Follow-the-Perturbed-Leader Achieves Best-of-Both-Worlds for Bandit Problems

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

Shinji Ito I-SHINJI@NEC.COM
NEC Corporation

Taira Tsuchiya TSUCHIYA@SYS.I.KYOTO-U.AC.JP
Kyoto University and RIKEN AIP

Editors: Shipra Agrawal and Francesco Orabona

Abstract

This paper discusses the adversarial and stochastic K-armed bandit problems. In the adversarial setting, the best possible regret is known to be $O(\sqrt{KT})$ for time horizon T. This bound can be achieved by several policies but they require to explicitly compute the arm-selection probabilities by solving optimization problems at each round, which becomes problematic in some settings. One promising candidate to avoid this issue is the Follow-The-Perturbed-Leader (FTPL) policy, which simply chooses the arm with the minimum cumulative estimated loss with a random perturbation. In particular, it has been conjectured that $O(\sqrt{KT})$ regret might be achieved by FTPL with a Fréchet-type perturbation. This paper affirmatively resolves this conjecture by showing that Fréchet perturbation indeed achieves this bound. We also show that FTPL achieves a logarithmic regret for the stochastic setting, meaning that FTPL achieves best-of-both-worlds regret bounds. The key to these bounds is the novel technique to evaluate the stability of the policy and to express the regret at each round in multiple forms depending on the estimated losses.

Keywords: multi-armed bandit, adversarial bandit, stochastic bandit, best-of-both-worlds, follow-the-perturbed-leader, Fréchet distribution

1. Introduction

The multi-armed bandit (MAB) is a model of a gampler playing slot machines and is one of the most fundamental problems of decision making under uncertainty. In this problem, there are K arms of slot machines and the player chooses one arm I_t based on his/her past observation at each round $t \in [T] = \{1, 2, \dots, T\}$ for time horizon T. The loss vector $\ell_t = (\ell_{t,1}, \ell_{t,2} \dots, \ell_{t,K})^\top \in [0,1]^K$ is determined by the environment and the player can only observe the loss ℓ_{t,I_t} of the pulled arm I_t.

The performance of the player is often measured by the *pseudo-regret* defined as $\mathrm{Regret}(T) = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,I_t}\right] - \min_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^T \ell_{t,i}\right]$, which is the gap of the expected cumulative loss between the policy and the best fixed arm. There are mainly two settings on the formulation of the environment to determine the loss vectors: the stochastic setting (Lai and Robbins, 1985; Auer et al., 2002a) and the adversarial setting (Auer et al., 2002b).

In the stochastic setting, we assume that loss vectors ℓ_t for $t \in [T]$ are i.i.d. from an unknown but fixed distribution $\mathcal{D}$ over $[0,1]^K$. The difficulty of this problem is usually measured by the *suboptimality gap* $\Delta_i = \mu_i - \mu_{i^*}$ for $\mu_i = \mathbb{E}_{\ell\sim\mathcal{D}}[\ell_i]$ and $i^* \in \mathrm{argmin}_{i\in[K]}\{\mu_i\}$, that is, Δ_i is the gap between the arm i and the best arm i^*. The optimal (problem-dependent) regret bound given the

suboptimality gap is expressed as $\mathrm{Regret}(T) = \sum_{i:\Delta_i>0} O(\frac{\log T}{\Delta_i})$ for fixed Δ_i with $T \to \infty$, which is called a problem-dependent regret bound. This bound can be achieved by several policies such as UCB (Auer et al., 2002a; Cappé et al., 2013) and Thompson sampling (Kaufmann et al., 2012; Agrawal and Goyal, 2013; Riou and Honda, 2020), some of which further improves the performance by considering dependence not only on $\{\Delta_i\}$ but also on $\mathcal{D}$ itself.

In the adversarial setting, no specific distribution of the loss is assumed and the environment may adversarially choose loss vectors ℓ_t depending on the history of the decisions $\{I_s\}_{s=1}^{t-1}$. For this setting, a family of policies called the Follow-The-Regularized-Leader (FTRL) with appropriate regularization functions achieves $O(\sqrt{KT})$ regret bounds (Audibert and Bubeck, 2009; Zimmert and Lattimore, 2019). In particular, the Tsallis-INF policy (Zimmert and Seldin, 2021) with adaptive learning rate has a special strength in its $O(\sum_{i\neq i^*} \frac{\log T}{\Delta_i})$ regret bound for the stochastic setting, though the leading factor of the $O(\log T)$ term is slightly worse than those purely for the stochastic setting and we need the assumption of the unique optimal arm.

In practice, it is difficult to know whether the environment is stochastic or adversarial (or in between), and it is desirable that the used policy has regret guarantees for both settings. This kind of guarantees is called Best-Of-Both-Worlds (BOBW, Bubeck and Slivkins, 2012), and Tsallis-INF (or other FTRL policies) is one of the most promising frameworks for BOBW policies for further complex settings such as partial monitoring and combinatorial-semi bandits (Zimmert et al., 2019).

1.1. Follow-The-Perturbed-Leader

A limitation of most existing BOBW policies is that we need to explicitly compute the list of arm-selection probabilities of arms. Though the computational complexity is usually $O(K)$ in most cases, it might become problematic when more complex settings are considered, such as the combinatorial bandits where the number of actions is exponential in K.

The Follow-The-Perturbed-Leader (FTPL) policy has been researched as a promising candidate to circumvent this limitation. This policy greedily selects the arm with the least estimated cumulative loss with a random perturbation at each round. To be more specific, FTPL chooses arm $\mathrm{argmin}_{i\in[K]}\{\hat{L}_{t,i} - r_{t,i}/\eta_t\}$, where $r_{t,i}$ is the random perturbation drawn from some distribution, η_t is a parameter corresponding to learning rate and $\hat{L}_{t,i}$ is an estimation of the cumulative loss $L_{t,i} = \sum_{s=1}^{t-1} \ell_{s,i}$. It was suggested in Kim and Tewari (2019) that the perturbation distribution to achieve $O(\sqrt{KT})$ adversarial regret (if exists) would have a Fréchet-type tail distribution. Nevertheless, this achievability has still been an open question, seemingly due to the following difficulties.

In the analysis of FTRL and FTPL, the key to the optimal adversarial regret bound is to evaluate how stably the arm-selection probability $w_{t,i} = \phi_i(\eta_t \hat{L}_t)$ behaves against the change of the estimated loss vector $\hat{L}_t$, where the function ϕ_i takes different forms depending on the policy. For this quantity, existing approaches try to uniformly bound $\phi_i'(\eta_t\hat{L}_t)/w_{t,i}$ (Abernethy et al., 2015) or $\phi_i'(\eta_t\hat{L}_t)/w_{t,i}^{3/2}$ (Bubeck, 2019) by a constant, where $\phi_i'(\lambda) = \mathrm{d}\phi_i(\lambda)/\mathrm{d}\lambda_i$. However, due to the complicated expression of $\phi_i(\cdot)$ of FTPL it is very difficult to obtain such a bound.

1.2. Contribution of the Paper

In this paper, we show that FTPL with Fréchet perturbation achieves $O(\sqrt{KT})$ adversarial regret and $\sum_{i\neq i^*} O(\frac{\log T}{\Delta_i})$ stochastic regret. In this analysis we derive a bound on $\phi_i'(\eta_t\hat{L}_t)/w_{t,i}$ depending on $\hat{L}_{t,i}$ rather than a uniform bound, which results in a tighter bound when its summation over arms

is taken. To be more specific, we show $\phi_i'(\eta_t \hat{L}_t)/w_{t,i} = O(1/\sqrt{\sigma_i})$ when the estimated cumulative loss $\hat{L}_{t,i}$ of arm i is the σ_i-th smallest among K arms.

We use the *self-bounding technique* (Zimmert and Seldin, 2021) to derive the stochastic regret bound, which is a typical tool for showing BOBW property of FTRL. However, in the case of FTPL the regret has a complicated dependency on $\hat{L}_t$ and the analysis for the adversarial setting does not immediately lead to bounds where the self-bounding technique is applicable. To solve this difficulty, we evaluate the regret depending on whether the estimated best arm is well-concentrated or not.

1.3. Related Work

After Hannan (1957) proposed FTPL in the context of game theory, Kalai and Vempala (2005) presented an analysis of FTPL as adversarial online linear optimization algorithms. Since then, FTPL has attracted much attention for its computational efficiency, and has been extended to a variety of settings of online learning and bandit problems, including adversarial MAB problems (Abernethy et al., 2015), linear bandits (McMahan and Blum, 2004), combinatorial semi-bandits (Neu, 2015; Neu and Bartók, 2016), online contextual learning problems (Syrgkanis et al., 2016), online learning with non-linear losses (Dudík et al., 2020) and MDP bandits (Dai et al., 2022). Most of the analysis in these studies is based on the approach of interpreting an FTPL policy as an FTRL policy with a regularizer function associated with the perturbation distribution (Abernethy et al., 2014, 2016). In particular, Abernethy et al. (2015) applied this approach to MAB problems to show that FTPL with a certain perturbation achieves regret bound of $O(\sqrt{KT \log K})$. However, finding a perturbation distribution that achieves an optimal regret bound of $O(\sqrt{KT})$ has been an open problem (Kim and Tewari, 2019), while FTRL with Tsallis-entropy regularization achieves optimal $O(\sqrt{KT})$-bound (Audibert and Bubeck, 2009; Abernethy et al., 2015; Zimmert and Seldin, 2021). A natural approach is to construct a perturbation distribution that induces the Tsallis entropy, but Kim and Tewari (2019) showed that no such distribution exists. This paper affirmatively resolves this open problem by showing that FTPL with Fréchet distributions achieves $O(\sqrt{KT})$ regret.

BOBW polices have been proposed for various online learning problems, such as MAB problems (Bubeck and Slivkins, 2012; Zimmert and Seldin, 2021), the problem of prediction with expert advice (De Rooij et al., 2014; Gaillard et al., 2014; Luo and Schapire, 2015), combinatorial semi-bandits (Zimmert et al., 2019; Ito, 2021a), online linear optimization (Huang et al., 2016), linear bandits (Lee et al., 2021), online learning with feedback graphs (Erez and Koren, 2021), and episodic Markov decision processes (Jin and Luo, 2020; Jin et al., 2021). Some of them are designed based on the FTRL framework. For such policies, stochastic regret bounds (e.g., of $O(\log T)$) are proved by combining regret upper bounds and *lower* bounds depending on arm-selection probabilities. This approach is called a *self-bounding technique*. This study differs from these existing studies in that the self-bounding technique is applied to FTPL rather than FTRL. It should be emphasized that this study provides the first BOBW bandit policy based of FTPL.

2. Problem Setup

In this section, we formulate the problem and explain the policy to analyze. At each round $t \in [T]$, the environment determines a loss vector $\ell_t = (\ell_{t,1}, \ell_{t,2}, \dots, \ell_{t,K})^\top \in [0,1]^K$ and the player pulls an arm $I_t \in [K]$. The player then observes the incurred loss ℓ_{t,I_t} of the pulled arm.

In the adversarial setting, we do not assume any model for the loss vector ℓ_t, which may depend on the history of loss vectors and chosen arms $\{(\ell_s, I_s)\}_{s=1}^{t-1}$. In the stochastic setting,

Algorithm 1: FTPL with geometric resampling.

$\hat{L}_1 := \mathbf{0}$.
for $t = 1, 2, \dots, T$ **do**
  Sample $r_t = (r_{t,1}, r_{t,2}, \dots, r_{t,K})$ i.i.d. from $\mathcal{F}_2$.
  Pull arm $I_t = \operatorname{argmin}_{i \in [K]} \{\hat{L}_{t,i} - r_{t,i}/\eta_t\}$ and observe ℓ_{t,I_t}.
  Set $m := 0$.
  repeat
    $m := m + 1$.
    Sample $r'_{t,1}, r'_{t,2}, \dots, r'_{t,K}$ i.i.d. from $\mathcal{F}_2$.
  until $I_t = \operatorname{argmin}_{i \in [K]} \{\hat{L}_{t,i} - r'_{t,i}/\eta_t\}$
  Set $\widehat{w^{-1}_{t,I_t}} := m$ and $\hat{L}_{t+1} := \hat{L}_t + \ell_{t,I_t} \widehat{w^{-1}_{t,I_t}} e_{I_t}$.

$\ell_1, \ell_2, \dots, \ell_T \in [0,1]^K$ are i.i.d. from an unknown but fixed distribution $\mathcal{D}$. The expected reward of arm i is denoted by $\mu_i = \mathbb{E}_{\ell \sim \mathcal{D}}[\ell_i] \in [0,1]$. The suboptimality gap of arm i is expressed by $\Delta_i = \mu_i - \mu_{i^*}$. Then an optimal arm is expressed by $i^* \in \operatorname{argmin}_{i \in [K]} \mu_i$.

