Optimism in the Face of Ambiguity Principle for Multi-Armed Bandits

Mengmeng Li* Daniel Kuhn† Bahar Taşkesen‡

October 7, 2024

Abstract

Follow-The-Regularized-Leader (FTRL) algorithms often enjoy optimal regret for adversarial as well as stochastic bandit problems and allow for a streamlined analysis. Nonetheless, FTRL algorithms require the solution of an optimization problem in every iteration and are thus computationally challenging. In contrast, Follow-The-Perturbed-Leader (FTPL) algorithms achieve computational efficiency by perturbing the estimates of the rewards of the arms, but their regret analysis is cumbersome. We propose a new FTPL algorithm that generates optimal policies for both adversarial and stochastic multi-armed bandits. Like FTRL, our algorithm admits a unified regret analysis, and similar to FTPL, it offers low computational costs. Unlike existing FTPL algorithms that rely on independent additive disturbances governed by a *known* distribution, we allow for disturbances governed by an *ambiguous* distribution that is only known to belong to a given set and propose a principle of optimism in the face of ambiguity. Consequently, our framework generalizes existing FTPL algorithms. It also encapsulates a broad range of FTRL methods as special cases, including several optimal ones, which appears to be impossible with current FTPL methods. Finally, we use techniques from discrete choice theory to devise an efficient bisection algorithm for computing the optimistic arm sampling probabilities. This algorithm is up to 10^4 times faster than standard FTRL algorithms that solve an optimization problem in every iteration. Our results not only settle existing conjectures but also provide new insights into the impact of perturbations by mapping FTRL to FTPL.

1 Introduction

We consider multi-armed bandit problems where a learner sequentially interacts with an environment for T rounds. In each round, the learner selects one of the K arms, observes and receives its reward. The goal of the learner is to minimize regret, which measures the absolute difference between the total reward received and the total reward that could have been received with perfect knowledge of the reward distribution. When the rewards received in each round are drawn independently from a fixed unknown reward distribution, the upper confidence bound (UCB) algorithm [Auer et al., 2002a] or the Thompson sampling algorithm [Thompson, 1933] achieve optimal regret $\mathcal{O}(\log T)$ [Bubeck and Cesa-Bianchi, 2012]. However, in an adversarial setting, where rewards are chosen strategically by an adversary, these methods suffer from linear regret [Zimmert

*Risk Analytics and Optimization Chair, EPFL mengmeng.li@epfl.ch
†Risk Analytics and Optimization Chair, EPFL daniel.kuhn@epfl.ch
‡Booth School of Business, University of Chicago bahar.taskesen@chicagobooth.edu

and Seldin, 2021]. Conversely, in an adversarial environment, the Follow-the-Regularized-Leader (FTRL) algorithm, introduced by Gordon [1999] as "generalized gradient descent", often achieves optimal regret $\mathcal{O}(\sqrt{KT})$ [Bubeck and Cesa-Bianchi, 2012].

Prior knowledge regarding the nature of the environment is typically unavailable. Therefore, an algorithm that simultaneously achieves optimal regret in both stochastic and adversarial settings is highly desirable. Recently, Zimmert and Seldin [2021] proved that an FTRL algorithm with Tsallis entropy regularizer can *simultaneously* achieve the optimal regret bound $\mathcal{O}(\sqrt{KT})$ in adversarial settings and the optimal regret bound $\mathcal{O}(\log T)$ in stochastic settings, without the need for parameter tuning. Algorithms of this kind are often said to display the "best-of-both-worlds" (BOBW) capability [Bubeck and Slivkins, 2012]. The results in [Zimmert and Seldin, 2021] have been extended along various directions, aimed at identifying the abstract properties of the regularizers that lead to the BOBW capability [Jin et al., 2024]. Nevertheless, FTRL algorithms require the solution of an expensive optimization problem in each round to compute the arm-sampling distribution. On the other hand, the Follow-the-Perturbed-Leader algorithms [Hannan, 1957] perturb the cumulative reward estimates with noise sampled from a given distribution and select the arm with the maximum perturbed reward estimate. They are popular for their superior computational efficiency [Abernethy et al., 2014; Lattimore and Szepesvári, 2020] compared to the FTRL algorithms. Recently, it was shown that FTPL with Fréchet perturbations has BOBW capability [Honda et al., 2023]. However, their analysis heavily relies on the specific form of the Fréchet distribution with a particular shape. A generalized analysis of the FTPL algorithms was later provided by Lee et al. [2024], further showcasing the strength of FTPL approaches. Nevertheless, their analysis still relies significantly on extreme value theory and shares no significant commonality with the FTRL-based analysis.

Note that any FTPL policy can be expressed as an FTRL policy [Abernethy et al., 2016; Hofbauer and Sandholm, 2002]. However, the reverse direction does not hold in general [Hofbauer and Sandholm, 2002, Proposition 2.2]. Establishing a one-to-one correspondence between meaningful subclasses of FTPL and FTRL policies remains open [Abernethy et al., 2016]. The Gradient-Based Prediction Algorithm (GBPA) framework [Abernethy et al., 2015] encompasses FTRL and FTPL as special cases, but whether the algorithm itself is FTRL or FTPL still demands specialized regret analysis. Moreover, an open question is posed by Kim and Tewari [2019] on whether there exists a noise distribution that matches the FTRL policy with the Tsallis entropy regularizer. Kim and Tewari [2019] even showed the impossibility of recovering Tsallis-entropy-regularized FTRL using FTPL with independent noise distributions across the arms. Constructing perturbations that exactly match the FTRL algorithm with BOBW capability is crucial to understanding the effects of regularization through perturbation. Answering this open question will also lead to the unification of FTRL and FTPL regret analysis.

In this paper, we bridge this gap by studying *ambiguous* noise-sampling distributions that allow for *correlation* across the arms. In addition, we introduce an "optimism in the face of ambiguity" principle, whereby arms are selected under the most advantageous noise distribution. This is in stark contrast to standard FTPL algorithms, which assume that the noise distribution is fixed. Using techniques from discrete choice theory, which is traditionally studied in economics and psychology, we show that the arm-sampling probabilities under the best noise distribution can be computed highly efficiently using bisection. Unlike standard FTRL algorithms that require the solution of an expensive optimization problem in every round, our approach is far more computationally efficient. As a result, our algorithm admits a unified regret analysis similar to FTRL and a computationally efficient implementation similar to FTPL. It also encompasses a broad range of FTRL methods as special cases, including several optimal ones, such as those with Tsallis entropy and hybrid regularizers. Notably, while it previously appeared impossible to unify these FTRL methods within the traditional FTPL framework, our approach successfully achieves this integration.

Related work. We relax the i.i.d. noise assumption in our paper, generalizing the traditional FTPL methods. The i.i.d. noise assumption underlying most FTPL algorithms is also relaxed in [Melo and Müller, 2023] by interpreting the arm-sampling probabilities as choice probabilities of a nested logit model commonly studied in discrete choice theory. However, the noise distribution considered in [Melo and Müller, 2023] must be a generalized extreme-value distribution, and the resulting algorithm does not have BOBW capability. In contrast, we work with a family of distributions and use ideas from discrete choice theory to devise a highly efficient bisection algorithm for computing the arm-sampling probabilities under the most advantageous noise distribution. This general framework encompasses several algorithms that enjoy BOBW regret bounds. The FTPL algorithm with i.i.d. Fréchet-distributed noise is also known to have the BOBW capability [Honda et al., 2023], but their regret analysis is tailored to the Fréchet distribution. The results in [Honda et al., 2023] are generalized to other noise distributions by Lee et al. [2024], but the regret analysis remains cumbersome. Meanwhile, our perturbation-based algorithm achieves BOBW regret bounds by exploiting the exact equivalence with FTRL algorithms that have the BOBW capability.

Notation. We denote by $[K] = \{1, \dots, K\}$ the set of all integers up to $K \in \mathbb{N}$. The probability simplex over $[K]$ is defined as $\Delta^K = \{\boldsymbol{p} \in \mathbb{R}_+^K : \sum_{k=1}^K p_k = 1\}$. We use $\boldsymbol{e}_i$ with $i \in [d]$ to denote the i^{th} standard basis vector of the d-dimensional Euclidean space. The Bregman divergence function induced by a differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$ can be expressed as $\mathbb{D}_\phi(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \boldsymbol{x} - \boldsymbol{y}, \nabla\phi(\boldsymbol{y})\rangle$.

2 Multi-Armed Bandits

We study the multi-armed bandit (MAB) problem, where a learner is given a fixed set of arms $[K]$ and interacts with an environment over $T \in \mathbb{N}$ rounds. In each round $t \in [T]$, the learner selects an action $a_t \in [K]$, and the environment generates a reward vector $\boldsymbol{r}_t = (r_{t,1}, r_{t,2}, \dots, r_{t,K}) \in [-1, 0]^K$. The learner observes and receives the reward associated with the chosen arm, r_{t,a_t}, and receives no information as to the values $r_{t,k}$ for $k \neq a_t$. The learner's objective is to minimize its (pseudo) regret, which measures the difference between the expected reward of the best arm in hindsight and the learner's expected cumulative reward

$$\mathcal{R}(T) = \max_{k \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} r_{t,k}\right] - \mathbb{E}\left[\sum_{t=1}^{T} r_{t,a_t}\right],$$

where the expectation is taken over the internal randomization of the algorithm and the stochastic nature of the environment. The nature of the learning experience depends on the reward generation paradigms employed by the environment. In the adversarial paradigm, the environment can choose the reward vectors arbitrarily, potentially using the history of the learner's actions to influence future rewards. In the stochastic paradigm, the rewards are sampled i.i.d. from a fixed distribution. There are also other regimes characterized by the varying levels of adversarial power exerted by the environment.

