}]N\alpha \max_{i\neq j_{t+1}} q_{t+1,i}^{1/\alpha} \\
&\leq \mathbb{1}[\mathcal{D}_{t+1,\alpha}]N\alpha 16^{1/\alpha} \max_{i\neq j_{t+1}} w_{t+1,i}^{1/\alpha} && \text{(by (29))} \\
&\leq \mathbb{1}[\mathcal{D}_{t+1,\alpha}]N\alpha 16^{1/\alpha} \sum_{i\neq j_{t+1}} w_{t+1,i}^{1/\alpha} \\
&\leq \mathbb{1}[\mathcal{D}_{t+1,\alpha}]N\alpha 16^{1/\alpha} \sum_{i\neq i^*} w_{t+1,i}^{1/\alpha} && (w_{t+1,j_{t+1}} \geq w_{t+1,i}, \forall i) \\
&= \mathbb{1}[\mathcal{D}_{t+1,\alpha}]N\alpha 16^{1/\alpha} \sum_{i\neq i^*} \frac{1}{\Delta_i^{1/\alpha}}(\Delta_i w_{t+1,i})^{1/\alpha} \\
&\leq \mathbb{1}[\mathcal{D}_{t+1,\alpha}]N\alpha 16^{1/\alpha} \left(\sum_{i\neq i^*}\frac{1}{\Delta_i^{1/(\alpha-1)}}\right)^{1-\frac{1}{\alpha}}\left(\sum_{i\neq i^*}\Delta_i w_{t+1,i}\right)^{1/\alpha}.
\end{aligned}$$

Therefore, we obtain

$$\mathbb{1}[\mathcal{D}_{t+1,\alpha}]h_t z_t \leq \mathbb{1}[\mathcal{D}_{t+1,\alpha}]\omega''(\Delta)\left\langle \Delta, w_{t+1}\right\rangle,$$

where

$$\omega''(\Delta) = 8N\frac{\alpha^2}{\alpha-1}\left(\sum_{i\neq i^*}\Delta_i^{-\frac{1}{\alpha-1}}\right)^{1-\frac{1}{\alpha}}\left(\sum_{i\neq i^*}\Delta_i^{1-\alpha}\right)^{\frac{1}{\alpha}},$$

By following the same steps in Appendix D.2 until (33), we obtain

$$\mathrm{Reg}(T) \leq \mathcal{O}\left(\sqrt{\log T\left(\omega''(\Delta) + \frac{\alpha^3 NK^{1/\alpha}}{(\alpha-1)\Delta_{\min}}\right)(\mathrm{Reg}(T)+C)}\right) + \mathcal{O}\left(\sqrt{\frac{NK^{1/\alpha}\alpha^3}{\alpha-1}} + 1/\nu\right),$$

where the last two terms are the term related to $z_1/\beta_1 + \beta_1 K^{1/\alpha}$ and $g_t(\alpha,\nu)$. Note that additional $z_{\max}/\beta_1$ term appear in Lemma 3 is $\mathcal{O}(1)$ in this case, so that we exclude. This result provides

$$\mathrm{Reg}(T) \leq \mathcal{O}\left(\omega'''(\Delta)\log T + \sqrt{C\omega'''(\Delta)\log T} + \sqrt{\frac{NK^{1/\alpha}\alpha^3}{\alpha-1}} + 1/\nu\right),$$

where

$$\begin{aligned}\omega'''(\Delta) &= \omega''(\Delta) + \frac{\alpha^3 N K^{1/\alpha}}{(\alpha-1)0.31\Delta \min} \\ &= \mathcal{O}\left(\frac{8N\alpha^2}{\alpha-1}\left(\sum_{i\neq i^*}\Delta_i^{-\frac{1}{\alpha-1}}\right)^{1-\frac{1}{\alpha}}\left(\sum_{i\neq i^*}\Delta_i^{1-\alpha}\right)^{\frac{1}{\alpha}} + \frac{\alpha^3 N K^{1/\alpha}}{(\alpha-1)0.31\Delta_{\min}}\right) \\ &= \mathcal{O}\left(\frac{\alpha^3 N K^{1/\alpha}}{(\alpha-1)\Delta_{\min}}\right).\end{aligned}$$