The performance of the player is evaluated in terms of the pseudo-regret $\mathrm{Regret}(T)$ defined as

$$\mathrm{Regret}(T) = \mathbb{E}\left[\sum_{t=1}^{T} (\ell_{t,I_t} - \ell_{t,i^*})\right], \qquad i^* \in \operatorname*{argmin}_{i \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right].$$

Note that the notion called *regret* instead of the pseudo-regret is sometimes considered in the adversarial setting. See Appendix A for the relation between them.

We consider the Follow-The-Perturbed-Leader (FTPL) policy given in Algorithm 1. This policy maintains the estimated cumulative loss $\hat{L}_t$ and pulls the arm minimizing the loss with a random perturbation r_t/η_t. Here $\eta_t = O(t^{-1/2})$ is the exploration parameter or the learning rate. The elements of $r_t = (r_{t,1}, r_{t,2}, \dots, r_{t,K})$ are i.i.d. from Fréchet distribution $\mathcal{F}_2$ with shape parameter $\alpha = 2$, whose density function $f(x)$ and cumulative distribution function $F(x)$ are expressed by

$$f(x) = 2x^{-3}\mathrm{e}^{-1/x^2}, \qquad F(x) = \mathrm{e}^{-1/x^2}, \qquad x \geq 0,$$

respectively. In the following, "Fréchet distribution" always refers to the distribution of this parameter. The probability of pulling arm i given $\hat{L}_t$ is expressed as $w_{t,i} = \phi_i(\eta_t \hat{L}_t)$ where, for $\lambda \in [0,\infty)^K$,

$$\phi_i(\lambda) := \mathbb{P}_{r \sim \mathcal{F}_2}\left[i = \operatorname*{argmin}_{j \in [K]} \{\lambda_j - r_j\}\right] = \int_{-\min_{j \in [K]} \lambda_j}^{\infty} \frac{2}{(z + \lambda_i)^3} \exp\left(-\sum_{i'} \frac{1}{(z + \lambda_{i'})^2}\right) \mathrm{d}z. \tag{1}$$

In general, FTRL and FTPL policies use an estimator $\hat{\ell}_t$ of the loss vector ℓ_t. Their cumulative versions are denoted by $L_t = \sum_{s=1}^{t-1} \ell_s$ and $\hat{L}_t = \sum_{s=1}^{t-1} \hat{\ell}_s$. The gap of the loss from its minimum is expressed by underlines, e.g., $\underline{\hat{L}}_t = \hat{L}_t - \mathbf{1} \min_i \hat{L}_{t,i} \in [0,\infty)^K$, where $\mathbf{1}$ is the all-one vector.

In FTRL we often use an unbiased estimator $\hat{\ell}_t = (\ell_{t,I_t}/w_{t,I_t}) e_{I_t}$ of ℓ_t, which is called Importance Weighted (IW) estimator. Here e_i is the unit vector whose i-th element is one and the others

are zero. On the other hand in FTPL, we do not explicitly compute $w_{t,i}$ and we instead use an estimator $\widehat{w_{t,i}^{-1}}$ of $w_{t,i}^{-1}$ by the technique called *geometric resampling* (Neu and Bartók, 2016).

Let us consider repeating resampling of r'_t from the same distribution until $\operatorname{argmin}_{i\in[K]}\{\hat{L}_{t,i} - r'_{t,i}/\eta_t\}$ coincides with $I_t = \operatorname{argmax}_{i\in[K]}\{\hat{L}_{t,i} - r_{t,i}/\eta_t\}$, and let $\widehat{w_{t,I_t}^{-1}}$ be the number of this resampling. Then its expectation is expressed as $1/w_{t,I_t}$, meaning that $\widehat{w_{t,I_t}^{-1}}$ is an unbiased estimator of $1/w_{t,I_t}$. Based on this estimator we define the loss estimator by $\hat{\ell}_t = (\ell_{t,I_t}\widehat{w_{t,I_t}^{-1}})e_{I_t}$, which corresponds to Lines 5–10 of Algorithm 1. The expected number of resampling at each round is $\sum_{i\in[K]} w_{t,i}\cdot 1/w_{t,i} = K$, which is independent of w_t. See Neu and Bartók (2016) for the technique to deterministically bound the number of resampling.

3. Regret Bounds

In this section, we summarize the regret bounds of FTPL and outline the evaluation of the regret.

3.1. Main Results

Theorem 1 *In the adversarial setting, FTPL with learning rate $\eta_t = c/\sqrt{t}$ for $c > 0$ satisfies*

$$\operatorname{Regret}(T) \le \left(12c\sqrt{\pi} + \frac{3.7}{c}\right)\sqrt{KT} + 10c + \frac{\sqrt{\pi K}}{c},$$

whose dominant term is optimized as $\operatorname{Regret}(T) \le 17.8\sqrt{KT} + O(\sqrt{K})$ when $c = 0.42$.

This result shows that FTPL achieves the optimal worst-case regret and affirmatively resolves the open question given in Kim and Tewari (2019).

We can also show that FTPL achieves logarithmic regret in the stochastic setting.

Theorem 2 *Assume that $i^* = \operatorname{argmin}_i \mu_i$ is unique and let $\Delta = \min_{i\neq i^*}\Delta_i$. Then, FTPL with learning rate $\eta_t = c/\sqrt{t}$ for $c > 0$ satisfies*

$$\operatorname{Regret}(T) \le \sum_{i\neq i^*}\frac{(25c + c^{-1})^2\log T}{0.075\Delta_i} + \frac{(121c + 12c^{-1})^2 K}{\Delta},$$

whose dominant term is optimized as $1400\sum_{i\neq i^*}\frac{\log T}{\Delta_i} + o(\log T)$ when $c = 0.2$.

Whereas these regret bounds suggest parameters $c = 0.42$ or $c = 0.2$, such choice is too small and not recommended in practice as we can see from the empirical results in Section 7. The reason can be explained as follows. The current evaluation is somewhat loose for the component of the regret called *stability term*, which corresponds to how stable the policy is against the change of the estimated loss. As a result, the parameter optimizing the current bound makes the policy too stable and harms the adaptivity (called the *penalty term*) of the policy.

In particular, the stochastic regret bound in Theorem 2 is about 3000 times worse than the optimal bound $\sum_{i\neq i^*}\frac{\log T}{2\Delta_i} + o(\log T)$ (seen from Lai and Robbins, 1985). Though the upper bound can be improved to $16\sum_{i\neq i^*}\frac{\log T}{\Delta_i} + o(\log T)$ at the cost of larger $o(\log T)$ term (see Remark 12 in Appendix D.3), it is still true that the current analysis suffers a large constant factor. How to

improve this evaluation is an important future work, which would fill the gap between theoretically and empirically good learning rates (see also simulation results in Section 7).

There are also studies to discuss the intermediate settings between stochastic and adversarial ones, many of which are captured by the *adversarial setting with self-bounding constraint*. We can also generalize the results to this setting but we do not give explicit bounds for simplicity since they require some additional arguments. See Appendix E for the outline of the generalization.

3.2. Regret Decomposition

Now we explain how to evaluate the regret of FTPL. First of all, the regret is expressed by

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \mathbb{E}\left[\langle \ell_t, w_t - e_{i^*} \rangle\right] = \sum_{t=1}^{T} \mathbb{E}\left[\left\langle \hat{\ell}_t, w_t - e_{i^*} \right\rangle\right].$$

This can be decomposed in the following way, whose proof is given in Appendix C.

Lemma 3

$$\mathrm{Regret}(T) \le \sum_{t=1}^{T} \mathbb{E}\left[\left\langle \hat{\ell}_t, w_t - w_{t+1} \right\rangle\right] + \sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \mathbb{E}\left[r_{t+1,I_{t+1}} - r_{t+1,i^*}\right] + \frac{\sqrt{\pi K}}{\eta_1}. \quad (2)$$

This decomposition is essentially similar to that based on the reduction to FTRL (Lattimore and Szepesvári, 2020, Exercise 28.12) but is slightly simpler. We refer to the first and second terms of (2) as *stability term* and *penalty term*, respectively. Here note that

$$\begin{aligned} w_t - w_{t+1} &= \phi(\eta_t \hat{L}_t) - \phi(\eta_{t+1}(\hat{L}_t + \hat{\ell}_t)) \\ &= (\phi(\eta_t \hat{L}_t) - \phi(\eta_t(\hat{L}_t + \hat{\ell}_t))) + (\phi(\eta_t(\hat{L}_t + \hat{\ell}_t)) - \phi(\eta_{t+1}(\hat{L}_t + \hat{\ell}_t))). \end{aligned} \quad (3)$$

We can show that the second term of (3) is not dominant and the stability term is bounded as follows.

Lemma 4

$$\sum_{t=1}^{T} \mathbb{E}\left[\left\langle \hat{\ell}_t, w_t - w_{t+1} \right\rangle\right] \le \sum_{t=1}^{T} \mathbb{E}\left[\left\langle \hat{\ell}_t, \phi(\eta_t \hat{L}_t) - \phi(\eta_t(\hat{L}_t + \hat{\ell}_t)) \right\rangle\right] + 10\eta_1. \quad (4)$$

The proof of this lemma is also given in Appendix C. In Sections 4 and 5 we bound (2) and (4) for the adversarial setting, which completes the proof of Theorem 1. We will also outline the analysis for the stochastic setting in Section 6, whose full proof is given in Appendix D.

4. Stability of Arm-selection Probability

In the analysis of FTPL and FTRL, it is known that the relation between the arm-selection probability function $\phi(\lambda)$ and its derivatives plays the central role in the regret analysis (Abernethy et al., 2015; Bubeck, 2019). To this end, this section derives properties of $\phi(\cdot)$ defined by (1) in our setting, which becomes the main difficulty of the analysis of FTPL. We can rewrite this function by

$$\phi_i(\lambda) = 2\int_0^\infty \frac{1}{(z + \underline{\lambda}_i)^3} \exp\left(-\sum_{i'} \frac{1}{(z + \underline{\lambda}_{i'})^2}\right) \mathrm{d}z, \quad (5)$$

where $\underline{\lambda} = \lambda - \mathbf{1}\min_{i\in[K]}\lambda_i \in [0,\infty)^K$. Define

$$\phi_i'(\lambda) = \frac{\partial \phi_i}{\partial \lambda_i}(\lambda), \qquad I_{i,n}(\lambda) = \int_0^\infty \frac{1}{(z+\lambda_i)^n} \exp\left(-\sum_{i'} \frac{1}{(z+\lambda_{i'})^2}\right) \mathrm{d}z > 0. \tag{6}$$

Here note that we used λ rather than $\underline{\lambda}$ in the RHS of (6) since we will sometimes consider $I_{i,n}(\lambda)$ for λ not in the form of $\underline{\lambda}$. By taking the derivative of (5) we immediately have

$$\phi_i(\lambda) = 2I_{i,3}(\underline{\lambda}), \qquad \phi_i'(\lambda) = -6I_{i,4}(\underline{\lambda}) + 4I_{i,6}(\underline{\lambda}). \tag{7}$$

One might think that (7) is not straightforward from (5) when $i \in \operatorname{argmin}_{j\in[K]}\lambda_j$, because $\{\underline{\lambda}_j\}_{j\neq i}$ rather than $\underline{\lambda}_i$ might change with λ_i. Nevertheless, by separately evaluating the derivative from each direction we can confirm that (7) eventually holds for all $i \in [K]$ (see Appendix F.1 for details).

Note that we have $\phi_i'(\lambda) \leq 0$, that is, $\phi_i(\lambda)$ is nonincreasing in λ_i thouth it is seemingly unclear from (7). This is because $\phi_i(\lambda)$ is the probability of $\lambda_i - r_i < \min_{i\neq j}\{\lambda_j - r_j\}$ when each r_i follows the Fréchet distribution, which is clearly nonincreasing in λ_i. By the same reason, $\phi_i(\lambda)$ is nondecreasing in λ_j for $j \neq i$. We can also see that $I_{i,4}(\lambda)$ is nonincreasing in λ_i and nondecreasing in λ_j for $j \neq i$. This is because $\frac{4}{\sqrt{\pi}}I_{i,4}(\lambda)$ is the probability of $\{\lambda_i - r_i < \min_{i\neq j}\{\lambda_j - r_j\} \wedge 0\}$ when r_j with $j \neq i$ follow the Fréchet distribution and r_i follows distribution with density

$$g(x) = \frac{x^{-4}\mathrm{e}^{-1/x^2}}{\int_0^\infty z^{-4}\mathrm{e}^{-1/z^2}\mathrm{d}z} = \frac{4x^{-4}\mathrm{e}^{-1/x^2}}{\sqrt{\pi}}.$$

The main result of this section and the key to the main theorems are the following lemma.