For clarity in presenting our results and to capture various learning paradigms, we adopt adversarial regime with a self-bounding constraint from [Zimmert and Seldin, 2021]. In this regime, for some $\Delta \in [0,1]^K$ and $C \geq 0$, the adversary selects rewards such that at time T, the learner's regret of any policy satisfies

$$\mathcal{R}(T) \geq \sum_{t=1}^{T} \sum_{k=1}^{K} \Delta_k \mathbb{P}(a_t = k) - C. \tag{1}$$

Note that the above condition must be satisfied at time T, but it is not required to hold for all times $t < T$. The stochastic bandit setting, where the rewards r_{t,a_t} are drawn i.i.d. from distributions with fixed means $\mathbb{E}[r_{t,a_t} | a_t = k] = \mu_k$, is an instance of the adversarial regime with a self-bounding constraint. The pseudo-regret in the stochastic regime can be written as

$$\mathcal{R}(T) = \sum_{t=1}^{T} \sum_{k=1}^{K} \left(\max_{j \in [K]} \mathbb{E}[r_{t,j}] - \mathbb{E}[r_{t,a_t} \mid a_t = k] \right) \mathbb{P}(a_t = k),$$

which satisfies (1) with $\Delta_k = \max_{j \in [K]} \mathbb{E}[r_{t,j}] - \mathbb{E}[r_{t,k}]$ for all $k \in [K]$ and $C = 0$. The adversarial regime with a self-bounding constraint also encompasses several other paradigms, including the stochastically constrained adversarial [Wei and Luo, 2018] and adversarially corrupted stochastic [Lykouris et al., 2018] settings. Finally, if the learner's regret for any policy is not required to satisfy the self-bounding constraint (1), then the learner operates within the adversarial setting.

Algorithm 1 Gradient-based prediction algorithm (GBPA) for MAB

Require: Differentiable and convex function ϕ with $\nabla_{\boldsymbol{u}} \phi(\boldsymbol{u}) \in \Delta^K$

```
û_{0,k} ← 0   ∀k ∈ [K]
for round t = 1, ..., T do
    Environment chooses a reward vector r_t ∈ [−1, 0]^K
    Learner chooses a_t ∼ p_t = ∇_u φ(u)|_{u = û_{t−1}}
    Learner receives r_{t,a_t}
    Learner estimates single-round reward vector r̂_t = (r_{t,a_t} / p_{t,a_t}) e_{a_t}
    û_t ← û_{t−1} + r̂_t
end for
```

We study the Gradient Based Prediction Algorithm (GBPA) [Abernethy et al., 2014, 2012, 2015; Kim and Tewari, 2019] for multi-armed bandits presented in Algorithm 1. At each round $t \in [T]$ of GBPA, the learner maintains an unbiased estimate of the cumulative reward $\hat{\boldsymbol{u}}_{t-1}$, updates $\hat{\boldsymbol{u}}_{t-1}$ by adding a single round estimate $\hat{\boldsymbol{r}}_t$, and uses the gradient of a convex potential function $\phi : \mathbb{R}^K \to \mathbb{R}$ evaluated at $\hat{\boldsymbol{u}}_{t-1}$ as an arm sampling distribution $\boldsymbol{p}_t$ from which the learner samples arm a_t. Although this framework might seem restrictive, it has proven foundational for several MAB algorithms, including but not limited to [Auer et al., 2002b; Kujala and Elomaa, 2005; Neu and Bartók, 2013]. Additionally, it encompasses various follow-the-leader type algorithms widely used in sequential decision-making processes with full information, differing primarily in the choice of the convex function ϕ used as an input to GBPA. In the following sections, we explain the policies of the learners in each setting given a cumulative reward estimate $\hat{\boldsymbol{u}}_t$.

Follow-the-leader (FTL). In the full information setting, GBPA with $\phi(\boldsymbol{u}) = \max_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top \boldsymbol{u}$ is known as FTL algorithm. The learner chooses the arm with the highest cumulative reward estimate, which can be equivalently written as $a_{t+1} \sim \boldsymbol{p}_t \in \operatorname{argmax}_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top \hat{\boldsymbol{u}}_t$.
Despite the simplicity of implementing Follow-the-Leader (FTL), it is well known that the regret of FTL can grow linearly with T even for the simple case of $K = 2$ when the adversary chooses the reward sequence to be $\boldsymbol{r}_1 = \{-1/2, 0\}$, $\boldsymbol{r}_t = \{-1, 0\}$ when $t > 1$ is odd and $\boldsymbol{r}_t = \{0, -1\}$ when t is even [Hazan, 2016, Chapter 5].

Follow-the-regularized-leader (FTRL). One of the most prominent approaches to stabilize the FTL algorithm is regularizing the linear objective $\boldsymbol{p}^\top \boldsymbol{u}$ with some convex function $\psi : \mathbb{R}^K \to \mathbb{R}$. In this case, the learner samples the next arm according to $\boldsymbol{p}_t$, where $\boldsymbol{p}_t \in \operatorname{argmax}_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top \hat{\boldsymbol{u}}_t - \psi(\boldsymbol{p})$. Then, GBPA with $\Phi^R(\boldsymbol{u}; \psi) = \max_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top \boldsymbol{u} - \psi(\boldsymbol{p})$ is known as the FTRL algorithm in the full information setting. In the adversarial regime, the FTRL method with Tsallis entropy, *i.e.,* GBPA$(\Phi^R(\cdot; \eta\psi^{\mathbb{T}}_\alpha))$, achieves the minimax optimal regret of $\mathcal{O}(\sqrt{KT})$ [Abernethy et al., 2015, Corollary 3.2], where $\eta = \sqrt{T(1-\alpha)/(2\alpha)}$ is the learning rate and $\psi^{\mathbb{T}}_\alpha$ is the Tsallis entropy with parameter $\alpha \in (0, 1)$

$$\psi^{\mathbb{T}}_\alpha(\boldsymbol{p}) = \frac{1 - \sum_{k=1}^K p_k^\alpha}{1-\alpha} \quad \forall \boldsymbol{p} \in \mathbb{R}^K. \tag{2}$$

Moreover, when the potential function is allowed to be adaptive, GBPA$(\Phi^R(\cdot; \eta_t\psi^{\mathbb{T}}_\alpha))$ achieves optimal regret of $\mathcal{O}(\log T)$ in the stochastic setting [Ito, 2021, Theorem 2]. An FTRL method using a hybrid regularizer combining Shannon entropy and Tsallis entropy is known to achieve optimal regret in both adversarial and stochastic settings [Zimmert et al., 2019].

Despite the widespread use of the FTRL framework with various choices of regularization, including optimal ones, computing arm sampling distributions at each iteration involves solving a convex optimization problem, making it computationally challenging.

Follow-the-perturbed-leader (FTPL). A promising candidate to circumvent the computational limitations of FTRL while maintaining the stability of FTRL is achieved by injecting stochastic noise $\boldsymbol{z} \sim \mathbb{Q}$ into the cumulative reward estimate. In that case, the learner samples the next arm according to $\boldsymbol{p}_t = \mathbb{E}_{\boldsymbol{z}\sim\mathbb{Q}}[\boldsymbol{e}_{k^\star(\boldsymbol{z})}]$, where $k^\star(\boldsymbol{z}) \in \operatorname{argmax}_{j\in[K]} \hat{u}_{t,j} + z_j$. Then, GBPA with $\Phi^P(\boldsymbol{u}; \mathbb{Q}) = \mathbb{E}_{\boldsymbol{z}\sim\mathbb{Q}}[\max_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top(\boldsymbol{u} + \boldsymbol{z})]$ is known as the FTPL algorithm in the full information setting.

While FTPL algorithms achieve computational efficiency by avoiding the need to solve complex optimization problems, their analysis is more cumbersome due to the perturbations introduced compared to the straightforward analysis of FTRL algorithms. Even though FTPL algorithms have shown BOBW capability [Honda et al., 2023; Lee et al., 2024], it is unclear whether it is possible to obtain a computationally efficient algorithm that simultaneously inherits the streamlined analysis of FTRL algorithms. One prominent approach to achieving this goal involves systematically identifying the perturbations for FTPL that coincide with the arm sampling distributions of FTRL, a task generally perceived as challenging [Honda et al., 2023]. Therefore, the following has been an important unresolved open problem seeking the existence of a bridge between regularization and perturbation-based algorithms.

Open Question: For some convex $\psi : \mathbb{R}^K \to \mathbb{R}$, is there a perturbation model with distribution $\mathbb{Q}$ that satisfies $\nabla_{\boldsymbol{u}} \Phi^P(\boldsymbol{u}; \mathbb{Q}) = \nabla_{\boldsymbol{u}} \Phi^R(\boldsymbol{u}; \psi)$ for all $\boldsymbol{u} \in \mathbb{R}^K$?

Since FTRL with Tsallis entropy regularizer can achieve the minimax optimal rate in adversarial bandits, a simpler version of the above open problem is posed by Kim and Tewari [2019] seeking the existence of an FTPL algorithm with the same arm sampling probability distribution as the FTRL algorithm with Tsallis entropy. Later, Kim and Tewari [2019, Theorem 8] shows that there is *no* stochastic perturbation that yields the same arm sampling probability distribution as the Tsallis entropy regularizer when the additive perturbations are mutually independent.

In the following section, we identify a general framework for constructing $\mathbb{Q}$ that positively answers the aforementioned open question. This was previously considered difficult or even impossible in the FTPL/GBPA literature [Honda et al., 2023; Kim and Tewari, 2019]. Our approach achieves this by studying *ambiguous* noise-sampling distributions that allow for *correlation* across the arms.

3 Distributionally Optimistic Perturbations

We now define the smooth potential function Φ as a best-case expected utility of the type studied in semi-parametric discrete choice theory, that is,

$$\Phi(\boldsymbol{u}; \mathcal{B}) = \sup_{\mathbb{Q} \in \mathcal{B}} \mathbb{E}_{\boldsymbol{z} \sim \mathbb{Q}} \left[\max_{k \in [K]} (u_k + z_k) \right], \tag{3}$$

where $\boldsymbol{z}$ represents a random vector of perturbations that are independent of $\boldsymbol{u}$. Specifically, we assume that $\boldsymbol{z}$ is governed by a Borel probability measure $\mathbb{Q}$ from within some *ambiguity set* $\mathcal{B} \subseteq \mathcal{P}(\mathbb{R}^K)$. Note that if $\mathcal{B}$ is a singleton that contains only the Dirac measure at the origin of $\mathbb{R}^K$, then Algorithm 1 with Φ coincides with FTL. If we denote by $\mathbb{Q}^\star$ an optimizer of (3), then $\Phi(\boldsymbol{u}; \{\mathbb{Q}^\star\})$ coincides with $\Phi^P(\boldsymbol{u}; \mathbb{Q}^\star)$ for all $\boldsymbol{u} \in \mathbb{R}^K$. Hence, GBPA with the potential function $\Phi(\cdot; \mathcal{B})$ generalizes traditional FTPL that injects i.i.d. noise into cumulative reward estimates. In particular, when $\mathcal{B}$ is a singleton joint probability measure with independent fixed marginals, GBPA with the potential function $\Phi(\cdot; \mathcal{B})$ is equivalent to conventional FTPL.

Remark 1 (Conventional FTPL as a special case)**.** *Fix any $\bar{\mathbb{Q}} \in \prod_{k=1}^K \mathcal{P}(\mathbb{R})$. If $\mathcal{B} = \{\mathbb{Q} \in \prod_{k=1}^K \mathcal{P}(\mathbb{R}) : \mathbb{Q}[z_k \leq s] = \bar{\mathbb{Q}}[z_k \leq s] \ \forall k \in [K]\}$, then we have $\Phi(\boldsymbol{u}; \mathcal{B}) = \Phi^P(\boldsymbol{u}; \bar{\mathbb{Q}})$.*

As a notable example within the FTPL family, our method also naturally encompasses the Exp3 algorithm [Auer et al., 1995]. This insight is detailed in the following remark.

Remark 2 (Exp3 algorithm as a special case)**.** *If $\mathcal{B}$ consists of a singleton distribution described by a Gumbel distribution, i.e., $\mathcal{B} = \{\mathbb{Q}\}$ where coordinates of $\boldsymbol{z} \sim \mathbb{Q}$ follow independent Gumbel distributions with means $\log(K)/\eta$ and variances $\pi^2/(6\eta^2)$, for some $\eta \in \mathbb{R}_{++}$. Then, the smooth potential $\Phi(\boldsymbol{u}; \mathcal{B})$ reduces to $\Phi(\boldsymbol{u}; \mathcal{B}) = \log(\sum_{k=1}^K \exp(\eta u_k))/\eta$, which follows from [Taşkesen et al., 2023, Proposition 3.4] and [McFadden, 1981, Theorem 5.2]. In this case, the arm-sampling probability vector $\boldsymbol{p}^\star(\boldsymbol{u}) \in \Delta^K$ admits a closed-form expression through $\boldsymbol{p}^\star(\boldsymbol{u})_k = \exp(\eta u_k)/(\sum_{j=1}^K \exp(\eta u_j))$, and GBPA$(\Phi(\cdot; \mathcal{B}))$ recovers the celebrated Exp3 algorithm [Auer et al., 1995].*

In the rest of our paper, we relax the i.i.d. noise assumption commonly adopted in the FTPL method and focus our attention on *marginal* ambiguity sets, also referred to as Fréchet ambiguity sets [Fréchet, 1951][1]. Marginal ambiguity sets completely specify the marginal distributions of the components of the random vector $\boldsymbol{z}$ but do not impose any constraints on their dependence structure. This relaxation allows us to recover many optimal FTRL methods in a systematic way.

Definition 1 (Marginal ambiguity set)**.** *For given cumulative distribution functions $\{F_k\}_{k=1}^K$, the marginal ambiguity set is defined as*

$$\mathcal{B} = \{\mathbb{Q} \in \mathcal{P}(\mathbb{R}^K) : \mathbb{Q}[z_k \leq s] = F_k(s) \ \forall s \in \mathbb{R}, \forall k \in [K]\}. \tag{4}$$

In the following, we will argue that marginal ambiguity sets explain most known as well as several new regularization methods in FTRL. As a first step, we state a known result initially established for discrete choice models that reformulates (3) as a regularized optimization problem, originally appeared in [Natarajan et al., 2009, Theorem 1] with an alternative proof provided in [Taşkesen et al., 2023, Proposition 3.6].

Lemma 3.1 ([Natarajan et al., 2009, Theorem 1])**.** *If $\mathcal{B}$ is a marginal ambiguity set of the form (4), and if the underlying cumulative distribution functions $F_k, k \in [K]$, are continuous, then the smooth potential function (3) can be equivalently expressed as*

$$\Phi(\boldsymbol{u}; \mathcal{B}) = \max_{\boldsymbol{p} \in \Delta^K} \sum_{k=1}^K u_k p_k + \sum_{k=1}^K \int_{1-p_k}^1 F_k^{-1}(t) \mathrm{d}t \tag{5}$$

for all $\boldsymbol{u} \in \mathbb{R}^K$. In addition, $\Phi(\boldsymbol{u}; \mathcal{B})$ is convex and differentiable with respect to $\boldsymbol{u} \in \mathbb{R}^K$, and $\nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u}; \mathcal{B})$ represents the unique solution of the convex program (5).

Note that the right-hand side of (5) is a sum of K strictly concave and differentiable functions $u_k p_k + \int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t$. Indeed, the derivative of the k^{th} function with respect to p_k is $u_k + F_k^{-1}(1-p_k)$, which is strictly decreasing in p_k because $F_k(s)$ is increasing in s and $1-p_k$ is strictly decreasing in p_k.

We are now positioned to formally address the open question presented in Section 2 with Corollary 3.2, which bridges the gap between regularization-based and perturbation-based algorithms for MAB problems.

Corollary 3.2. *Suppose that $\psi : \mathbb{R}^K \to \mathbb{R}$ is of the form $\psi(\boldsymbol{p}) = \sum_{k=1}^K \psi_k(p_k)$ where each ψ_k is strictly convex, differentiable and satisfies $\psi_k(0) = 0$. If $\mathcal{B}$ is a marginal ambiguity set of the form (4) with cumulative distribution functions F_k satisfying $-\int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t = \psi_k(p_k)$ for all $k \in [K]$, then $\Phi(\boldsymbol{u}; \mathcal{B}) = \Phi^R(\boldsymbol{u}; \psi(\boldsymbol{p}))$ for all $\boldsymbol{u} \in \mathbb{R}^K$.*

To differentiate our approach from traditional algorithms, we refer to our algorithm GBPA$(\Phi(\cdot; \mathcal{B}))$ prescribed by ambiguity set $\mathcal{B}$, as the Distributionally Optimistic Perturbation Algorithm (DOPA). The performance of DOPA, in terms of regret, varies with different ambiguity sets. This variation will be discussed in Section 4.

[1] Note that to the best of our knowledge, Fréchet ambiguity sets have no obvious relationship with Fréchet distributions.

4 Regret of DOPA

A fundamental principle in algorithm design is stability, which dictates that small perturbations in the input should not dramatically alter the algorithm's output. For GBPA with a convex differentiable potential function ϕ, the output corresponds to $\nabla_{\boldsymbol{u}}\phi(\boldsymbol{u})$, and stability is reflected through the Lipschitz continuity of the gradient. Unfortunately, not every regularizer leads to a convergent regret bound. Abernethy et al. [2014] demonstrated that a uniform bound on $\nabla^2_{\boldsymbol{u}}\phi(\boldsymbol{u})$ ensures a regret guarantee for GBPA(ϕ) in the full information setting. However, this condition does not directly transfer to the bandit setting, where the inverse scaling with respect to the arm sampling probability affects the cumulative reward estimation. Hence, an additional regularity condition on the potential function Φ, known as differential consistency [Abernethy et al., 2015], is required to ensure sublinear regret for DOPA.

Definition 2 (Differential consistency)**.** *For $\gamma, B > 0$, a function $g : \mathbb{R}^K \to \mathbb{R}$ is (γ, B)-differentially-consistent if for all $\boldsymbol{u} \in \mathbb{R}^K$*

$$(\nabla^2_{\boldsymbol{u}} g(\boldsymbol{u}))_{kk} \leq B(\nabla_{\boldsymbol{u}} g(\boldsymbol{u}))_k^{\gamma} \quad \forall k \in [K].$$

The following lemma translates the differential consistency condition on the potential function $\Phi(\cdot;\mathcal{B})$ into requirements on the marginal cumulative distribution functions that prescribe the ambiguity set $\mathcal{B}$.

Lemma 4.1. *Suppose $\mathcal{B}$ is a marginal ambiguity set of the form (4) where the marginal cumulative distribution functions F_k, $k \in [K]$ are twice differentiable on the interior of their respective supports. The potential function $\Phi(\cdot;\mathcal{B})$ is (γ, B)-differentially-consistent if*

$$F_k'(F_k^{-1}(1-p)) \leq Bp^{\gamma} \quad \forall p \in (0,1),\ \forall k \in [K]. \tag{6}$$

Equipped with Lemma 4.1, we are now prepared to present an upper bound on the expected regret for GBPA($\Phi(\cdot;\mathcal{B})$).