Lemma 5 *If λ_i is the σ_i-th smallest among $\lambda_1, \lambda_2, \ldots, \lambda_K$ (ties are broken arbitrarily) then*

$$\frac{I_{i,4}(\underline{\lambda})}{I_{i,3}(\underline{\lambda})} \leq \frac{2}{3\underline{\lambda}_i} \wedge \frac{\sqrt{\pi/\sigma_i}}{2}. \tag{8}$$

As we can see from (6) and (7), the LHS of (8) is an upper bound of $-\phi_i'(\lambda)/3\phi_i(\lambda)$. On the other hand, the RHS is roughly related to $\sqrt{\phi_i(\lambda)}$, which will be seen from, e.g., (11). Therefore this bound plays (though not in the strict sense) a role similar to showing $\phi_i'(\lambda)/(\phi_i(\lambda))^{3/2} = O(1)$, which is the desired scenario for the optimal adversarial regret (Bubeck, 2019).

In the rest of this section we always assume $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_K$ so that $\sigma_i = i$ for notational simplicity. The following lemma is the key property for the proof of Lemma 5.

Lemma 6 *$I_{i,4}(\lambda)/I_{i,3}(\lambda)$ is monotonically increasing in λ_j for any $j \neq i$.*

Proof Define

$$I_{i,j,n}(\lambda) = \int_0^\infty \frac{1}{(z+\lambda_i)^n}\frac{1}{(z+\lambda_j)^3} \exp\left(-\sum_{i'} \frac{1}{(z+\lambda_{i'})^2}\right) \mathrm{d}z.$$

The derivative of $I_{i,4}(\lambda)/I_{i,3}(\lambda)$ is expressed as

$$\frac{\mathrm{d}}{\mathrm{d}\lambda_j} I_{i,4}(\lambda)/I_{i,3}(\lambda) = \frac{2I_{i,j,4}(\lambda)I_{i,3}(\lambda) - 2I_{i,4}(\lambda)I_{i,j,3}(\lambda)}{I_{i,3}^2(\lambda)}. \tag{9}$$

By letting $f(z) = (z+\lambda_i)^{-3}\mathrm{e}^{-\sum_{i'}\frac{1}{(z+\lambda_{i'})^2}} > 0$, each term of the numerator of (9) is expressed as

$$\begin{aligned}
I_{i,j,4}(\lambda)I_{i,3}(\lambda) &= \iint_{z,w\geq 0} \frac{f(z)f(w)}{(z+\lambda_i)(z+\lambda_j)^3}\mathrm{d}z\mathrm{d}w \\
&= \frac{1}{2}\iint_{z,w\geq 0} f(z)f(w)\left(\frac{1}{(z+\lambda_i)(z+\lambda_j)^3} + \frac{1}{(w+\lambda_i)(w+\lambda_j)^3}\right)\mathrm{d}z\mathrm{d}w, \\
I_{i,4}(\lambda)I_{i,j,3}(\lambda) &= \iint_{z,w\geq 0} \frac{f(z)f(w)}{(z+\lambda_i)(w+\lambda_j)^3}\mathrm{d}z\mathrm{d}w \\
&= \frac{1}{2}\iint_{z,w\geq 0} f(z)f(w)\left(\frac{1}{(z+\lambda_i)(w+\lambda_j)^3} + \frac{1}{(w+\lambda_i)(z+\lambda_j)^3}\right)\mathrm{d}z\mathrm{d}w.
\end{aligned}$$

Here, by an elementary calculation we can see

$$\begin{aligned}
&\frac{1}{(z+\lambda_i)(z+\lambda_j)^3} + \frac{1}{(w+\lambda_i)(w+\lambda_j)^3} - \frac{1}{(z+\lambda_i)(w+\lambda_j)^3} - \frac{1}{(w+\lambda_i)(z+\lambda_j)^3} \\
&= \frac{(w+\lambda_i)(w+\lambda_j)^3 + (z+\lambda_i)(z+\lambda_j)^3 - (w+\lambda_i)(z+\lambda_j)^3 - (z+\lambda_i)(w+\lambda_j)^3}{(z+\lambda_i)(w+\lambda_i)(z+\lambda_j)^3(w+\lambda_j)^3} \\
&= (w-z)^2\frac{(w+\lambda_j)^2 + (w+\lambda_j)(z+\lambda_j) + (z+\lambda_j)^2}{(z+\lambda_i)(w+\lambda_i)(z+\lambda_j)^3(w+\lambda_j)^3} > 0,
\end{aligned}$$

which means that $I_{i,j,4}(\lambda)I_{i,3}(\lambda) - I_{i,4}(\lambda)I_{i,j,3}(\lambda)$ can be expressed as an integral of a positive function, implying that $I_{i,4}(\lambda)/I_{i,3}(\lambda)$ is increasing in λ_j. ■

Proof of Lemma 5 By the monotonicity of $I_{i,4}(\lambda)/I_{i,3}(\lambda)$ in Lemma 6, we have

$$I_{i,4}(\underline{\lambda})/I_{i,3}(\underline{\lambda}) \leq I_{i,4}(\lambda^*)/I_{i,3}(\lambda^*), \qquad \text{where } \lambda_j^* = \begin{cases} \underline{\lambda}_i & j \leq i, \\ \infty & j > i. \end{cases} \tag{10}$$

The RHS of (10) is expressed by incomplete gamma function $\gamma(k,x) = \int_0^x \mathrm{e}^{-t}t^{k-1}\mathrm{d}t$ as

$$\begin{aligned}
I_{i,n}(\lambda^*) &= \int_0^\infty \frac{1}{(z+\underline{\lambda}_i)^n}\mathrm{e}^{-i/(z+\underline{\lambda}_i)^2}\mathrm{d}z \\
&= i^{-(n-1)/2}\int_0^{i/\underline{\lambda}_i^2} u^{(n-3)/2}\mathrm{e}^{-u}\mathrm{d}u \\
&= i^{-(n-1)/2}\gamma((n-1)/2, i/\underline{\lambda}_i^2).
\end{aligned} \tag{11}$$

We can see that $\gamma(3/2,x)/\gamma(1,x) \leq 2\sqrt{x}/3 \wedge \sqrt{\pi}/2$ for $x > 0$ by an elementary calculation (see Appendix F.2). Then we have

$$\frac{I_{i,4}(\underline{\lambda})}{I_{i,3}(\underline{\lambda})} \leq i^{-1/2}\frac{\gamma(3/2, i/\underline{\lambda}_i^2)}{\gamma(1, i/\underline{\lambda}_i^2)} \leq i^{-1/2}\left(\frac{2\sqrt{i}}{3\underline{\lambda}_i} \wedge \frac{\sqrt{\pi}}{2}\right) = \frac{2}{3\underline{\lambda}_i} \wedge \frac{\sqrt{\pi/i}}{2}.$$

■

5. Optimal Adversarial Bound

In this section we complete the proof of Theorem 1.

5.1. Stability Term

By using the result of the last section, we can express the stability term as follows.

Lemma 7 *For any $i \in [K]$, if $\hat{L}_{t,i}$ is the $\sigma_{t,i}$-th smallest among $\{\hat{L}_{t,j}\}$ then*

$$\mathbb{E}\left[\hat{\ell}_{t,i}\left(\phi_i(\eta_t\hat{L}_t) - \phi_i(\eta_t(\hat{L}_t + \hat{\ell}_t))\right)\middle|\hat{L}_t\right] \le \frac{4}{\underline{\hat{L}}_{t,i}} \wedge 3\eta_t\sqrt{\pi/\sigma_{t,i}}.$$

As can be seen from the proof of this lemma, this bound is two times larger than the one for the case where $1/w_{t,i}$ is exactly computed instead of $\widehat{w_{t,i}^{-1}}$, which is the cost for easier computation.

Proof First we have

$$\begin{aligned}
\phi_i(\eta_t\hat{L}_t) - \phi_i(\eta_t(\hat{L}_t + (\ell_{t,i}\widehat{w_{t,i}^{-1}})e_i)) &= \int_0^{\eta_t\ell_{t,i}\widehat{w_{t,i}^{-1}}} (-\phi_i'(\eta_t\hat{L}_t + xe_i))\mathrm{d}x \\
&\le 6\int_0^{\eta_t\ell_{t,i}\widehat{w_{t,i}^{-1}}} I_{i,4}(\underline{\eta_t\hat{L}_t + xe_i})\mathrm{d}x \quad \text{(by (7))} \\
&\le 6\int_0^{\eta_t\ell_{t,i}\widehat{w_{t,i}^{-1}}} I_{i,4}(\eta_t\underline{\hat{L}_t})\mathrm{d}x \qquad (12) \\
&= 6\eta_t\ell_{t,i}I_{i,4}(\eta_t\underline{\hat{L}_t})\widehat{w_{t,i}^{-1}},
\end{aligned}$$

where (12) follows from the monotonicity of $I_{i,4}$. Here note that $\widehat{w_{t,i}^{-1}}$ follows the geometric distribution with expectation $1/w_{t,i}$ given $\hat{L}_t$ and I_t, which satisfies

$$\mathbb{E}\left[\widehat{w_{t,I_t}^{-1}}^2\middle|\hat{L}_t, I_t\right] = \frac{2}{w_{t,I_t}^2} - \frac{1}{w_{t,I_t}} \le \frac{2}{w_{t,I_t}^2}. \qquad (13)$$

Since $\hat{\ell}_t = (\ell_{t,i}\widehat{w_{t,i}^{-1}})e_i$ when $I_t = i$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\hat{\ell}_{t,i}(\phi_i(\eta_t\hat{L}_t) - \phi_i(\eta_t(\hat{L}_t + \hat{\ell}_t)))\middle|\hat{L}_t\right] &\le \mathbb{E}\left[\mathbb{1}\left[I(t) = i\right]\ell_{t,i}\widehat{w_{t,i}^{-1}} \cdot 6\eta_t\ell_{t,i}I_{i,4}(\eta_t\underline{\hat{L}_t})\widehat{w_{t,i}^{-1}}\middle|\hat{L}_t\right] \\
&\le 12\eta_t\mathbb{E}\left[w_{t,i}\frac{\ell_{t,i}^2 I_{i,4}(\eta_t\underline{\hat{L}_t})}{w_{t,i}^2}\middle|\hat{L}_t\right] \\
&\le 6\eta_t\mathbb{E}\left[\frac{I_{i,4}(\eta_t\underline{\hat{L}_t})}{I_{i,3}(\eta_t\underline{\hat{L}_t})}\middle|\hat{L}_t\right] \quad \left(\text{by } w_{t,i} = 2I_{i,3}(\eta_t\underline{\hat{L}_t})\right) \\
&\le \frac{4\eta_t}{\eta_t\underline{\hat{L}}_{t,i}} \wedge 3\eta_t\sqrt{\pi/\sigma_{t,i}} \quad \text{(by Lemma 5)}.
\end{aligned}$$

■

Lemma 7 immediately leads to the following bound, which is used for both the adversarial and stochastic settings.

Lemma 8 *For any $\hat{L}_t$,*

$$\mathbb{E}\left[\left\langle\hat{\ell}_t, \phi(\eta_t\hat{L}_t) - \phi(\eta_t(\hat{L}_t + \hat{\ell}_t))\right\rangle\middle|\hat{L}_t\right] \le 6\eta_t\sqrt{\pi K}.$$

Proof By Lemma 7,

$$\begin{aligned}\mathbb{E}\left[\left\langle \hat{\ell}_t, \phi(\eta_t \hat{L}_t) - \phi(\eta_t(\hat{L}_t + \hat{\ell}_t))\right\rangle \middle| \hat{L}_t\right] &\le \sum_{i=1}^{K} 3\eta_t \sqrt{\pi/\sigma_{t,i}} \\ &\le 3\eta_t\sqrt{\pi}\left(1 + \int_1^K x^{-1/2}\mathrm{d}x\right) \\ &= 3\eta_t\sqrt{\pi}(1 + 2(\sqrt{K} - 1)) \le 6\eta_t\sqrt{\pi K}. \end{aligned}$$

■

5.2. Penalty Term

For the penalty term we have the following bound.

Lemma 9

$$\mathbb{E}\left[r_{t,I(t)} - r_{t,i^*}\middle|\hat{L}_t\right] \le \left(2\sum_{i\neq i^*}\frac{1}{\eta_t \underline{\hat{L}}_{t,i}}\right) \wedge 3.7\sqrt{K}.$$

Here we only use the bound $3.7\sqrt{K}$ for the adversarial setting, whereas we use both of the bounds for the stochastic setting.