Theorem 4.2. *Suppose $\mathcal{B}$ is a marginal ambiguity set of the form (4) where the marginal cumulative distribution functions F_k, $k \in [K]$ are twice differentiable on the interior of their respective supports. If additionally $\mathcal{B}$ encompasses distributions with zero mean and F_k is $(\gamma, B(T))$-differentially consistent for all $k \in [K]$ with $\gamma \in (1,2)$, then the regret of GBPA($\Phi(\cdot;\mathcal{B})$) ensures*

$$\mathcal{R}(T) \leq \sum_{k=1}^{K} \int_{1-p_{0,k}}^{1} F_k^{-1}(t)\mathrm{d}t + \frac{1}{2}B(T)TK^{2-\gamma}.$$

5 Optimal Ambiguity Sets

This section identifies the instances of the marginal ambiguity sets $\mathcal{B}$ that allow GBPA($\Phi(\cdot;\mathcal{B})$) to recover FTRL algorithms that achieve optimal regret bounds in designated regimes. Within these specific settings, the unresolved conjectures that DOPA addresses become particularly relevant concerning the recoverability of FTRL algorithms through the application of FTPL methods. First, we introduce a structured approach to defining the marginal cumulative distribution functions that prescribe ambiguity sets $\mathcal{B}$ and systematically demonstrate the corresponding forms of regularization.

Theorem 5.1 (Fréchet regularization)**.** *Suppose that $\mathcal{B}$ is a marginal ambiguity set of the form* (4) *and that the marginal cumulative distribution functions are defined through*

$$F_k(s) = \min\{1, \max\{0, 1 - F(-s/\eta_k)\}\} \tag{7}$$

for some vector $\boldsymbol{\eta} \in \mathbb{R}^K_{++}$ and strictly increasing function $F : \mathbb{R} \to \mathbb{R}$ with $\int_0^1 F^{-1}(t)\mathrm{d}t = 0$. Then, $\Phi(\boldsymbol{u}; \mathcal{B})$ is equivalent to $\Phi^R(\boldsymbol{u}; \psi)$, where $\psi(\boldsymbol{p}) = \sum_{k=1}^K \eta_k f(p_k)$ and $f(s) = \int_0^s F^{-1}(t)\mathrm{d}t$.

The function $f(s)$ introduced in Theorem 5.1 is smooth and convex because its derivative $\mathrm{d}f(s)/\mathrm{d}s = F^{-1}(s)$ is strictly increasing. From now on we will refer to F as the marginal generating function. We now tailor the regret upper bound presented in Theorem 4.2 to cases where marginal cumulative functions are generated by F.

Corollary 5.2 (Regret bound for marginal ambiguity sets of the Fréchet form)**.** *Suppose $\mathcal{B}$ is a marginal ambiguity set of the form* (7) *for some vector $\boldsymbol{\eta} = \eta\mathbf{1} \in \mathbb{R}^K_{++}$ encompassing distributions with zero mean and $\Phi(\cdot; \mathcal{B})$ is (γ, B)-differentially-consistent. Then, GBPA$(\Phi(\cdot; \mathcal{B}))$ ensures*

$$\mathcal{R}(T) \leq -\eta K f(1/K) + \frac{BTK^{2-\gamma}}{2\eta}.$$

In the adversarial setting, the optimal regret is established as $\mathcal{O}(\sqrt{KT})$ [Audibert and Bubeck, 2009, 2010], and is achievable by FTRL methods using Tsallis entropy [Abernethy et al., 2012]. Recently, there has been significant interest in determining whether perturbation-based methods can match the efficacy of FTRL techniques. Kim and Tewari [2019] has shown that no stochastic perturbation can reproduce the choice probability function of the Tsallis entropy regularizer when the additional random noise in each arm utility is independent. Despite this, it was conjectured that $\mathcal{O}(\sqrt{KT})$ regret might be attainable through FTPL with Fréchet-type perturbations. More recently, Honda et al. [2023] demonstrated that an FTPL with a Fréchet perturbation indeed achieves $\mathcal{O}(\sqrt{KT})$ regret. Nevertheless, whether an FTRL algorithm with Tsallis entropy, achieving this optimal rate, can be replicated by a perturbation-based algorithm has remained open until now [Honda et al., 2023; Kim and Tewari, 2019] and is resolved by the following theorem.

Theorem 5.3. *Suppose that $\mathcal{B}$ is a marginal ambiguity set with (shifted) Pareto distributed marginals of the form* (7) *induced by the marginal generating function $F(s) = (s(\alpha-1)/\alpha + 1/\alpha)^{\frac{1}{\alpha-1}}$ with $\alpha \in (0, 1)$. Then, GBPA$(\Phi(\cdot; \mathcal{B}))$ with $\eta_k = \sqrt{(T(1-\alpha))/(2\alpha)}K^{\alpha-\frac{1}{2}}$ for all $k \in [K]$ satisfies $\mathcal{R}(T) \leq \sqrt{KT/(\alpha(1-\alpha))}$. In particular, when $\alpha = 1/2$, GBPA$(\Phi(\cdot; \mathcal{B}))$ ensures $\mathcal{R}(T) \leq 2\sqrt{2KT}$.*

Note that the Exp3 algorithm can be realized as a special case of DOPA when $\mathcal{B}$ is a marginal ambiguity set with shifted exponential marginals. Therefore, the Exp3 algorithm is not only induced by a singleton distribution as in Remark 2, but also induced by marginal ambiguity sets of the form (4) with exponential marginals.

Remark 3 (Exp3 algorithm revisited)**.** *Suppose that $\mathcal{B}$ is a marginal ambiguity set of the form* (7)*, where $\boldsymbol{\eta} = \eta\mathbf{1} \in \mathbb{R}^K_{++}$ and $F(s) = \exp(-s-1)$. Then, DOPA with $\mathcal{B}$ is equivalent to FTRL with $\psi(\boldsymbol{p}) = \sum_{k=1}^K p_k \log(p_k)$ [Abernethy et al., 2014, Section 3].*

Due to the mathematical equivalence between employing FTRL with Tsallis regularization and using DOPA with shifted Pareto marginals, the attractive BOBW capability of FTRL can be directly extended to DOPA. Additionally, the algorithm exhibits anytime properties; it does not require knowledge of the time horizon T nor the use of doubling schemes. This relationship is detailed in Theorem 5.4.

Algorithm 2 Anytime GBPA for MAB

Require: $(\phi_t)_{t=1,2,\ldots}$ with $\nabla_{\boldsymbol{u}}\phi_t(\boldsymbol{u}) \in \Delta^K$

$\hat{u}_{0,k} \leftarrow 0 \quad \forall k \in [K]$
for round $t = 1, \ldots$ **do**
  A reward vector $\boldsymbol{r}_t \in [-1, 0]^K$ is chosen by the environment
  Learner chooses $a_t \sim \boldsymbol{p}_t = \nabla_{\boldsymbol{u}}\phi_t(\boldsymbol{u})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{t-1}}$
  Learner receives r_{t,a_t}
  Learner estimates single-round reward vector $\hat{\boldsymbol{r}}_t = (r_{t,a_t}/p_{t,a_t})\boldsymbol{e}_{a_t}$
  $\hat{\boldsymbol{u}}_t \leftarrow \hat{\boldsymbol{u}}_{t-1} + \hat{\boldsymbol{r}}_t$
end for

Theorem 5.4 (Anytime BOBW algorithm with adaptive perturbations)**.** *Suppose that $\mathcal{B}_t$, $t \in \mathbb{Z}_+$ is a marginal ambiguity set of the form*

$$\mathcal{B}_t = \{\mathbb{Q} \in \mathcal{P}(\mathbb{R}^K) : \mathbb{Q}[z_k \leq s] = F_{t,k}(s) \ \forall s \in \mathbb{R}, \forall k \in [K]\}$$

with (shifted) Pareto-distributed marginals,

$$F_{t,k}(s) = \min\{1, \max\{0, 1 - (s/\eta_t + 2)^{-2}\}\}.$$

Then, for any $T_0 \in \mathbb{N}$, GBPA$(\Phi(\cdot; \mathcal{B}_t))$ with $\eta_t = 2\sqrt{t}$ ensures $\mathcal{R}(T_0) \leq 4\sqrt{KT_0}+1$ always, and simultaneously the following regret bound if the adversary satisfies (1): $\mathcal{R}(T_0) \leq \mathcal{O}(\sum_{k\in[K]:\Delta_k>0} \log(T_0)/\Delta_k)$.

In addition to replicating FTRL using a regularization function derived from a single marginal generator function, DOPA also effectively replicates hybrid regularizers. These are systematically derived from two marginal generator functions, as detailed in Corollary 5.5.

Corollary 5.5 (Hybrid Fréchet regularization)**.** *Suppose that $\mathcal{B}$ is a marginal ambiguity set of the form (4), and fix $\boldsymbol{\gamma}, \boldsymbol{\eta} \in \mathbb{R}^K_+$. Suppose further that the marginal cumulative distribution functions are defined through*

$$F_k(s) = \min\{1, \max\{0, 1 - (\gamma_k G_1^{-1} + \eta_k G_2^{-1})^{-1}(-s)\}\} \tag{8}$$

where $G_1, G_2 : \mathbb{R} \to \mathbb{R}$ are strictly increasing functions with $\int_0^1 G_1^{-1}(t)\mathrm{d}t = \int_0^1 G_2^{-1}(t)\mathrm{d}t = 0$. Then, $\Phi(\boldsymbol{u}; \mathcal{B})$ is equivalent to $\Phi^R(\boldsymbol{u}; \psi)$, where $\psi(\boldsymbol{p}) = \sum_{k=1}^K(\eta_k f(p_k) + \gamma_k g(p_k))$, $f(s) = \int_0^s G_1^{-1}(t)\mathrm{d}t$, and $g(s) = \int_0^s G_2^{-1}(t)\mathrm{d}t$.

Remark 4 (Generalization to N regularization functions)**.** *All existing theoretically optimal algorithms incorporate a hybrid regularizer that combines only two regularization terms [Jin et al., 2024; Zimmert et al., 2019]. However, it is worth noting that instead of having two generating functions G_1, G_2, one can define marginal cumulative distribution functions F_k through N generating functions $G_1, \ldots, G_N$ and obtain a regularizer as a sum of N integrals of quantile functions G_i^{-1} for $i \in [N]$.*

The following corollary demonstrates that our generalized mixed perturbation formulation can indeed achieve theoretically optimal BOBW results through its equivalent formulation as hybrid regularized FTRL. To the best of our knowledge, it has not been known whether any FTPL algorithm could recover an FTRL algorithm with hybrid regularizers.

Corollary 5.6 (Hybrid adaptive regularizers for bandits)**.** *Suppose that $\mathcal{B}_t$, $t \in \mathbb{Z}_+$ is a marginal ambiguity set of the form*

$$\mathcal{B}_t = \{\mathbb{Q} \in \mathcal{P}(\mathbb{R}^K) : \mathbb{Q}[z_k \leq s] = F_{t,k}(s) \ \forall s \in \mathbb{R}, \forall k \in [K]\},$$

with

$$F_{t,k}(s) = \min\{1, \max\{0, 1 - (\gamma_{t,k} G_1^{-1} + \eta_{t,k} G_2^{-1})^{-1}(-s)\}\}.$$

If $G_1(s) = 1 - \exp(-(s+1))$, $G_2(s) = (-2s)^{-2}$, with $\gamma_{t,k} = \eta_{t,k} = \sqrt{t}$ for all $t \in [T]$ and $k \in [K]$, then, GBPA($\Phi(\cdot; \mathcal{B}_t)$) ensures $\mathcal{R}(T) \leq \mathcal{O}(\sqrt{KT})$ always, and simultaneously the following regret bound if the optimal arm is unique and the adversary satisfies (1): $\mathcal{R}(T) \leq \mathcal{O}(\sum_{k \neq k^\star} \log T/\Delta_k) + \mathcal{O}(\sum_{k \neq k^\star} (\log K)^2/\Delta_k)$.

Our algorithmic framework extends beyond the K-armed bandit setting while maintaining BOBW capability. Notable examples include the decoupled-exploitation-exploration setting [Jin et al., 2024], where the learner can choose to receive a reward from one arm while obtaining information about the reward from another arm. Another example where our algorithm can be applied and achieve BOBW regret bound is the dueling bandit setting [Zimmert and Seldin, 2021], where, in each round, two arms are chosen to "duel" and feedback is received for the arm with the higher reward. Additionally, our framework recovers the hybrid Tsallis entropy regularizers used in an FTRL-type algorithm with BOBW capability [Ito et al., 2024] for both K-armed bandit and linear bandit problems.

6 Numerical Experiments

FTPL-type algorithms have been popular because the arm sampling probability distributions appearing in Line 4 of Algorithm 1 can be computed efficiently when the perturbations z_k are i.i.d. [Neu and Bartók, 2016]. On the other hand, FTRL-type algorithms fall short of this computational benefit because an optimization problem at each round has to be solved to compute the arm-sampling probabilities. This section discusses how the arm-sampling probabilities of DOPA admit an efficient computation even when we relax the usual assumption in FTPL that the additive noise components are independent.

Surprisingly, one can still apply a computationally efficient bisection algorithm to find the optimal choice probabilities [Taşkesen et al., 2023, Algorithm 2]. This bisection method uses techniques from discrete choice theory to exploit the structure of the marginal ambiguity set. As a result, it is inherently faster than solving an expensive optimization problem. Algorithm 3 enjoys the following convergence guarantee.

Theorem 6.1 (Convergence guarantee of Algorithm 3 [Taşkesen et al., 2023, Theorem 4.9])**.** *If $\mathcal{B}$ is a marginal ambiguity set of the form (4) and the cumulative distribution function F_k is continuous for every $k \in [K]$, then, for any $\boldsymbol{u} \in \mathbb{R}^K$ and $\varepsilon > 0$, Algorithm 3 outputs $\boldsymbol{p} \in \mathbb{R}_+^K$ with $\sum_{k=1}^K p_k \leq 1$ and $\|\nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u}, \mathcal{B}) - \boldsymbol{p}\| \leq \varepsilon$.*

Algorithm 3 Bisection method to approximate $\nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u}; \mathcal{B})$

Require: error tolerance ε, utility vector $\boldsymbol{u}$, marginal ambiguity set $\mathcal{B}$

Set $\bar{\tau} \leftarrow \max_{k \in [K]} \{-u_k - F_k^{-1}(1 - 1/K)\}$
Set $\underline{\tau} \leftarrow \min_{k \in [K]} \{-u_k - F_k^{-1}(1 - 1/K)\}$
Evaluate $\delta(\varepsilon) = \min_{k \in [K]} \{\max_\delta \{\delta : |F_k(t_1) - F_k(t_2)| \leq \varepsilon / \sqrt{K}\ \forall t_1, t_2 \in \mathbb{R} \text{ with } |t_1 - t_2| \leq \delta\}\}$

for $k = 1, 2, \ldots, \lceil \log_2((\bar{\tau} - \underline{\tau})/\delta(\varepsilon)) \rceil$ **do**
  Set $\tau \leftarrow (\tau + \underline{\tau})/2$
  Set $p_k \leftarrow 1 - F_k(-u_k - \tau)$ for $k \in [K]$
  if $\sum_{k \in [K]} p_k > 1$ **then** $\bar{\tau} \leftarrow \tau$
  else $\underline{\tau} \leftarrow \tau$ **end if**
end for
return $\boldsymbol{p}$ with $p_k = 1 - F_k(-u_k - \underline{\tau}), k \in [K]$

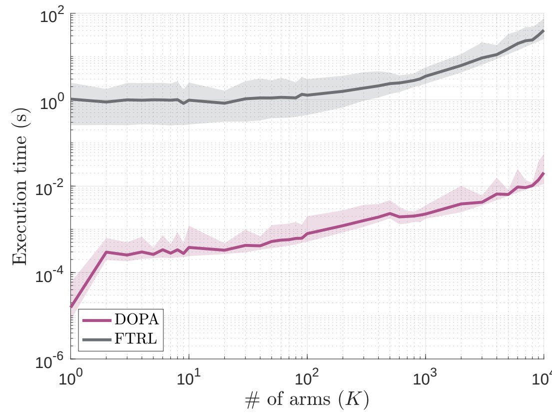

Figure 1: Runtime of computing arm-sampling probabilities using FTRL (gray) and DOPA (purple) over 10 simulation runs (solid lines show the mean and the shaded areas correspond to 1 standard deviation) as a function of number of arms K.

In the following, we empirically demonstrate the computational efficiency of DOPA relative to FTRL by evaluating the runtimes required to compute arm-sampling probabilities. All experiments are run on an Intel i7-8700 CPU (3.2 GHz) computer with 16GB RAM. The optimization problems are modelled in MATLAB via YALMIP [McCormick, 1976]. The code is publicly available at https://github.com/RAO-EPFL/DOPA.

For DOPA, we choose $\mathcal{B}$ of the form (7) with $F(s) = (-s+2)^{-2}$ and $\eta = 1$. In the case of FTRL, we utilize Tsallis regularization and establish that the arm sampling distributions of both algorithms are equivalent, such that $\nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u}; \mathcal{B}) = \nabla_{\boldsymbol{u}} \Phi^R(\boldsymbol{u}; \eta \psi^{\mathbb{T}}_{1/2})$ for all $\boldsymbol{u} \in \mathbb{R}^K$. We calculate $\nabla_{\boldsymbol{u}} \Phi^R(\boldsymbol{u}; \psi^{\mathbb{T}}_{1/2}) = \operatorname{argmax}_{\boldsymbol{p} \in \Delta^K} \boldsymbol{p}^\top \boldsymbol{u} - \psi^{\mathbb{T}}_{1/2}(\boldsymbol{p})$ by solving the corresponding second-order-cone program using MOSEK [Mosek ApS, 2019]. We employ Algorithm 3 to approximate $\nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u}; \mathcal{B})$ with an error tolerance of $\varepsilon = 10^{-8}$, matching the optimality tolerance used by MOSEK for conic problems. Figure 1 highlights that DOPA achieves running times that are uniformly lower than FTRL with Tsallis entropy across all numbers of arms and is, in fact, up to 10^4 times faster.

7 Concluding Remarks and Limitations

We introduce a distributional "optimism in the face of ambiguity" principle to determine the noise distribution for FTPL-type algorithms in multi-armed bandit problems. This principle allows us to establish a one-to-one correspondence between FTRL algorithms with separable strictly convex regularizers and FTPL algorithms. Hence, our algorithm bypasses the difficulties in analyzing FTPL-type algorithms and lifts the computational burden of FTRL by devising an efficient bisection algorithm using ideas from modern discrete choice theory. DOPA aims to provide a unified regret analysis for perturbation-based methods through FTRL and opens doors to the discovery of new algorithms. We find it promising to study other types of ambiguity sets or other types of regularizers induced by marginal ambiguity sets, such as hyperbolic perturbations [Taşkesen et al., 2023, Example 3.11].