Proof By letting $f(z) = \sum_i \frac{1}{(z+\eta_t \underline{\hat{L}}_{t,i})^2} \in \left(0, \frac{K}{z^2}\right]$ we have

$$\begin{aligned}\mathbb{E}[r_{t,I(t)} - r_{t,i^*}|\hat{L}_t] &\le \sum_{i\neq i^*}\mathbb{E}[\mathbb{1}\left[I(t) = i\right] r_{t,i}|\hat{L}_t] \\ &= 2\int_0^\infty \sum_{i\neq i^*}\frac{1}{(z+\eta_t\underline{\hat{L}}_{t,i})^2}\mathrm{e}^{-f(z)}\mathrm{d}z \qquad (14) \\ &\le 2\int_0^\infty \sum_{i\neq i^*}\frac{1}{(z+\eta_t\underline{\hat{L}}_{t,i})^2}\mathrm{d}z = 2\sum_{i\neq i^*}\frac{1}{\eta_t\underline{\hat{L}}_{t,i}}.\end{aligned}$$

We can also bound (14) by

$$\begin{aligned}2\int_0^\infty \sum_{i\neq i^*}\frac{1}{(z+\eta\hat{L}_i)^2}\mathrm{e}^{-f(x)}\mathrm{d}z &\le 2\int_0^\infty f(z)\mathrm{e}^{-f(z)}\mathrm{d}z \\ &= 2\int_0^{\sqrt{K}} f(z)\mathrm{e}^{-f(z)}\mathrm{d}z + 2\int_{\sqrt{K}}^\infty f(z)\mathrm{e}^{-f(z)}\mathrm{d}z \\ &\le 2\int_0^{\sqrt{K}} \mathrm{e}^{-1}\mathrm{d}z + 2\int_{\sqrt{K}}^\infty \frac{K}{z^2}\exp\left(-\frac{K}{z^2}\right)\mathrm{d}z \qquad (15) \\ &= 2\mathrm{e}^{-1}\sqrt{K} + \sqrt{K}\int_0^1 w^{-1/2}\mathrm{e}^w\mathrm{d}w \le 3.7\sqrt{K}.\end{aligned}$$

Here, for the first term of (15) we used the fact $x\mathrm{e}^{-x} \le \mathrm{e}^{-1}$. For the second term, we used the fact that $x\mathrm{e}^{-x}$ is increasing in x for $x \le 1$ and $f(z) \le 1$ holds for $z \ge \sqrt{K}$. ■

5.3. Proof of Theorem 1

By combining Lemmas 3, 4, 8 and 9 with $\eta_t = c/\sqrt{t}$ we have

$$\begin{aligned}\mathrm{Regret}(T) &\le 6c\sqrt{\pi K}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} + \frac{3.7\sqrt{K}}{c}\sum_{t=1}^{T}\left(\sqrt{t+1}-\sqrt{t}\right) + 10c + \frac{\sqrt{\pi K}}{c}\\ &\le \left(12c\sqrt{\pi T} + \frac{3.7(\sqrt{T+1}-1)}{c}\right)\sqrt{K} + 10c + \frac{\sqrt{\pi K}}{c}\\ &\le \left(12c\sqrt{\pi} + \frac{3.7}{c}\right)\sqrt{KT} + 10c + \frac{\sqrt{\pi K}}{c},\end{aligned}$$

where the last inequality follows from $\sqrt{T+1}-1 \le \sqrt{T}$. ■

6. Outline for Stochastic Regret Bound

In this section we explain how to derive a logarithmic regret bound for the stochastic setting by the self-bounding technique.

The regret in the stochastic setting is expressed as

$$\mathrm{Regret}(T) = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*} w_{t,i}\Delta_i\right]. \tag{16}$$

A typical analysis of FTRL (see, e.g., Zimmert and Seldin, 2021) derives a bound of form $\mathrm{Regret}(T) \le \mathbb{E}\left[\sum_{t\in[T]}\sum_{i\neq i^*} O(\sqrt{w_{t,i}/t})\right]$ by excluding the regret associated to the optimal arm i^* from the adversarial bound. Subtracting (16)/2 from this bound, we obtain

$$\frac{\mathrm{Regret}(T)}{2} \le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*} O\left(\sqrt{w_{t,i}/t} - \frac{\Delta_i w_{t,i}}{2}\right)\right].$$

This yields an upper bound $\sum_{i\neq i^*} O(\frac{\log T}{\Delta_i})$ when the worst case of $w_{t,i}$ is taken.

In the analysis of FTPL, it is difficult to exclude the regret from the optimal arm in the same way. Still, we can derive a similar result when we consider t such that $A_t = \{\sum_{i\neq i^*}(\eta_t \underline{\hat{L}}_{t,i})^{-2} \le 1\}$, which is the event that the estimated losses $\hat{L}_{t,i}$ of the suboptimal arms $i \neq i^*$ are sufficiently large compared with $\hat{L}_{t,i^*}$. As we will see in Appendix D.2, we can bound the stability given $\underline{\hat{L}}_t$ associated to the optimal arm by $\sum_{i\neq i^*} 1/\underline{\hat{L}}_{t,i}$ when A_t holds. Combining this with Lemmas 7–9, we can obtain

$$\mathrm{Regret}(T) \le \mathbb{E}\left[\sum_{t=1}^{T} O\left(\mathbb{1}\left[A_t\right]\sum_{i\neq i^*}\frac{1}{\underline{\hat{L}}_{t,i}} + \mathbb{1}\left[A_t^c\right]\sqrt{K/t}\right)\right].$$

Similarly, by the analysis depending on A_t and A_t^c, we can also obtain a regret lower bound of form

$$\mathrm{Regret}(T) \ge \mathbb{E}\left[\sum_{t=1}^{T} O\left(\mathbb{1}\left[A_t\right]\sum_{i\neq i^*}\frac{t\Delta_i}{\underline{\hat{L}}_{t,i}^2} + \mathbb{1}\left[A_t^c\right]\Delta\right)\right]$$

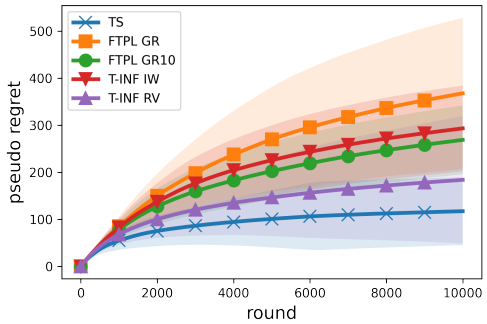

Figure 1: Results for stochastic setting.

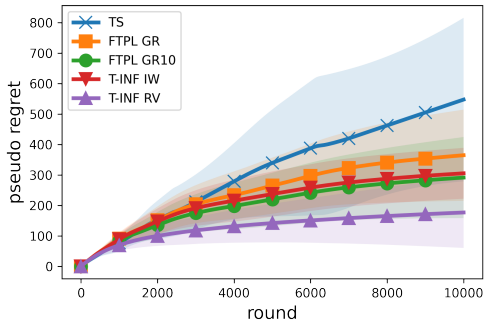

Figure 2: Results for adversarial setting.

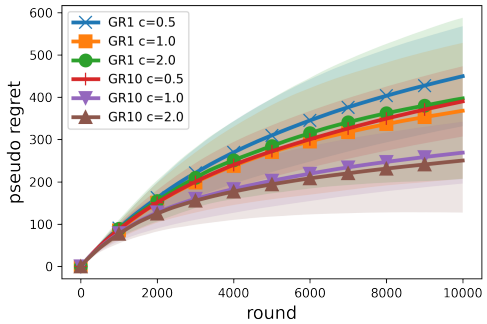

Figure 3: Results for stochastic setting.

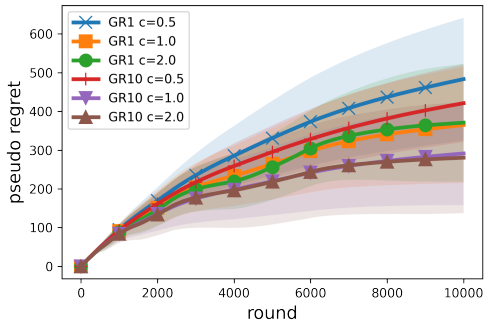

Figure 4: Results for adversarial setting.

as shown in Appendix D.1. By combining these bounds we obtain

$$\frac{\mathrm{Regret}(T)}{2} \leq \mathbb{E}\left[\sum_{t=1}^{T} O\left(\mathbb{1}\left[A_t\right] \sum_{i \neq i^*} \left(\frac{1}{\underline{\hat{L}}_{t,i}} - \frac{t\Delta_i}{2\underline{\hat{L}}_{t,i}^2}\right) + \mathbb{1}\left[A_t^c\right]\left(\sqrt{K/t} - \Delta/2\right)\right)\right],$$

which yields a logarithmic regret bound when the worst case of $\mathbb{1}\left[A_t\right]$ and $\hat{L}_t$ is taken. By formalizing this discussion we prove Theorem 2 in Appendix D. See Appendix E for extension to the adversarial setting with a self-bounding constraint.

7. Experiments

In this section we examine the empirical performance of FTPL and other policies. Following Zimmert and Seldin (2021), we consider the eight-armed bandit under the stochastic setting and the stochastically constrained adversarial setting. See Appendix B for details of the settings as well as the runtime for the experiments.

We compare FTPL with geometric resampling (FTPL GR) with Thomson sampling (TS) and Tsallis-INF. For Tsallis-INF, there is also a loss estimator called Reduced-Variance (RV, Zimmert and Seldin, 2021) estimator as well as the IW estimator considered in this paper. We write T-INF IW and T-INF RV for Tsallis-INF with these estimators. Note that the RV estimator explicitly requires the value of $w_{t,i}$ and is not applicable to FTPL. We also consider a stable variant of geometric resampling (GR10), where resampling (Lines 6–9 in Algorithm 1) is repeated ten times and its average is taken, which makes the variance of $\widehat{w_{t,i}^{-1}}$ ten times smaller.

Figures 1 and 2 are the results to compare FTPL with other policies. For T-INF, we used the same learning rate as Zimmert and Seldin (2021) ($\eta_t = c/\sqrt{t}$ with $c = 2$ for IW and $c = 4$ for RV). For FTPL we used the same learning rate as that for T-INF IW, since FTPL with Fréchet perturbation is designed to mimic it (see Kim and Tewari, 2019). As shown there, FTPL and T-INF perform stably in both settings, while TS designed for the stochastic setting performs poorly in the adversarial setting. We can also see that the behavior of FTPL GR10 is very similar to T-INF IW and FTPL GR is a little worse, which seems to be due to the larger variance of the loss estimator by GR. Since T-INF RV performed much better than T-INF IW, it is an important future work to devise a counterpart of T-INF RV applicable to FTPL.

Figures 3 and 4 are the results to check the effect of learning rate and the geometric resampling. In this experiment, we considered FTPL with parameter $c = 0.5,\ 1,\ 2$. We can see from the figures that the performance becomes better for larger c compared with theoretically suggested small ones and the stable version improves the performance for each choice of c.

8. Conclusion

In this paper we tackled the open problem on the optimality of FTPL by Kim and Tewari (2019), and affirmatively resolved it by showing that FTPL with Fréchet perturbation achieves $O(\sqrt{KT})$ adversarial regret. We also derived $O(\sum_{i \neq i^*} \frac{\log T}{\Delta_i})$ stochastic regret bound, meaning that FTPL has the BOBW property. Still, the constant factor of the bound is very large and we confirmed that there are currently some gap between the bound-optimizing learning rate and empirically suggested one.

One of the most important future work is to extend this result to the setting with exponentially many actions like the combinatorial bandits. As a first step, we expect that FTPL with Frechét perturbation can be used for m-sets semi-bandits to achieve $\sqrt{mKT}$ optimal adversarial regret (Audibert et al., 2014), where m arms are pulled at each round. Still, its analysis is currently open since the arm-selection probability $\phi_i(\cdot)$ becomes much more complicated than the non-combinatorial setting, though we believe that the technique of this paper becomes a clue to the analysis.

Extension of the BOBW result is further nontrivial. This is because the existing BOBW policies for the combinatorial semi-bandits use a hybrid regularization in addition to the Tsallis-entropy (Zimmert et al., 2019). Since Fréchet distribution roughly (though not exactly) corresponds to Tsallis-entropy regularization, it would be a very challenging task to realize the effect of hybrid regularization by FTPL.

Related to the above point, another remaining open problem is to answer to the question raised by Kim and Tewari (2019) in a more general form. Whereas this paper showed that the Fréchet perturbation achieves $O(\sqrt{KT})$ regret, the current analysis heavily depends on the specific form of this distribution, and it is unclear how this result can be extended to general distributions (seemingly with Frechét-type tails). Since many regularization functions have been considered in FTRL policies with BOBW properties, revealing more general conditions for the optimal regret would be helpful for construction of the counterparts of such FTRL policies.

Acknowledgments

JH was supported by JSPS, KAKENHI Grant Number JP21K11747, Japan. TT was supported by JST, ACT-X Grant Number JPMJAX210E, Japan and JSPS, KAKENHI Grant Number JP21J21272, Japan.