At the same time, we acknowledge the limitations of this work. First, certain types of regularizers cannot be recovered by marginal ambiguity sets $\mathcal{B}$ of the form (4), with a notable example of the log-barrier regularizer as considered in [Jin et al., 2024]. While whether the log-barrier regularizer is essential in showing the BOBW guarantee of FTRL algorithms remains unclear [Jin et al., 2024], [Hofbauer and Sandholm, 2002, Proposition 2.2] demonstrates that it is *im*possible to recover the FTRL algorithm with the log-barrier regularizer using any FTPL algorithm with a stochastic perturbation whose distribution is independent of the underlying utilities. Second, the bisection algorithm presented in Algorithm 3 is efficient as long as the computation of the marginal cumulative distributions F_k and the quantile function F_k^{-1} are efficient. However, for hybrid regularizers, the computation of F_k could be cumbersome. As a result, bisection method might not be computationally efficient for some choices of hybrid marginal generating functions G_1 and G_2.

Acknowledgements

ML and DK are supported by the Swiss National Science Foundation under the NCCR *Automation*, grant agreement 51NF40_180545.

References

J. Abernethy, C. Lee, A. Sinha, and A. Tewari. Online linear optimization via smoothing. *Conference on Learning Theory*, 2014.

J. Abernethy, C. Lee, and A. Tewari. Perturbation techniques in online learning and optimization. *Perturbations, Optimization, and Statistics*, 233:17, 2016.

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

J. D. Abernethy, C. Lee, and A. Tewari. Fighting bandits with a new kind of smoothness. *Advances in Neural Information Processing Systems*, 2015.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. *Conference on Learning Theory*, 2009.

J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *IEEE 36th Annual Foundations of Computer Science*, 1995.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. *Conference on Learning Theory*, 2012.

M. Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Annales de lUniversité de Lyon, Sciences*, 4(1/2):13–84, 1951.

G. J. Gordon. Regret bounds for prediction problems. *Conference on Computational Learning Theory*, 1999.

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

E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

J. Hofbauer and W. H. Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

J. Honda, S. Ito, and T. Tsuchiya. Follow-the-Perturbed-Leader Achieves Best-of-Both-Worlds for Bandit Problems. *International Conference on Algorithmic Learning Theory*, 2023.

S. Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. *Conference on Learning Theory*, 2021.

S. Ito, T. Tsuchiya, and J. Honda. Adaptive learning rate for follow-the-regularized-leader: Competitive ratio analysis and best-of-both-worlds. *arXiv:2403.00715*, 2024.

T. Jin, J. Liu, and H. Luo. Improved best-of-both-worlds guarantees for multi-armed bandits: FTRL with general regularizers and multiple optimal arms. *Advances in Neural Information Processing Systems*, 2024.

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

J. Kujala and T. Elomaa. On following the perturbed leader in the bandit setting. *International Conference on Algorithmic Learning Theory*, 2005.

T. Lattimore and C. Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

J. Lee, J. Honda, S. Ito, and M.-h. Oh. Follow-the-perturbed-leader with Fréchet-type tail distributions: Optimality in adversarial bandits and best-of-both-worlds. *arXiv:2403.05134*, 2024.

T. Lykouris, V. Mirrokni, and R. Paes Leme. Stochastic bandits robust to adversarial corruptions. *ACM Symposium on Theory of Computing*, 2018.

G. P. McCormick. Computability of global solutions to factorable nonconvex programs: Part IConvex underestimating problems. *Mathematical programming*, 10(1):147–175, 1976.

D. McFadden. Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications*, 1981.

E. Melo and D. Müller. Discrete choice multi-armed bandits. *arXiv:2310.00562*, 2023.

Mosek ApS. Mosek optimization toolbox for matlab. *User's Guide and Reference Manual, Version*, 4(1), 2019.

K. Natarajan, M. Song, and C.-P. Teo. Persistency model and its applications in choice modeling. *Management Science*, 55(3):453–469, 2009.

G. Neu and G. Bartók. An efficient algorithm for learning with semi-bandit feedback. *International Conference on Algorithmic Learning Theory*, 2013.

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

T. Sun and Q. Tran-Dinh. Generalized self-concordant functions: A recipe for Newton-type methods. *Mathematical Programming*, 178(1):145–213, 2019.

B. Taşkesen, S. Shafieezadeh-Abadeh, and D. Kuhn. Semi-discrete optimal transport: Hardness, regularization and numerical solution. *Mathematical Programming*, 199(1-2):1033–1106, 2023.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

C.-Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. *Conference on Learning Theory*, 2018.

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

J. Zimmert, H. Luo, and C.-Y. Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. *International Conference on Machine Learning*, 2019.

A Auxiliary Results

The following corollary sheds light on the condition of the properties of the cumulative distribution functions F_k that induce a strongly convex regularizer. Its proof follows from the smoothness/strong convexity duality and specifically the proof of [Taşkesen et al., 2023, Proposition 4.8]. For completeness, we include the full proof here.

Corollary A.1. *If $\mathcal{B}$ is a marginal ambiguity set of the form (4), and if the cumulative distribution functions $F_k, k \in [K]$, are Lipschitz continuous with Lipschitz constant L, then $\sum_{k=1}^K \int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t$ is L-strongly concave.*

Proof of Corollary A.1. We aim to show that $-\sum_{k=1}^K \int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t - p_k^2/(2L)$ is convex. By the assumed L-Lipschitz continuity of F_k, we have

$$L \geq \sup_{\substack{s_1,s_2\in\mathbb{R}\\ s_1\neq s_2}} \frac{|F_k(s_1) - F_k(s_2)|}{|s_1 - s_2|} = \sup_{\substack{s_1,s_2\in\mathbb{R}\\ s_1> s_2}} \frac{F_k(s_1) - F_k(s_2)}{s_1 - s_2} \geq \sup_{\substack{p_k,q_k\in(0,1)\\ p_k>q_k}} \frac{p_k - q_k}{F_k^{-1}(p_k) - F_k^{-1}(q_k)}$$

where the second inequality follows from restricting s_1 and s_2 to the preimage of $(0,1)$ with respect to F_k. Rearranging terms in the above inequality then yields

$$-F_k^{-1}(1-q_k) - \frac{q_k}{L} \leq -F_k^{-1}(1-p_k) - \frac{p_k}{L}$$

for all $p_k, q_k \in (0,1)$ such that $q_k < p_k$. Consequently, the function $-F_k^{-1}(1-p_k) - p_k/L$ is non-decreasing and its primitive $-\int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t - p_k^2/(2L)$ is convex in p_k on $(0,1)$. The claim then follows because convexity is preserved under summation. $\square$

We use $\mathbb{E}_t[\cdot]$ as a shorthand for the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$, where $\mathcal{F}_t$ is the σ-algebra $\sigma(a_1, \boldsymbol{r}_1, \ldots, a_t, \boldsymbol{r}_t)$ generated by the history of actions and rewards.

Lemma A.2 (Regret bound for GBPA [Abernethy et al., 2015, Lemma 2.1])**.** *For some convex potential function $\phi : \mathbb{R}^K \to \mathbb{R}$, the expected regret of GBPA(ϕ) enjoys the following upper bound expressed through*

$$\mathcal{R}(T) \leq \phi(\mathbf{0}) + \mathbb{E}\left[\max_{k\in[K]} \hat{u}_{T,k} - \phi(\hat{\boldsymbol{u}}_T) + \sum_{t=1}^T \mathbb{E}_t[\mathbb{D}_\phi(\hat{\boldsymbol{u}}_t, \hat{\boldsymbol{u}}_{t-1})]\right].$$

Theorem A.3 (Regret bound for marginal ambiguity sets)**.** *Suppose $\mathcal{B}$ is a marginal ambiguity set of the form (4) where the marginal cumulative distribution functions $F_k, k \in [K]$ are twice differentiable on the interior of their respective supports. Define $\boldsymbol{p}^\star(\boldsymbol{u}) = \nabla_{\boldsymbol{u}}\Phi(\boldsymbol{u}; \mathcal{B})$ for any $\boldsymbol{u} \in \mathbb{R}^K$. The regret of GBPA($\Phi(\cdot;\mathcal{B})$) satisfies*

$$\begin{aligned}\mathcal{R}(T) \leq &\sum_{k=1}^K \int_{1-p^\star_{0,k}}^1 F_k^{-1}(t)\mathrm{d}t + \mathbb{E}\left[\max_{k\in[K]} \hat{u}_{T,k} - \Phi(\hat{\boldsymbol{u}}_T;\mathcal{B})\right] \\ &+ \mathbb{E}\left[\sum_{t=1}^T \sum_{k=1,\ p_{t,k}>0}^K p_{t,k} \int_0^{\|\hat{r}_t\|}\int_0^x F_k'(F_k^{-1}(1-\boldsymbol{p}^\star(\hat{\boldsymbol{u}}_{t-1} - s\boldsymbol{e}_{a_t})_k))\ \mathrm{d}s\ \mathrm{d}x\right].\end{aligned} \tag{9}$$

Proof of Theorem A.3. By Lemma 3.1 and Lemma A.2, we have

$$\mathcal{R}(T) \leq \sum_{k=1}^{K} \int_{1-p_{0,k}}^{1} F_k^{-1}(t)\mathrm{d}t + \mathbb{E}\left[\max_{k\in[K]} \hat{u}_{T,k} - \Phi(\hat{\boldsymbol{u}}_T;\mathcal{B}) + \sum_{t=1}^{T} \mathbb{E}_t[\mathbb{D}_\Phi(\hat{\boldsymbol{u}}_t, \hat{\boldsymbol{u}}_{t-1})]\right].$$

In what follows, we will establish an upper bound on the term $\mathbb{E}_t[\mathbb{D}_\Phi(\hat{\boldsymbol{u}}_t, \hat{\boldsymbol{u}}_{t-1})]$. For any $t \in [T]$, conditioning on the event that arm $a_t \in [K]$ is chosen, we define $h : \mathbb{R}_+ \to \mathbb{R}$ as

$$h(s) = \mathbb{D}_\Phi(\hat{\boldsymbol{u}}_{t-1} + s\hat{\boldsymbol{r}}_t/\|\hat{\boldsymbol{r}}_t\|, \hat{\boldsymbol{u}}_{t-1}).$$