References

Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *Conference on Learning Theory*, pages 807–823. PMLR, 2014.

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

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

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for Thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 99–107. PMLR, 2013.

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

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002a.

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

Sébastien Bubeck. Five Miracles of Mirror Descent. Lecture note of HDPA-2019: High dimensional probability and algorithms, 2019. URL https://hdpa2019.sciencesconf.org/data/pages/bubeck_hdpa.pdf.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, volume 23, pages 42.1–42.23. PMLR, 2012.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013. ISSN 0090-5364.

Yan Dai, Haipeng Luo, and Liyu Chen. Follow-the-perturbed-leader for adversarial Markov decision processes with bandit feedback. In *Advances in Neural Information Processing Systems*, 2022.

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

Miroslav Dudík, Nika Haghtalab, Haipeng Luo, Robert E Schapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. Oracle-efficient online learning and auction design. *Journal of the ACM (JACM)*, 67(5):1–57, 2020.

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

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.

James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3(2):97–139, 1957.

Ruitong Huang, Tor Lattimore, András György, and Csaba Szepesvári. Following the leader and fast rates in linear prediction: Curved constraint sets and other regularities. In *Advances in Neural Information Processing Systems*, pages 4970–4978, 2016.

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

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

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

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

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

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: an asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213, 2012.

Baekjin Kim and Ambuj Tewari. On the optimality of perturbations in stochastic and adversarial multi-armed bandit problems. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pages 6142–6151, 2021.

Haipeng Luo and Robert E Schapire. Achieving all with no parameters: AdaNormalHedge. In *Conference on Learning Theory*, pages 1286–1304. PMLR, 2015.

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

H Brendan McMahan and Avrim Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *International Conference on Computational Learning Theory*, pages 109–123, 2004.

Gergely Neu. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, pages 1360–1375. PMLR, 2015.

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

Charles Riou and Junya Honda. Bandit algorithms based on Thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory*, pages 777–826. PMLR, 2020.

Vasilis Syrgkanis, Akshay Krishnamurthy, and Robert Schapire. Efficient algorithms for adversarial contextual learning. In *International Conference on Machine Learning*, pages 2159–2168. PMLR, 2016.

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

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. Relation between Pseudo-regret and Regret

In this paper we consider the pseudo-regret

$$\mathbb{E}\left[\sum_{t=1}^{T}(\ell_{t,I_t}-\ell_{t,i^*})\right], \qquad i^* \in \operatorname*{argmin}_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,i}\right].$$

On the other hand, in the adversarial bandit the *regret* given by

$$\mathbb{E}\left[\sum_{t=1}^{T}(\ell_{t,I_t}-\ell_{t,i^\star})\right], \qquad i^\star \in \operatorname*{argmin}_{i\in[K]} \sum_{t=1}^{T}\ell_{t,i}.$$

is also sometimes considered, where the best arm $i^\star$ becomes a random variable. In BOBW literature, Zimmert and Seldin (2021) consider the pseudo-regret as in this paper, while Kim and Tewari (2019) consider the regret. There is no essential difference between these styles; in these papers and this paper, if we consider the regret rather than the pseudo-regret then we need to assume an *oblivious adversary*, which determines all losses before the game begins. In this paper we consider the pseudo-regret just for a unified discussion of the adversarial and stochastic settings.

Appendix B. Experiment Details

The empirical evaluation presented in Section 7 is performed in problem setups similar to those by Zimmert and Seldin (2021). Bandit policies are evaluated in the *stochastic setting* as well as the *stochastically constrained adversarial setting* (or adversarial setting in short). In both settings, values of losses $\ell_{t,i} \in \{0,1\}$ are generated independently from Bernoulli distributions for $K = 8$ arms including a single optimal arm. Difference between stochastic and adversarial settings is summarized as follows:

Stochastic setting In the stochastic setting, the mean losses are chosen to be $(1-\Delta)/2$ for the optimal arm and $(1+\Delta)/2$ for the other suboptimal arms, where $\Delta > 0$ is a parameter corresponding to the suboptimality gap.

Stochastically constrained adversarial setting In the stochastically constrained adversarial setting, the mean losses for the optimal arm and the other suboptimal arms switch between $(1-\Delta, 1)$ and $(0, \Delta)$. The time between alternations increases exponentially with factor 1.6 after each switch, similarly to that in Zimmert and Seldin (2021).

Table 1: Runtime (sec) of policies for 10,000 rounds and different K.

	$K=4$	$K=8$	$K=16$	$K=32$
Thompsom sampling	0.34	0.42	0.58	0.91
Tsallis-INF	2.16	2.29	2.45	2.68
FTPL	0.85	1.27	2.19	4.15

In our experiments, the suboptimality gap parameter is chosen to be $\Delta = 0.125$ for both settings. The time horizon is chosen to be 10,000. In all experiments, the pseudo-regret is estimated by 100 repetitions. We also compute standard deviations of these empirical pseudo-regret, which are depicted by the shaded areas. For Tsallis-INF, w_t is computed by Newton's method. The runtime for the algorithms is listed in Table 1.

Appendix C. Regret Decomposition

In this appendix we provide proofs of lemmas on the decomposition of the regret in Section 3.2.

Lemma 3 (restated)

$$\mathrm{Regret}(T) \le \sum_{t=1}^{T} \mathbb{E}\left[\left\langle \hat{\ell}_t, w_t - w_{t+1} \right\rangle\right] + \sum_{t=1}^{T} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \mathbb{E}\left[r_{t+1,I_{t+1}} - r_{t+1,i^*}\right] + \frac{\sqrt{\pi K}}{\eta_1}.$$

Proof of Lemma 3 Let us consider random variable $r \in [0,\infty)^K$ that independently follows Fréche distribution and is independent from the randomness $\{\ell_t, r_t\}_{t=1}^T$ of the environment and the policy. Define $u_t = \mathrm{argmin}_{w \in \mathcal{P}_K} \left\langle \eta_t \hat{L}_t - r, w \right\rangle$, where $\mathcal{P}_K = \{p \in [0,1]^K : \sum_{i\in[K]} p_i = 1\}$ is the $(K-1)$-dimensional probability simplex. Then, since r_t and r are identically distributed given $\hat{L}_t$, we have

$$\mathbb{E}[u_t | \hat{L}_t] = w_t, \quad \mathbb{E}[\langle r, u_t \rangle | \hat{L}_t] = \mathbb{E}[\langle r_t, e_{I(t)} \rangle | \hat{L}_t] = \mathbb{E}[r_{t,I(t)} | \hat{L}_t]. \tag{17}$$

Recalling $\hat{L}_t = \sum_{s=1}^{t-1} \hat{\ell}_s$ we have

$$\begin{aligned}
\sum_{t=1}^{T} \left\langle \hat{\ell}_t, e_{i^*} \right\rangle &= \left\langle \hat{L}_{T+1}, e_{i^*} \right\rangle \\
&= \left\langle \hat{L}_{T+1} - \frac{1}{\eta_{T+1}} r, e_{i^*} \right\rangle + \frac{1}{\eta_{T+1}} \langle r, e_{i^*} \rangle \\
&\ge \left\langle \hat{L}_{T+1} - \frac{1}{\eta_{T+1}} r, u_{T+1} \right\rangle + \frac{1}{\eta_{T+1}} \langle r, e_{i^*} \rangle \\
&= \left\langle \hat{L}_{T} - \frac{1}{\eta_{T}} r, u_{T+1} \right\rangle + \left\langle \hat{\ell}_T, u_{T+1} \right\rangle - \left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_T}\right) \langle r, u_{T+1} \rangle + \frac{1}{\eta_{T+1}} \langle r, e_{i^*} \rangle \\
&\ge \left\langle \hat{L}_{T} - \frac{1}{\eta_{T}} r, u_{T} \right\rangle + \left\langle \hat{\ell}_T, u_{T+1} \right\rangle - \left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_T}\right) \langle r, u_{T+1} \rangle + \frac{1}{\eta_{T+1}} \langle r, e_{i^*} \rangle
\end{aligned}$$

and recursively applying this relation we obtain

$$\sum_{t=1}^{T} \left\langle \hat{\ell}_t, e_{i^*} \right\rangle \ge \left\langle -\frac{1}{\eta_1} r, u_1 \right\rangle + \sum_{t=1}^{T} \left(\left\langle \hat{\ell}_t, u_{t+1} \right\rangle - \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \langle r, u_{t+1} \rangle \right) + \frac{1}{\eta_{T+1}} \langle r, e_{i^*} \rangle$$

and therefore

$$\sum_{t=1}^{T} \left\langle \hat{\ell}_t, u_t - e_{i^*} \right\rangle \le \frac{1}{\eta_1} \langle r, u_1 - e_{i^*} \rangle + \sum_{t=1}^{T} \left(\left\langle \hat{\ell}_t, u_t - u_{t+1} \right\rangle + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \langle r, u_{t+1} - e_{i^*} \rangle \right).$$

By using (17) and taking the expectation with respect to r we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \left\langle \hat{\ell}_t, w_t - e_{i^*} \right\rangle \\
&\leq \frac{1}{\eta_1} \mathbb{E}_{r \sim \mathcal{F}_2} \left[\langle r, u_1 - e_{i^*} \rangle \right] + \sum_{t=1}^{T} \left(\left\langle \hat{\ell}_t, w_t - w_{t+1} \right\rangle + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \mathbb{E}_{r \sim \mathcal{F}_2} \left[\langle r, u_{t+1} - e_{i^*} \rangle \right] \right) \\
&\leq \frac{1}{\eta_1} \mathbb{E}_{r_1 \sim \mathcal{F}_2} \left[r_{1,I_1} \right] + \sum_{t=1}^{T} \left(\left\langle \hat{\ell}_t, w_t - w_{t+1} \right\rangle + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \mathbb{E}_{r_{t+1} \sim \mathcal{F}_2} \left[r_{t+1,I_{t+1}} - r_{t+1,i^*} \middle| \hat{L}_{t+1} \right] \right).
\end{aligned}$$

Note that $r_{1,I_1} = \max_{i \in [K]} r_{1,i}$. Since its cumulative distribution function is given by e^{-K/z^2} with density $2Kz^{-3}\mathrm{e}^{-K/z^2}$, we have

$$\frac{1}{\eta_1} \mathbb{E}_{r \sim \mathcal{F}_2} \left[r_{1,I_1} \right] = \frac{2K}{\eta_1} \int_0^\infty \frac{1}{z^3} \mathrm{e}^{-K/z^2} \mathrm{d}z = \frac{\sqrt{\pi K}}{\eta_1},$$

which completes the proof. ■

Lemma 4 (restated)

$$\sum_{t=1}^{T} \mathbb{E} \left[\left\langle \hat{\ell}_t, w_t - w_{t+1} \right\rangle \right] \leq \sum_{t=1}^{T} \mathbb{E} \left[\left\langle \hat{\ell}_t, \phi(\eta_t \hat{L}_t) - \phi(\eta_t (\hat{L}_t + \hat{\ell}_t)) \right\rangle \right] + 10\eta_1.$$

Proof of Lemma 4 By (3) we have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E} \left[\left\langle \hat{\ell}_t, w_t - w_{t+1} \right\rangle \right] &= \sum_{t=1}^{T} \mathbb{E} \left[\left\langle \hat{\ell}_t, \phi(\eta_t \hat{L}_t) - \phi(\eta_t (\hat{L}_t + \hat{\ell}_t)) \right\rangle \right] \\
&\quad + \sum_{t=1}^{T} \mathbb{E} \left[\left\langle \hat{\ell}_t, \phi(\eta_t (\hat{L}_t + \hat{\ell}_t)) - \phi(\eta_{t+1} (\hat{L}_t + \hat{\ell}_t)) \right\rangle \right] \qquad (18)
\end{aligned}$$

and we bound the second term of (18) in the following. Each component of this term is expressed as

$$\begin{aligned}
&\mathbb{E} \left[\left\langle \hat{\ell}_t, \phi(\eta_t (\hat{L}_t + \hat{\ell}_t)) - \phi(\eta_{t+1} (\hat{L}_t + \hat{\ell}_t)) \right\rangle \right] \\
&= \sum_{i \in [K]} \mathbb{E} \left[\mathbb{1} \left[I_t = i \right] \delta_i \left(\phi_i(\eta_t (\hat{L}_t + \delta_i e_i)) - \phi_i(\eta_{t+1} (\hat{L}_t + \delta_i e_i)) \right) \right], \qquad (19)
\end{aligned}$$

where we write $\delta_i = \ell_{t,i} \widehat{w_{t,i}^{-1}}$.

On the other hand, for generic $L \in \mathbb{R}^K$ we have

$$\begin{aligned}
\frac{\partial}{\partial \eta}\phi_i(\eta L) &= 2\int_0^\infty \left(\frac{1}{(z+\eta \underline{L}_i)^3}\sum_{i'}\frac{2\underline{L}_{i'}}{(z+\eta\underline{L}_{i'})^3} - \frac{3\underline{L}_i}{(z+\eta\underline{L}_i)^4}\right)\exp\left(-\sum_{i'}\frac{1}{(z+\eta\underline{L}_{i'})^2}\right)\mathrm{d}z \\
&\le 4\int_0^\infty \frac{1}{(z+\eta \underline{L}_i)^3}\sum_{i'}\frac{1}{(z+\eta\underline{L}_{i'})^2}\exp\left(-\sum_{i'}\frac{1}{(z+\eta\underline{L}_{i'})^2}\right)\mathrm{d}z \qquad (20)\\
&\le 4\int_0^\infty \left(\sum_{i'}\frac{1}{(z+\eta \underline{L}_{i'})^3}\right)\left(\sum_{i'}\frac{1}{(z+\eta\underline{L}_{i'})^2}\right)\exp\left(-\sum_{i'}\frac{1}{(z+\eta\underline{L}_{i'})^2}\right)\mathrm{d}z \\
&\le 2\int_0^\infty w\mathrm{e}^{-w}\mathrm{d}w = 2. \qquad (21)
\end{aligned}$$

Now consider the case where L_i is not the unique minimizer of $\{L_j\}$. In this case, by (20) we have

$$\begin{aligned}
\frac{\partial}{\partial \eta}\phi_i(\eta(L+\delta_i e_i)) &\le 4\int_0^\infty \frac{1}{(z+\eta \underline{L}_i)^3}\sum_{i'}\frac{1}{(z+\eta\underline{L}_{i'})^2}\exp\left(-\sum_{i'\neq i}\frac{1}{(z+\eta\underline{L}_{i'})^2}\right)\mathrm{d}z \\
&\le 4\int_0^\infty \frac{1}{(z+\eta \underline{L}_i)^3}\sum_{i'}\frac{1}{(z+\eta\underline{L}_{i'})^2}\exp\left(-\sum_{i'}\frac{1}{2(z+\eta\underline{L}_{i'})^2}\right)\mathrm{d}z, \qquad (22)
\end{aligned}$$

where the last inequality holds since there exists $i' \neq i$ such that $\underline{L}_{i'} = 0$ when L_i is not the unique minimizer of $\{L_j\}$.

Now, let us write $\hat{i}_t^* \in \mathrm{argmin}_i \hat{L}_{t,i}$ for an arbitrarily broken tie and consider bounding $\frac{\partial}{\partial\eta}\phi_i(\cdot)$ by (21) if i is the unique minimizer of $\{L_j\}$ and by (22) otherwise. Then (21) is applied at most once and therefore (19) is bounded by

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}\left[\left\langle \hat{\ell}_t, \phi(\eta_t(\hat{L}_t+\hat{\ell}_t)) - \phi(\eta_{t+1}(\hat{L}_t+\hat{\ell}_t))\right\rangle\right] \\
&\le 2(\eta_t-\eta_{t+1})\mathbb{E}\left[\mathbb{1}\left[I_t = \hat{i}_t^*\right]\delta_{\hat{i}_t^*}\right] \\
&\quad + 4\sum_{i\in[K]}\mathbb{E}\left[\int_{\eta_{t+1}}^{\eta_t}\mathbb{1}\left[I_t=i\right]\delta_i\int_0^\infty \frac{1}{(z+\eta \underline{L}_i)^3}\sum_{i'}\frac{1}{(z+\eta\underline{L}_{i'})^2}\exp\left(-\sum_{i'}\frac{1}{2(z+\eta\underline{L}_{i'})^2}\right)\mathrm{d}z\mathrm{d}\eta\right] \\
&= 2(\eta_t-\eta_{t+1})\mathbb{E}\left[\ell_{t,\hat{i}_t^*}\right] \\
&\quad + 4\sum_{i\in[K]}\mathbb{E}\left[\int_{\eta_{t+1}}^{\eta_t}\ell_{t,i}\int_0^\infty \frac{1}{(z+\eta \underline{L}_i)^3}\sum_{i'}\frac{1}{(z+\eta\underline{L}_{i'})^2}\exp\left(-\sum_{i'}\frac{1}{2(z+\eta\underline{L}_{i'})^2}\right)\mathrm{d}z\mathrm{d}\eta\right] \\
&\le 2(\eta_t-\eta_{t+1}) \\
&\quad + 4\mathbb{E}\left[\int_{\eta_{t+1}}^{\eta_t}\int_0^\infty \left(\sum_{i\in[K]}\frac{1}{(z+\eta \underline{L}_i)^3}\right)\sum_{i'}\frac{1}{(z+\eta\underline{L}_{i'})^2}\exp\left(-\sum_{i'}\frac{1}{2(z+\eta\underline{L}_{i'})^2}\right)\mathrm{d}z\mathrm{d}\eta\right] \\
&\le 2(\eta_t-\eta_{t+1}) + 2\int_{\eta_{t+1}}^{\eta_t}\int_0^\infty w\mathrm{e}^{-w/2}\mathrm{d}w\mathrm{d}\eta \\
&= 10(\eta_t-\eta_{t+1}),
\end{aligned}$$

where we used the relation same as (21) in the last inequality. We obtain the lemma by taking its summation over $t \in [T]$. ■

Appendix D. Regret Bound for Stochastic Setting

In this appendix we prove Theorem 2 on the logarithmic regret bound of FTPL for the stochastic setting.

D.1. Regret Lower Bounds

In the regret analysis for the stochastic setting, we use the self-bounding technique (Zimmert and Seldin, 2021), which requires a regret lower bound of the policy. In our analysis, a lower bound in terms of w_t is not useful and instead we use the following bound.

Lemma 10 (i) *If* $\sum_{i \neq i^*} \frac{1}{(\eta_t \underline{\hat{L}}_{t,i})^2} \geq 1$ *then the instantaneous regret satisfies* $\sum_{i \neq i^*} \Delta_i w_{t,i} \geq 0.14\Delta$. (ii) *If* $\sum_{i \neq i^*} \frac{1}{(\eta_t \underline{\hat{L}}_{t,i})^2} \leq 1$ *then* $\sum_{i \neq i^*} \Delta_i w_{t,i} \geq 0.075 \sum_{i \neq i^*} \frac{\Delta_i}{(\eta_t \underline{\hat{L}}_{t,i})^2}$ *and* $w_{t,i^*} \geq 1/\mathrm{e}$.

Proof of Lemma 10 Let $\underline{\hat{L}}' = \min_{i \neq i^*} \underline{\hat{L}}_{t,i}$. Then, for any $a > 0$ we have

$$\begin{aligned}
\sum_{i \neq i^*} \Delta_i w_{t,i} &= 2 \int_0^\infty \left(\sum_{i \neq i^*} \frac{\Delta_i}{(z + \eta_t \underline{\hat{L}}_{t,i})^3} \right) \exp\left(- \sum_i \frac{1}{(z + \eta_t \underline{\hat{L}}_{t,i})^2} \right) \mathrm{d}z \\
&\geq 2 \int_{a\eta_t \underline{\hat{L}}'}^\infty \left(\sum_{i \neq i^*} \frac{\Delta_i}{(z + \eta_t \underline{\hat{L}}_{t,i})^3} \right) \exp\left(- \sum_i \frac{1}{(z + \eta_t \underline{\hat{L}}_{t,i})^2} \right) \mathrm{d}z.
\end{aligned}$$

(i) Consider the case $\sum_{i \neq i^*} \frac{1}{(\eta_t \underline{\hat{L}}_{t,i})^2} \geq 1$. It holds for any $z \geq a\eta_t \underline{\hat{L}}'$ that

$$\begin{aligned}
\sum_i \frac{1}{(z + \eta_t \underline{\hat{L}}_{t,i})^2} &\leq \sum_{i \neq i^*} \frac{1}{(z + \eta_t \underline{\hat{L}}_{t,i})^2} + \frac{1}{z^2} \\
&\leq \sum_{i \neq i^*} \frac{1}{(z + \eta_t \underline{\hat{L}}_{t,i})^2} + \frac{1}{(\frac{z + a\eta_t \underline{\hat{L}}'}{2})^2} \\
&\leq \sum_{i \neq i^*} \frac{1}{(z + \eta_t \underline{\hat{L}}_{t,i})^2} + 4 \sum_{i \neq i^*} \frac{1}{(z + a\eta_t \underline{\hat{L}}_{t,i})^2}.
\end{aligned}$$

Therefore, letting $a = 1$ we obtain

$$\begin{aligned}
\sum_{i \neq i^*} \Delta_i w_{t,i} &\geq 2\Delta \int_{\eta_t \underline{\hat{L}}'}^\infty \left(\sum_{i \neq i^*} \frac{1}{(z + \eta_t \underline{\hat{L}}_{t,i})^3} \right) \exp\left(-5 \sum_{i \neq i^*} \frac{1}{(z + \eta_t \underline{\hat{L}}_{t,i})^2} \right) \mathrm{d}z \\
&= \frac{\Delta}{5} \left(1 - \exp\left(-5 \sum_{i \neq i^*} \frac{1}{(\eta_t \underline{\hat{L}}' + \eta_t \underline{\hat{L}}_{t,i})^2} \right) \right)
\end{aligned}$$

$$\geq \frac{\Delta}{5}\left(1-\exp\left(-\sum_{i\neq i^*}\frac{5}{4(\eta_t\underline{\hat{L}}_{t,i})^2}\right)\right)$$
$$\geq \frac{\Delta}{5}\left(1-\mathrm{e}^{-5/4}\right) \geq 0.14\Delta.$$

(ii) When $\sum_{i\neq i^*}\frac{1}{(\eta_t\underline{\hat{L}}_{t,i})^2} \leq 1$ we have

$$\begin{aligned}
\sum_{i\neq i^*}\Delta_i w_{t,i} &\geq 2\int_{a\eta_t\underline{\hat{L}}'}^{\infty}\left(\sum_{i\neq i^*}\frac{\Delta_i}{(z+\eta_t\underline{\hat{L}}_{t,i})^3}\right)\exp\left(-\sum_i\frac{1}{(z+\eta_t\underline{\hat{L}}_{t,i})^2}\right)\mathrm{d}z \\
&\geq 2\int_{a\eta_t\underline{\hat{L}}'}^{\infty}\left(\sum_{i\neq i^*}\frac{\Delta_i}{(z+\eta_t\underline{\hat{L}}_{t,i})^3}\right)\exp\left(-\frac{1}{(a\eta_t\underline{\hat{L}}')^2}-\sum_{i\neq i^*}\frac{1}{(\eta_t\underline{\hat{L}}_{t,i})^2}\right)\mathrm{d}z \\
&\geq 2\int_{a\eta_t\underline{\hat{L}}'}^{\infty}\left(\sum_{i\neq i^*}\frac{\Delta_i}{(z+\eta_t\underline{\hat{L}}_{t,i})^3}\right)\exp\left(-\left(1+\frac{1}{a^2}\right)\sum_{i\neq i^*}\frac{1}{(\eta_t\underline{\hat{L}}_{t,i})^2}\right)\mathrm{d}z \\
&\geq 2\int_{a\eta_t\underline{\hat{L}}'}^{\infty}\sum_{i\neq i^*}\frac{\Delta_i}{(z+\eta_t\underline{\hat{L}}_{t,i})^3}\exp\left(-\left(1+\frac{1}{a^2}\right)\right)\mathrm{d}z \\
&= \sum_{i\neq i^*}\frac{\Delta_i}{(a\eta_t\underline{\hat{L}}'+\eta_t\underline{\hat{L}}_{t,i})^2}\exp\left(-\left(1+\frac{1}{a^2}\right)\right) \\
&\geq \sum_{i\neq i^*}\frac{\Delta_i}{((1+a)\eta_t\underline{\hat{L}}_{t,i})^2}\exp\left(-\left(1+\frac{1}{a^2}\right)\right). \qquad (23)
\end{aligned}$$

We obtain the desired bound by letting $a = 1.3$. Note that $\sum_{i\neq i^*}\frac{1}{(\eta_t\underline{\hat{L}}_{t,i})^2} \leq 1$ implies $\underline{\hat{L}}_{t,i^*} = 0$. Therefore in this case we also have

$$\begin{aligned}
w_{t,i^*} &= 2\int_0^{\infty}\frac{1}{z^3}\exp\left(-\sum_i\frac{1}{(z+\eta_t\underline{\hat{L}}_{t,i})^2}\right)\mathrm{d}z \\
&\geq 2\int_0^{\infty}\frac{1}{z^3}\exp\left(-\frac{1}{z^2}-\sum_{i\neq i^*}\frac{1}{(\eta_t\underline{\hat{L}}_{t,i})^2}\right)\mathrm{d}z \\
&\geq 2\mathrm{e}^{-1}\int_0^{\infty}\frac{1}{z^3}\exp\left(-\frac{1}{z^2}\right)\mathrm{d}z = 1/\mathrm{e}, \qquad (24)
\end{aligned}$$

which is the desired result. ■

D.2. Regret for Optimal Arm

When we apply the self-bounding technique, we need to express regret arising from the optimal arm in terms of statstics of the other arms. For this purpose we use the following lemma.