A direct calculation reveals that the second derivative of h satisfies

$$\begin{aligned} h''(s) &= (\hat{\boldsymbol{r}}_t/\|\hat{\boldsymbol{r}}_t\|)^\top (\nabla^2_{\boldsymbol{u}}\Phi(\boldsymbol{u};\mathcal{B})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{t-1}+s\hat{\boldsymbol{r}}_t/\|\hat{\boldsymbol{r}}_t\|})(\hat{\boldsymbol{r}}_t/\|\hat{\boldsymbol{r}}_t\|) \\ &= \boldsymbol{e}_{a_t}^\top (\nabla^2_{\boldsymbol{u}}\Phi(\boldsymbol{u};\mathcal{B})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{t-1}-s\boldsymbol{e}_{a_t}})\boldsymbol{e}_{a_t}, \end{aligned}$$

where the second equality follows because $r_t \in [-1,0]^K$ and thus $\boldsymbol{e}_{a_t} = -\hat{\boldsymbol{r}}_t/\|\hat{\boldsymbol{r}}_t\|$. Then, we have

$$\begin{aligned} \mathbb{E}_t[\mathbb{D}_\Phi(\hat{\boldsymbol{u}}_t, \hat{\boldsymbol{u}}_{t-1})] &= \sum_{k=1,\ p_{t,k}>0}^{K} p_{t,k} \int_0^{\|\hat{r}_t\|}\int_0^x h''(s)\ \mathrm{d}s\ \mathrm{d}x \\ &= \sum_{k=1,\ p_{t,k}>0}^{K} p_{t,k} \int_0^{\|\hat{r}_t\|}\int_0^x \boldsymbol{e}_{a_t}^\top (\nabla^2_{\boldsymbol{u}}\Phi(\boldsymbol{u};\mathcal{B})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{t-1}-s\boldsymbol{e}_{a_t}})\boldsymbol{e}_{a_t}\ \mathrm{d}s\ \mathrm{d}x \\ &= \sum_{k=1,\ p_{t,k}>0}^{K} p_{t,k} \int_0^{\|\hat{r}_t\|}\int_0^x F_k'(F_k^{-1}(1-\boldsymbol{p}^\star(\hat{\boldsymbol{u}}_{t-1}-s\boldsymbol{e}_{a_t})_k))\ \mathrm{d}s\ \mathrm{d}x, \end{aligned}$$

where the first equality holds thanks to the fundamental theorem of calculus, and the third equality follows from [Sun and Tran-Dinh, 2019, Proposition 6].

□

Similar to Lemma 4.1, we can translate the differential consistency condition into requirements on the noise distribution.

Lemma A.4 (Differential consistency condition for marginal ambiguity sets of the Fréchet form)**.** *Suppose $\mathcal{B}$ is a marginal ambiguity set of the form (7) for some vector $\boldsymbol{\eta} = \eta\boldsymbol{1} \in \mathbb{R}^K_{++}$ where the marginal generating function F is twice differentiable on the interior of its support. The corresponding potential function $\Phi(\cdot;\mathcal{B})$ is (γ, B)-differentially-consistent if $f(s) = \int_0^s F^{-1}(t)\mathrm{d}t < \infty$ satisfies*

$$(\eta f''(p))^{-1} \leq Bp^\gamma \quad \forall p \in (0,1). \tag{10}$$

Proof of Lemma A.4. Denote by $\boldsymbol{p}^\star(\boldsymbol{u}) = \nabla_{\boldsymbol{u}}\Phi(\boldsymbol{u};\mathcal{B})$ which represents the unique solution of the optimization problem (5) by Lemma 3.1. By [Sun and Tran-Dinh, 2019, Proposition 6], the Hessian of $\Phi(\cdot;\mathcal{B})$ can be expressed through the Hessian of its convex conjugate. We then have

$$\begin{aligned} (\nabla^2_{\boldsymbol{u}}\Phi(\boldsymbol{u};\mathcal{B}))_{kk} &= ((\nabla^2_{\boldsymbol{p}}\Phi^*(\boldsymbol{p};\mathcal{B})|_{\boldsymbol{p}=\boldsymbol{p}^\star(\boldsymbol{u})})^{-1})_{kk} \\ &= \frac{1}{\eta f''(\boldsymbol{p}^\star(\boldsymbol{u})_k)} \leq B(\boldsymbol{p}^\star(\boldsymbol{u})_k)^\gamma = (\nabla_{\boldsymbol{u}}(\Phi(\boldsymbol{u};\mathcal{B}))_k)^\gamma, \end{aligned}$$

where the second equality follows from Theorem 5.1, and the inequality holds because f satisfies (10). Hence, the claim follows. □

B Omitted Proofs

Proof of Corollary 3.2. By the strict convexity of ψ_k, we deduce that $\psi_k'(s)$ is strictly increasing in s. Hence, there exist strictly increasing functions G_k such that $G_k(s) = -\psi_k'(1-s)$ for any $s \in [0,1]$. In addition, we have $-\int_{1-p_k}^1 G_k(t)\mathrm{d}t = \psi_k(p_k)$ thanks to the fundamental theorem of calculus and the assumption that $\psi_k(0) = 0$. Choosing $F_k(s) = \min\{1, \max\{0, G_k^{-1}(s)\}\}$ gives $F_k^{-1}(s) = G_k(s)$ for all $s \in (0,1)$ and the desired relation $-\int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t = -\int_{1-p_k}^1 G_k(t)\mathrm{d}t = \psi_k(p_k)$. Applying Lemma 3.1 concludes the proof. □

Proof of Lemma 4.1. Denote by $\boldsymbol{p}^\star(\boldsymbol{u}) = \nabla_{\boldsymbol{u}}\Phi(\boldsymbol{u};\mathcal{B})$ which represents the unique solution of the optimization problem (5) by Lemma 3.1. By [Sun and Tran-Dinh, 2019, Proposition 6], the Hessian of $\Phi(\cdot;\mathcal{B})$ can be expressed through the Hessian of its convex conjugate, and thus we have

$$\begin{aligned}(\nabla_{\boldsymbol{u}}^2\Phi(\boldsymbol{u};\mathcal{B}))_{kk} &= ((\nabla_{\boldsymbol{p}}^2\Phi^*(\boldsymbol{p};\mathcal{B})|_{\boldsymbol{p}=\boldsymbol{p}^\star(\boldsymbol{u})})^{-1})_{kk}\\ &= F_k'(F_k^{-1}(1-\boldsymbol{p}^\star(\boldsymbol{u})_k))\\ &\le B(\boldsymbol{p}^\star(\boldsymbol{u})_k)^\gamma = B(\nabla_{\boldsymbol{u}}(\Phi(\boldsymbol{u};\mathcal{B}))_k)^\gamma,\end{aligned}$$

where the second equality follows by the inverse function theorem together with the fact that $\Phi^*(\boldsymbol{p};\mathcal{B}) = -\int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t$, and the third equality holds because $\boldsymbol{p}^\star(\boldsymbol{u})_k = (\nabla_{\boldsymbol{u}}\Phi(\boldsymbol{u};\mathcal{B}))_k$ by Lemma 3.1. Finally, the inequality holds because F_k's satisfy (6). This observation concludes our proof.

□

Proof of Theorem 4.2. Note that $\mathcal{B}$ encompasses distributions of zero mean, and thus any $\mathbb{Q} \in \mathcal{B}$ satisfies $\mathbb{E}_{\boldsymbol{z}\sim\mathbb{Q}}[\boldsymbol{z}] = \boldsymbol{0}$. Then, for any $\boldsymbol{u} \in \mathbb{R}^K$, we have

$$\max_{k\in[K]} u_k = \max_{k\in[K]} \mathbb{E}_{\boldsymbol{z}\sim\mathbb{Q}}\left[u_k + z_k\right] \le \mathbb{E}_{\boldsymbol{z}\sim\mathbb{Q}}\left[\max_{k\in[K]} u_k + z_k\right] \le \Phi(\boldsymbol{u}_T;\mathcal{B}),$$

where the first inequality follows by Jensen's inequality. The above inequality implies

$$\mathbb{E}\left[\max_{k\in[K]} \hat{u}_{T,k} - \Phi(\hat{\boldsymbol{u}}_T;\mathcal{B})\right] \le 0.$$

Therefore, the second term in the regret bound in (9) is upper bounded by 0.

As for the third term in (9), we have

$$\begin{aligned}&\sum_{k=1,p_{t,k}>0}^K p_{t,k}\int_0^{\|\hat{\boldsymbol{r}}_t\|}\int_0^x F_k'(F_k^{-1}(1-\boldsymbol{p}^\star(\hat{\boldsymbol{u}}_{t-1} - s\boldsymbol{e}_{a_t})_k))\ \mathrm{d}s\ \mathrm{d}x\\ &\le B\sum_{k=1,p_{t,k}>0}^K p_{t,k}\int_0^{\|\hat{\boldsymbol{r}}_t\|}\int_0^x (\boldsymbol{p}^\star(\hat{\boldsymbol{u}}_{t-1} - s\boldsymbol{e}_{a_t})_k)^\gamma \mathrm{d}s\ \mathrm{d}x\\ &\le B\sum_{k=1,p_{t,k}>0}^K p_{t,k}\int_0^{\|\hat{\boldsymbol{r}}_t\|}\int_0^x \boldsymbol{p}^\star(\hat{\boldsymbol{u}}_{t-1})_k^\gamma \mathrm{d}s\ \mathrm{d}x\\ &= B\sum_{k=1,p_{t,k}>0}^K {p_{t,k}}^{1+\gamma}\int_0^{\|\hat{\boldsymbol{r}}_t\|}\int_0^x \mathrm{d}s\ \mathrm{d}x = \frac{B}{2}\sum_{k=1}^K {p_{t,k}}^{\gamma-1} r_{t,a_t}^2,\end{aligned}$$

where the first inequality follows as $\Phi(\cdot;\mathcal{B})$ is (γ, B)-differentially consistent and the second inequality follows because $\boldsymbol{p}_k^\star(\boldsymbol{u})$ is non-decreasing and $s \geq 0$.