Lemma 11 *Assume that $\hat{L}_t$ satisfies $\sum_{i\neq i^*} \frac{1}{(\eta_t \underline{\hat{L}}_{t,i})^2} \leq 1$. Then, for any $\alpha \in (0,1)$,*

$$\mathbb{E}\left[\hat{\ell}_{t,i^*}\left(\phi_{i^*}(\eta_t \hat{L}_t) - \phi_{i^*}(\eta_t(\hat{L}_t + \hat{\ell}_t))\right)\middle| \hat{L}_t\right] \leq \frac{4\mathrm{e}}{(1-\alpha)^3}\sum_{i\neq i^*}\frac{1}{\underline{\hat{L}}_{t,i}} + \frac{(1-\mathrm{e}^{-1})^{-\alpha/\eta_t}\,(\alpha/\eta_t + \mathrm{e})}{1-\mathrm{e}^{-1}}.$$

Proof $\sum_{i\neq i^*} \frac{1}{(\eta_t \underline{\hat{L}}_{t,i})^2} \leq 1$ implies that any $i \neq i^*$ satisfies

$$\underline{\hat{L}}_{t,i} \geq 1/\eta_t, \tag{25}$$

which also implies that i^* is the unique minimizer of $\{\hat{L}_{t,j}\}$. We separately consider the cases $\widehat{w_{t,i^*}^{-1}} \leq \alpha/\eta_t$ and $\widehat{w_{t,i^*}^{-1}} > \alpha/\eta_t$.

Let us consider the former case. In this case, we also have $\hat{\ell}_{t,i^*} \leq \alpha/\eta_t$. On the other hand, for any $x \leq \alpha/\eta_t$ and $i \neq i^*$ we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}x}\phi_i(\eta_t(\hat{L}_t + e_{i^*}x)) \\
&= 4\eta_t \int_0^\infty \frac{1}{z^3}\frac{1}{(z+\eta_t(\underline{\hat{L}}_{t,i} - x))^3}\exp\left(-\frac{1}{z^2} - \sum_{i'\neq i^*}\frac{1}{(z+\eta_t(\underline{\hat{L}}_{t,i'} - x))^2}\right)\mathrm{d}z \\
&\leq \frac{4\eta_t}{(1-\alpha)^3}\int_0^\infty \frac{1}{z^3}\frac{1}{(\eta_t \underline{\hat{L}}_{t,i})^3}\exp\left(-\frac{1}{z^2}\right)\mathrm{d}z \\
&= \frac{2\eta_t}{(1-\alpha)^3(\eta_t \underline{\hat{L}}_{t,i})^3} \\
&\leq \frac{2}{(1-\alpha)^3 \underline{\hat{L}}_{t,i}} \quad \text{(by (25))}\,.
\end{aligned}$$

Combining this with the fact that $\sum_i \phi_i(\lambda) = 1$ holds for any λ, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left[\hat{\ell}_{t,i^*} \leq \alpha/\eta_t\right]\hat{\ell}_{t,i^*}\left(\phi_{i^*}(\eta_t \hat{L}_t) - \phi_{i^*}(\eta_t(\hat{L}_t + \hat{\ell}_t))\right)\middle|\hat{L}_t\right] \\
&= \mathbb{E}\left[\mathbb{1}\left[\hat{\ell}_{t,i^*} \leq \alpha/\eta_t\right]\hat{\ell}_{t,i^*}\sum_{i\neq i^*}\left(\phi_i(\eta_t(\hat{L}_t + \hat{\ell}_t)) - \phi_i(\eta_t \hat{L}_t)\right)\middle|\hat{L}_t\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\left[\hat{\ell}_{t,i^*} \leq \alpha/\eta_t\right]\hat{\ell}_{t,i^*}^2\sum_{i\neq i^*}\frac{2}{(1-\alpha)^3\underline{\hat{L}}_{t,i}}\middle|\hat{L}_t\right] \\
&\leq \mathbb{E}\left[\hat{\ell}_{t,i^*}^2\sum_{i\neq i^*}\frac{2}{(1-\alpha)^3\underline{\hat{L}}_{t,i}}\middle|\hat{L}_t\right] \\
&\leq \mathbb{E}\left[\frac{2\ell_{t,i^*}^2}{w_{t,i^*}}\sum_{i\neq i^*}\frac{2}{(1-\alpha)^3\underline{\hat{L}}_{t,i}}\middle|\hat{L}_t\right] \quad \text{(by (13))} \\
&\leq 4\mathrm{e}\sum_{i\neq i^*}\frac{1}{(1-\alpha)^3\underline{\hat{L}}_{t,i}},
\end{aligned} \tag{26}$$

where the last inequality follows from Lemma 10.

On the other hand we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left[\hat{\ell}_{t,i^*} > \alpha/\eta_t\right]\hat{\ell}_{t,i^*}\left(\phi_{i^*}(\eta_t\hat{L}_t) - \phi_{i^*}(\eta_t(\hat{L}_t+\hat{\ell}_t))\right)\middle|\hat{L}_t\right] \\
&\le \mathbb{E}\left[\mathbb{1}\left[\hat{\ell}_{t,i^*} > \alpha/\eta_t\right]\hat{\ell}_{t,i^*}\middle|\hat{L}_t\right] \\
&= \mathbb{E}\left[\mathbb{1}\left[I_t = i^*,\, \hat{\ell}_{t,i^*} > \alpha/\eta_t\right]\hat{\ell}_{t,i^*}\middle|\hat{L}_t\right] \\
&\le \mathbb{E}\left[\mathbb{1}\left[I_t = i^*,\, \widehat{w^{-1}_{t,i^*}} > \alpha/\eta_t\right]\widehat{w^{-1}_{t,i^*}}\middle|\hat{L}_t\right] \quad \left(\text{by } \hat{\ell}_{t,i^*} = \ell_{t,i}\widehat{w^{-1}_{t,i^*}} \le \widehat{w^{-1}_{t,i^*}} \text{ when } I_t = i^*\right) \\
&= \mathbb{E}\left[w_{t,i^*}\sum_{m=\lfloor\alpha/\eta_t\rfloor+1}^{\infty} m(1-w_{t,i^*})^{m-1}\right] \\
&= \mathbb{E}\left[(1-w_{t,i^*})^{\lfloor\alpha/\eta_t\rfloor}\left(\lfloor\alpha/\eta_t\rfloor + \frac{1}{w_{t,i^*}}\right)\right] \\
&\le (1-\mathrm{e}^{-1})^{\lfloor\alpha/\eta_t\rfloor}\left(\lfloor\alpha/\eta_t\rfloor + \mathrm{e}\right) \\
&\le \frac{1}{1-\mathrm{e}^{-1}}(1-\mathrm{e}^{-1})^{\alpha/\eta_t}\left(\alpha/\eta_t + \mathrm{e}\right). \qquad (27)
\end{aligned}
$$

We obtain the lemma by combining (26) and (27). ■

D.3. Proof of Theorem 2

Define event $A_t = \{\sum_{i\neq i^*}\frac{1}{(\eta_t\hat{L}_{t,i})^2} \le 1\}$. By putting the results so far the regret is bounded from above by

$$
\begin{aligned}
&\mathrm{Regret}(T) \\
&\le \sum_{t=1}^{T}\mathbb{E}\left[\left\langle\hat{\ell}_t, \phi(\eta_t\hat{L}_t) - \phi(\eta_t(\hat{L}_t+\hat{\ell}_t))\right\rangle\right] + \sum_{t=1}^{T}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\mathbb{E}\left[r_{t+1,I_{t+1}} - r_{t+1,i^*}\right] + C_1 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by Lemmas 3 and 4)} \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\left\langle\hat{\ell}_t, \phi(\eta_t\hat{L}_t) - \phi(\eta_t(\hat{L}_t+\hat{\ell}_t))\right\rangle + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\left(r_{t+1,I_{t+1}} - r_{t+1,i^*}\right)\middle|\hat{L}_t\right]\right] + C_1 \\
&\le \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\left\langle\hat{\ell}_t, \phi(\eta_t\hat{L}_t) - \phi(\eta_t(\hat{L}_t+\hat{\ell}_t))\right\rangle + \frac{r_{t+1,I_{t+1}} - r_{t+1,i^*}}{2c\sqrt{t}}\middle|\hat{L}_t\right]\right] + C_1 \qquad (28)
\end{aligned}
$$

for $C_1 = 10\eta_1 + \frac{\sqrt{\pi K}}{\eta_1}$, where the last equality follows from

$$
\begin{aligned}
\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} &= \frac{1}{c}(\sqrt{t+1} - \sqrt{t}) \\
&= \frac{\sqrt{t}}{c}(\sqrt{1+1/t} - 1). \\
&\le \frac{1}{2c\sqrt{t}}.
\end{aligned}
$$

If $\hat{L}_t$ satisfies A_t then the inner expectation is bounded by

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \hat{\ell}_t, \phi(\eta_t \hat{L}_t) - \phi(\eta_t(\hat{L}_t + \hat{\ell}_t))\right\rangle + \frac{r_{t+1,I_{t+1}} - r_{t+1,i^*}}{2c\sqrt{t}} \middle| \hat{L}_t\right] \\
&\leq \sum_{i \neq i^*} \left(\frac{4}{\underline{\hat{L}}_{t,i}} + \frac{4\mathrm{e}}{(1-\alpha)^3 \underline{\hat{L}}_{t,i}} + \frac{1}{c^2 \underline{\hat{L}}_{t,i}} \right) + C_{2,t} \quad \text{(by Lemmas 7, 9 and 11)} \\
&= \sum_{i \neq i^*} \frac{25 + c^{-2}}{\underline{\hat{L}}_{t,i}} + C_{2,t}, \qquad (29)
\end{aligned}$$

where $C_{2,t} = \frac{(1-\mathrm{e}^{-1})^{\alpha/\eta_t}(\alpha/\eta_t + \mathrm{e})}{1-\mathrm{e}^{-1}}$ and we chose[1] $\alpha = 1 - (4\mathrm{e}/21)^{1/3}$.

On the other hand, if $\hat{L}_t$ does not satisfy A_t then it is bounded by

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \hat{\ell}_t, \phi(\eta_t \hat{L}_t) - \phi(\eta_t(\hat{L}_t + \hat{\ell}_t))\right\rangle + \frac{r_{t+1,I_{t+1}} - r_{t+1,i^*}}{2c\sqrt{t}} \middle| \hat{L}_t\right] \\
&\leq 6c\sqrt{\frac{\pi K}{t}} + \frac{3.7}{2c}\sqrt{\frac{K}{t}} \quad \text{(by Lemmas 8 and 9)} \\
&\leq (11c + 2/c)\sqrt{\frac{K}{t}}. \qquad (30)
\end{aligned}$$

Combining (29) and (30) with (28) we obtain

$$\mathrm{Regret}(T) \leq \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}[A_t] \sum_{i \neq i^*} \frac{25 + c^{-2}}{\hat{L}_{t,i}} + \mathbb{1}[A_t^c]\,(11c + 2/c)\sqrt{K/t} \right] + C_1 + C_2. \qquad (31)$$

Here we defined $C_2 = \sum_{t=1}^{\infty} C_{2,t}$, which satisfies

$$\begin{aligned}
C_2 &= \sum_{t=1}^{\infty} \frac{(1-\mathrm{e}^{-1})^{\alpha/\eta_t}\,(\alpha/\eta_t + \mathrm{e})}{1-\mathrm{e}^{-1}} \\
&= \sum_{t=1}^{\infty} \frac{\mathrm{e}^{-\rho\alpha\sqrt{t}/c}\left(\alpha\sqrt{t}/c + \mathrm{e}\right)}{1-\mathrm{e}^{-1}} \quad \left(\text{by letting } \rho = -\log(1-\mathrm{e}^{-1}) > 0\right) \\
&\leq \int_0^{\infty} \frac{\mathrm{e}^{-\rho\alpha\sqrt{t}/c}\left(\alpha(\sqrt{t}+1)/c + \mathrm{e}\right)}{1-\mathrm{e}^{-1}} \mathrm{d}t \\
&\leq \int_0^{\infty} \frac{\mathrm{e}^{-u}\,(u/\rho + \alpha/c + \mathrm{e})}{1-\mathrm{e}^{-1}} \frac{2c^2 u}{\rho^2\alpha^2} \mathrm{d}u \quad \left(\text{by letting } u = \rho\alpha\sqrt{t}/c\right) \\
&= \frac{2/\rho + \alpha/c + \mathrm{e}}{1-\mathrm{e}^{-1}} \frac{2c^2}{\rho^2\alpha^2} \\
&\leq 2743c^2 + 77c.
\end{aligned}$$

On the other hand, by the lower bound in Lemma 10 we have

$$\mathrm{Regret}(T) \geq \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}[A_t]\, 0.075 \sum_{i \neq i^*} \frac{t\Delta_i}{c^2 \underline{\hat{L}}_{t,i}^2} + \mathbb{1}[A_t^c]\, 0.14\Delta \right]. \qquad (32)$$

1. This choice is just for simpler main term and not essential.

By considering (31) $-$ (32)$/2$ we have

$$
\begin{aligned}
&\frac{\mathrm{Regret}(T)}{2} \\
&\le \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left[A_t\right] \sum_{i \neq i^*} \left(\frac{15 + c^{-2}}{\hat{L}_{t,i}} - \frac{0.075 t \Delta_i}{2c^2 \hat{L}_{t,i}^2} \right) + \mathbb{1}\left[A_t^c\right] \left((11c + 2/c)\sqrt{K/t} - 0.07\Delta \right) \right] + C_1 + C_2 \\
&\le \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left[A_t\right] \sum_{i \neq i^*} \frac{(15c + c^{-1})^2}{0.15 t \Delta_i} + \mathbb{1}\left[A_t^c\right] \left((11c + 2/c)\sqrt{K/t} - 0.07\Delta \right) \right] + C_1 + C_2 \\
&\qquad\qquad \left(\text{by } ax - bx^2 \le a^2/4b \text{ for } b > 0\right) \\
&\le \sum_{t=1}^{T} \sum_{i \neq i^*} \frac{(15c + c^{-1})^2}{0.15 t \Delta_i} + \sum_{t=1}^{T} \max\left\{ (11c + 2/c)\sqrt{K/t} - 0.07\Delta, 0 \right\} + C_1 + C_2 \\
&\le \sum_{i \neq i^*} \frac{(25c + c^{-1})^2 (1 + \log T)}{0.15 \Delta_i} + \frac{(11c + 2/c)^2 K}{0.07\Delta} + C_1 + C_2 \\
&\le \sum_{i \neq i^*} \frac{(25c + c^{-1})^2 \log T}{0.15 \Delta_i} + \frac{(25c + c^{-1})^2 K}{0.15\Delta} + \frac{(11c + 2/c)^2 K}{0.07\Delta} + \frac{(2743c^2 + 87c + \frac{\sqrt{\pi/2}}{c})K}{2\Delta} \qquad (33) \\
&\le \sum_{i \neq i^*} \frac{(25c + c^{-1})^2 \log T}{0.15 \Delta_i} + \frac{(121c + 12/c)^2 K}{2\Delta},
\end{aligned}
$$

where (33) follows from $K \ge 2$ and $\Delta \in (0, 1]$, and the last inequality can be confirmed by comparison of the coefficients of $c^2, c, 1, c^{-1}, c^{-2}$. ■

Remark 12 In this analysis, we separately considered cases $\sum_{i \neq i^*} \frac{1}{(\eta_t \hat{L}_{t,i})^2} \le h$ and $\sum_{i \neq i^*} \frac{1}{(\eta_t \hat{L}_{t,i})^2} > h$ for $h = 1$. Here choice of the threshold h is not important and we can run the same argument under any choice of $h > 0$. In particular, if we take $h > 0$ arbitrarily close to 0 then the coefficient 0.075 in Lemma 10 approaches 1 and the term corresponding to Lemma 11 becomes negligible. As a result, we can see that the resulting regret bound approaches

$$
\mathrm{Regret}(T) \le \sum_{i \neq i^*} \frac{(4c + c^{-1})^2 \log T}{\Delta_i} + o(\log T)
$$

with diverging $o(\log T)$ term, which is optimized as $\sum_{i \neq i^*} \frac{16 \log T}{\Delta_i} + o(\log T)$ for $c = 1/2$.

Similarly, if we consider the version where $w_{t,i}$ is exactly computed instead of geometric resampling then the bound becomes

$$
\mathrm{Regret}(T) \le \sum_{i \neq i^*} \frac{(2c + c^{-1})^2 \log T}{\Delta_i} + o(\log T),
$$

which is optimized as $\sum_{i \neq i^*} \frac{8 \log T}{\Delta_i} + o(\log T)$ for $c = \sqrt{2}/2$.

Appendix E. Extension to Intermediate Settings

Several models have been proposed as intermediates between the stochastic and adversarial settings. Most of them are expressed as the *adversarial setting with a self-bounding constraint*, which is formulated as follows.

A setting is said to satisfy (Δ, C, T) self-bounding constraint for $\Delta \in [0,1]^K$ and $C \geq 0$ with time horizon T if the regret satisfies

$$\mathrm{Regret}(T) \geq \sum_{t=1}^{T} \sum_{i} \Delta_i \mathbb{P}[I_t = i] - C.$$

For example, the corrupted stochastic setting (Lykouris et al., 2018; Ito, 2021b; Jin et al., 2021; Ito, 2021a; Ito et al., 2022; Erez and Koren, 2021) and the stochastically constrained adversarial setting (Zimmert and Seldin, 2021) as well as the stochastic setting ($C = 0$) are expressed within this formulation. For this setting, the self-bounding technique can be applied in the following way.

Assume that the setting satisfies (Δ, C, T) self-bounding constraint for Δ such that $\Delta_i = 0$ holds for unique i, denoted by i^*. Let $\Delta = \min_{i \neq i^*} \Delta_i > 0$ and $A_t = \{\sum_{i \neq i^*} \frac{1}{(\eta_t \hat{L}_{t,i})^2} \leq 1\}$. As explained in Section 6, the regret is bounded by

$$\mathrm{Regret}(T) \leq O\left(\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}\left[A_t\right]\sum_{i \neq i^*}\frac{1}{\underline{\hat{L}}_{t,i}} + \mathbb{1}\left[A_t^c\right]\sqrt{K/t}\right)\right]\right).$$

Here, it is implicitly proved in (23) that under A_t we have

$$w_{t,i} \geq \frac{1}{((1+a)\eta_t \underline{\hat{L}}_{t,i})^2} \exp\left(-\left(1 + \frac{1}{a^2}\right)\right) \geq \frac{0.075}{(\eta_t \underline{\hat{L}}_{t,i})^2}$$

for any $i \neq i^*$, where we set $a = 1.3$. Therefore, we have

$$\begin{aligned}
\mathrm{Regret}(T) &\leq O\left(\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}\left[A_t\right]\sum_{i \neq i^*}\frac{\eta_t\sqrt{w_{t,i}}}{\sqrt{0.075}} + \mathbb{1}\left[A_t^c\right]\sqrt{K/t}\right)\right]\right) \\
&= O\left(\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}\left[A_t\right]\sum_{i \neq i^*}\sqrt{w_{t,i}/t} + \mathbb{1}\left[A_t^c\right]\sqrt{K/t}\right)\right]\right). \qquad (34)
\end{aligned}$$

On the other hand, since $\mathbb{P}[I_t = i] = \mathbb{E}\left[w_{t,i}\right]$ we have

$$\begin{aligned}
\mathrm{Regret}(T) &\geq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i \neq i^*}\Delta_i w_{t,i}\right] - C \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}\left[A_t\right]\sum_{i \neq i^*}\Delta_i w_{t,i} + \mathbb{1}\left[A_t^c\right]\sum_{i \neq i^*}\Delta_i w_{t,i}\right)\right] - C \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}\left[A_t\right]\sum_{i \neq i^*}\Delta_i w_{t,i} + \mathbb{1}\left[A_t^c\right]\Delta(1 - w_{t,i^*})\right)\right] - C.
\end{aligned}$$

Here, by following the same discussion as (24), we can show that $w_{t,i^*} \le \mathrm{e}^{-1}$ under A_t^c. Therefore we have

$$\mathrm{Regret}(T) \ge \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}\left[A_t\right]\sum_{i\neq i^*}\Delta_i w_{t,i} + \mathbb{1}\left[A_t^c\right](1-\mathrm{e}^{-1})\Delta\right)\right] - C. \tag{35}$$

By considering (34) $-\ \alpha \times$ (35) for $\alpha \in (0,1)$ we have

$$\begin{aligned}&(1-\alpha)\mathrm{Regret}(T)\\ &= O\left(\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}\left[A_t\right]\sum_{i\neq i^*}\left(\sqrt{w_{t,i}/t} - \alpha\Delta_i w_{t,i}\right) + \mathbb{1}\left[A_t^c\right]\left(\sqrt{K/t} - \alpha(1-\mathrm{e}^{-1})\Delta\right)\right)\right]\right) + \alpha C.\end{aligned}$$

and taking the worst case of A_t and $w_{t,i}$ we obtain

$$(1-\alpha)\mathrm{Regret}(T) = O\left(\sum_{i\neq i^*}\frac{\log T}{\alpha\Delta_i} + \frac{K}{\alpha\Delta}\right) + \alpha C.$$

By optimizing $\alpha \in (0,1)$ we obtain

$$\mathrm{Regret}(T) = O\left(\left(\sum_{i\neq i^*}\frac{\log T}{\Delta_i} + \frac{K}{\Delta}\right) + \sqrt{C\left(\sum_{i\neq i^*}\frac{\log T}{\Delta_i} + \frac{K}{\Delta}\right)}\right).$$

Appendix F. Miscellaneous Calculation

In this appendix we cover omitted elementary calculation for the results of this paper.

F.1. Derivation of (7)

When $\lambda_i \neq \min_j \lambda_j$, it holds from (1) that

$$\phi_i'(\lambda) = 2\int_{-\min_{j\in[K]}\lambda_j}^{\infty}\left(\frac{2}{(z+\lambda_i)^6} - \frac{3}{(z+\lambda_i)^4}\right)\exp\left(-\sum_{i'}\frac{1}{(z+\lambda_i)^2}\right)\mathrm{d}z \tag{36}$$

and (7) immediately follows.

When λ_i is the unique minimizer of $\{\lambda_j\}$, we have

$$\begin{aligned}\phi_i'(\lambda) &= 2\int_{-\min_{j\in[K]}\lambda_j}^{\infty}\left(\frac{2}{(z+\lambda_i)^6} - \frac{3}{(z+\lambda_i)^4}\right)\exp\left(-\sum_{i'}\frac{1}{(z+\lambda_i)^2}\right)\mathrm{d}z\\ &\quad + 2\lim_{z\to-\min_{j\in[K]}\lambda_j}\frac{1}{(z+\lambda_i)^3}\exp\left(-\sum_{i'}\frac{1}{(z+\lambda_i)^2}\right)\\ &= 2\int_{-\min_{j\in[K]}\lambda_j}^{\infty}\left(\frac{2}{(z+\lambda_i)^6} - \frac{3}{(z+\lambda_i)^4}\right)\exp\left(-\sum_{i'}\frac{1}{(z+\lambda_i)^2}\right)\mathrm{d}z,\end{aligned} \tag{37}$$

which again recovers (7).

When λ_i is a non-unique minimizer of $\{\lambda_j\}$, the right and left derivatives of $\phi_i(\lambda)$ are expressed by (36) and (37), respectively.

F.2. Ratio of Incomplete Gamma Functions

In this appendix we show that it holds for $x > 0$ that

$$\frac{\gamma(3/2, x)}{\gamma(1, x)} \leq \frac{2\sqrt{x}}{3} \wedge \frac{\sqrt{\pi}}{2},$$

which is, by $\gamma(1, x) = 1 - \mathrm{e}^{-x}$, also expressed as

$$\gamma(3/2, x) - 2(1 - \mathrm{e}^{-x})\sqrt{x}/3 \leq 0, \tag{38}$$
$$\gamma(3/2, x) - \sqrt{\pi}(1 - \mathrm{e}^{-x})/2 \leq 0. \tag{39}$$

The derivative of the LHS of (38) is expressed as

$$\mathrm{e}^{-x}\sqrt{x} - \frac{2\mathrm{e}^{-x}\sqrt{x}}{3} - \frac{1 - \mathrm{e}^{-x}}{3\sqrt{x}} = \frac{1}{3\sqrt{x}}\left(x\mathrm{e}^{-x} + \mathrm{e}^{-x} - 1\right) \leq 0.$$

Therefore the LHS of (38) is decreasing and we obtain (38) since (38) holds with equality at $x = 0$.

Similarly, the derivative of the LHS of (39) is expressed as

$$\mathrm{e}^{-x}\sqrt{x} - \sqrt{\pi}\mathrm{e}^{-x}/2 = \mathrm{e}^{-x}(\sqrt{x} - \sqrt{\pi}/2),$$

which means that the LHS of (39) is maximized at $x = 0$ or $x \to \infty$. We obtain (39) since it holds with equality at these x's.