Note that $1/(2-\gamma)$-norm and $1/(\gamma-1)$-norm are duals for $\gamma \in (1,2)$. Then, Hölder's inequality yields

$$\sum_{k=1}^{K} {p_{t,k}}^{\gamma-1} = \sum_{k=1}^{K} {p_{t,k}}^{\gamma-1} \cdot 1 \leq \left(\sum_{k=1}^{K} {p_{t,k}}^{\frac{\gamma-1}{\gamma-1}}\right)^{\gamma-1} \left(\sum_{k=1}^{K} 1^{\frac{1}{2-\gamma}}\right)^{2-\gamma} = (1)^{\gamma-1} K^{2-\gamma} = K^{2-\gamma}.$$

This observation together with the assumption that $r_{t,a_t}^2 \in [0,1]$ completes our proof. $\square$

Proof of Theorem 5.1. By Lemma 3.1, $\Phi(\boldsymbol{u};\mathcal{B})$ is equivalent to

$$\Phi(\boldsymbol{u};\mathcal{B}) = \max_{\boldsymbol{p}\in\Delta^K} \sum_{k=1}^{K} u_k p_k + \sum_{k=1}^{K} \int_{1-p_k}^{1} F_k^{-1}(t)\mathrm{d}t$$

As F is strictly increasing, we have $F_k^{-1}(s) = -F^{-1}(1-s)\eta_k$ for all $s \in (0,1)$. Thus, we find

$$f(s) = \int_0^s F^{-1}(t)\mathrm{d}t = -\int_1^{1-s} F^{-1}(1-x)\,\mathrm{d}x = -\frac{1}{\eta_k}\int_{1-s}^{1} F_k^{-1}(x)\mathrm{d}x,$$

where the second equality follows from the variable substitution $x \leftarrow 1-t$. This integral representation of $f(s)$ then allows us to reformulate $\Phi(\boldsymbol{u};\mathcal{B})$ as

$$\Phi(\boldsymbol{u};\mathcal{B}) = \max_{\boldsymbol{p}\in\Delta^K} \sum_{k=1}^{K} u_k p_k - \sum_{k=1}^{K} \eta_k\, f(p_k).$$

This concludes our proof. $\square$

Proof of Corollary 5.2. By Lemma A.3 and Lemma A.4, we have

$$\mathcal{R}(T) \leq -\eta \sum_{k=1}^{K} f(p_{0,k}) + \frac{BTK^{2-\gamma}}{2\eta}, \tag{11}$$

where $\boldsymbol{p}_0 = \nabla_{\boldsymbol{u}}\Phi(\boldsymbol{u};\mathcal{B})|_{\boldsymbol{u}=\boldsymbol{0}}$ by Line 4 of Algorithm 1. We now show that $\boldsymbol{p}_0 = [1/K,\ldots,1/K]$, which coincides with the unique optimizer of

$$\max_{\boldsymbol{p}\in\Delta^K} -\eta \sum_{k=1}^{K} f(p_k)$$

Denote by $H(\boldsymbol{p}) = -\sum_{k=1}^{K} f(p_k)$ and $\Pi(\boldsymbol{p}_0)$ the set of all permuted copies of $\boldsymbol{p}_0$. Take K elements $\{\boldsymbol{x}^{(i)}\}_{i=1}^{K} \subseteq \Pi(\boldsymbol{p}_0)$ by cyclic permutation, *i.e.,* let $\boldsymbol{x}^{(1)} = \boldsymbol{p}_0$, $\boldsymbol{x}^{(K)} = [p_{0,K}, p_{0,1}, \ldots, p_{0,K-1}]$, and $\boldsymbol{x}^{(i)} = [p_{0,i}, \ldots, p_{0,K}, p_{0,1}, \ldots, p_{0,i-1}] \in \mathbb{R}^K$ for $i = 2, \ldots, K-1$. Observe as well that $\boldsymbol{p}' = [1/K, \ldots, 1/K] \in \Delta^K$ can be represented as $\boldsymbol{p}' = \sum_{i=1}^{K} \boldsymbol{x}^{(i)}/K$. We then have

$$-Kf\left(\frac{1}{K}\right) = H(\boldsymbol{p}') = H\left(\frac{1}{K}\sum_{i=1}^{K} \boldsymbol{x}^{(i)}\right) \geq \frac{1}{K}\sum_{i=1}^{K} H(\boldsymbol{x}^{(i)}) = -\sum_{k=1}^{K} f(p_{0,k}) = \max_{\boldsymbol{p}\in\Delta^K} -\sum_{k=1}^{K} f(p_k),$$

where the inequality is due to Jensen, and the third equality holds because H is a permutation invariant function, *i.e.*, $H(\boldsymbol{x}^{(i)}) = -\sum_{k=1}^{K} f(x_k^{(i)}) = -\sum_{k=1}^{K} f(p_{0,k})$ for any $\boldsymbol{x}^{(i)} \in \Pi(\boldsymbol{p}_0)$. By Corollary 3.1, (11) admits a unique maximizer $\boldsymbol{p}_0$. As $\boldsymbol{p}' \in \Delta^K$ is feasible in (11), the upper bound above is in fact tight and $\boldsymbol{p}_0 = \boldsymbol{p}'$. This observation concludes our proof. □

Proof of Theorem 5.3. We denote $\eta = \eta_k$ within this proof. Thanks to Theorem 5.1 we have $\Phi(\boldsymbol{u}; \mathcal{B}) = \Phi^R(\boldsymbol{u}; \eta\psi_\alpha^{\mathbb{T}})$ for all $\boldsymbol{u} \in \mathbb{R}^K$. The claim follows as GBPA$(\Phi^R(\cdot; \eta\psi_\alpha^{\mathbb{T}}))$ has $\mathbb{R}(T) \leq \sqrt{KT/(\alpha(1-\alpha))}$ by [Abernethy et al., 2015, Theorem 3.1]. □

Proof of Theorem 5.4. Thanks to Theorem 5.1 for any $\alpha \in (0,1)$, we have $\Phi(\boldsymbol{u}; \mathcal{B}_t) = \Phi^R(\boldsymbol{u}; \eta_t\psi_{1/2}^{\mathbb{T}})$ for all $\boldsymbol{u} \in \mathbb{R}^K$. The first part of the claim then follows because when $\eta_t = 2\sqrt{t}$, the regret of GBPA$(\Phi^R(\cdot; \eta_t\psi_{1/2}^{\mathbb{T}}))$ in the *adversarial* setting is upper bounded by $4\sqrt{KT_0}+1$ thanks to [Zimmert and Seldin, 2021, Theorem 1]. By the refined analysis of Tsallis-INF algorithm in [Ito, 2021, Theorem 2], we further have that $\mathcal{R}(T_0) \leq \mathcal{O}(\sum_{k\in[K]:\Delta_k>0} \log(T_0)/\Delta_k)$ if adversary satisfies (1). □

Proof of Corollary 5.5. By Lemma 3.1, $\Phi(\boldsymbol{u}; \mathcal{B})$ is equivalent to

$$\Phi(\boldsymbol{u}; \mathcal{B}) = \max_{\boldsymbol{p}\in\Delta^K} \sum_{k=1}^{K} u_k p_k + \sum_{k=1}^{K} \int_{1-p_k}^{1} F_k^{-1}(t)\mathrm{d}t.$$

By construction for all $k \in [K]$, we have

$$\begin{aligned}\eta_k f(s) + \gamma_k g(s) &= \eta_k \int_0^s G_1^{-1}(t)\mathrm{d}t + \gamma_k \int_0^s G_2^{-1}(t)\mathrm{d}t \\ &= -\int_1^{1-s} (\eta_k G_1^{-1} + \gamma_k G_2^{-1})(1-x)\mathrm{d}x = -\int_{1-s}^{1} F_k^{-1}(x)\mathrm{d}x,\end{aligned}$$

where the second equality follows from the variable substitution $x \leftarrow 1-t$ and last equality follows by construction of F_k. The final representation of $f_k(s)$ above then allows us to reformulate $\Phi(\boldsymbol{u}; \mathcal{B})$ as

$$\Phi(\boldsymbol{u}; \mathcal{B}) = \max_{\boldsymbol{p}\in\Delta^K} \sum_{k=1}^{K} u_k p_k - \sum_{k=1}^{K} (\eta_k f(p_k) + \gamma_k g(p_k)).$$

This observation concludes our proof. □

Proof of Corollary 5.6. Note first that by Corollary 5.5, for every $t \in [T]$ we have $\Phi(\boldsymbol{u}; \mathcal{B}_t) = \Phi^R(\boldsymbol{u}; \psi_t)$, where $\psi(\boldsymbol{p}) = \sum_{k=1}^{K} (\eta_{t,k} f(p_k) + \gamma_{t,k} g(p_k))$, $f(s) = \int_0^s G_1^{-1}(t)\mathrm{d}t$, and $g(s) = \int_0^s G_2^{-1}(t)\mathrm{d}t$. Moreover, thanks to our choice of G_1 and G_2, we have that $\eta_{t,k} f(s) + \gamma_{t,k} g(s) = -\eta_{t,k}(\sqrt{s} + (s-1)\log(1-s))$. The claim then follows from [Zimmert et al., 2019, Theorem 3]. □