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

Mengmeng Li* Daniel Kuhn† Bahar Taşkesen‡

February 13, 2025

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. However, 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. Similar to 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 in which a learner interacts with an environment over T rounds. In each round, the learner selects one of K arms and then observes and receives an uncertain reward associated with the chosen arm. The learner's objective is to minimize regret, which we define as the absolute difference between the total expected reward obtained and the total expected reward that could have been achieved with perfect knowledge of the reward distribution. In the stochastic setting, where the rewards in each round are drawn independently from an unknown but fixed distribution, the *Upper Confidence Bound* algorithm [Auer et al., 2002a] as well as

*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

the *Thompson Sampling* algorithm [Thompson, 1933] achieve the optimal $\mathcal{O}(\log T)$ regret [Bubeck and Cesa-Bianchi, 2012]. In the adversarial setting, where rewards are strategically chosen by a malicious adversary, however, these methods suffer from linear regret [Zimmert and Seldin, 2021]. In contrast, the Follow-the-Regularized-Leader (FTRL) algorithm by Gordon [1999], which uses the iterates of a gradient descent-type algorithm as arm-sampling distributions, often achieves the optimal $\mathcal{O}(\sqrt{KT})$ regret in the adversarial setting [Bubeck and Cesa-Bianchi, 2012].

Prior knowledge about the nature of the environment is typically unavailable. Therefore, an algorithm that achieves optimal regret in both stochastic and adversarial settings simultaneously is highly desirable. Recently, Zimmert and Seldin [2021] proved that an FTRL algorithm with a Tsallis entropy regularizer can *simultaneously* achieve the optimal $\mathcal{O}(\log T)$ regret in the stochastic setting as well as the optimal $\mathcal{O}(\sqrt{KT})$ regret in the adversarial setting, without requiring parameter tuning. Algorithms of this type are often said to exhibit the "best-of-both-worlds" (BOBW) property [Bubeck and Slivkins, 2012]. The results by Zimmert and Seldin [2021] have been extended in various directions, aiming to identify the key properties of regularizers that induce the BOBW property [Jin et al., 2024]. However, FTRL algorithms require solving a computationally expensive optimization problem in each round to determine the arm-sampling distribution.

Follow-the-Perturbed-Leader (FTPL) algorithms [Hannan, 1957] select an arm with a maximal perturbed reward estimate, where the perturbation is sampled from a prescribed noise distribution. These algorithms are widely favored for their superior computational efficiency compared to FTRL approaches [Abernethy et al., 2014; Lattimore and Szepesvári, 2020]. Recently, it was shown that FTPL with Fréchet perturbations possesses the BOBW property [Honda et al., 2023]. However, this analysis heavily relies on the specifics of a Fréchet distribution with a particular shape parameter. While this paper was under review, a more systematic analysis of FTPL algorithms was provided by Lee et al. [2024], further highlighting the effectiveness of FTPL methods.

It is well known that any FTPL policy can be expressed as an FTRL policy [Abernethy et al., 2017; Hofbauer and Sandholm, 2002]. However, the reverse does not hold in general [Hofbauer and Sandholm, 2002, Proposition2.2]. Establishing a one-to-one correspondence between meaningful subclasses of FTPL and FTRL policies remains an open problem [Abernethy et al., 2017]. Although all FTRL and FTPL methods can be viewed as instances of a Gradient-Based Prediction Algorithm (GBPA)[Abernethy et al., 2015], their regret analyses require separate techniques. Furthermore, an open question posed by Kim and Tewari [2019] asks whether there exists a noise distribution such that the corresponding FTPL policy exactly matches the FTRL policy with the Tsallis entropy regularizer. Kim and Tewari [2019] also proved that FTPL with independent and identically distributed (i.i.d.) noise across the arms cannot recover Tsallis-entropy-regularized FTRL. Designing perturbations that precisely replicate the FTRL algorithm with BOBW capability is essential for understanding the role of regularization through perturbation. Resolving this open question would also facilitate the unification of FTRL and FTPL regret analysis.

Contributions. In this paper, we bridge the gap between FTRL and FTPL methods by studying *ambiguous* noise distributions that allow for *correlations* across the arms. Additionally, we introduce a new "optimism in the face of ambiguity" principle, whereby the perturbations in FTPL are sampled from the most advantageous noise distribution within a prescribed ambiguity set. This contrasts sharply with standard FTPL algorithms, which rely on a single fixed noise distribution. We derive explicit formulas for this most advantageous noise distribution, thus resolving the open problem posed by Kim and Tewari [2019]. Leveraging techniques from discrete choice theory [Natarajan et al., 2009; Feng et al., 2017]traditionally studied in economics and psychologywe show that the arm-sampling probabilities under the optimal noise distribution can be computed highly efficiently using bisection. Unlike standard FTRL algorithms, which require solving an expensive optimization problem in every round, our approach is significantly more computationally efficient and its runtime remains comparable to that of FTPL, up to logarithmic factors. As a result, our algorithm combines the unified regret analysis of FTRL with the computational efficiency of FTPL. Moreover, it encompasses a broad class of FTRL methods as special cases, including several optimal ones, such as those based on Tsallis entropy and hybrid regularizers. Notably, while unifying these FTRL methods within the traditional FTPL framework was previously considered infeasible, our approach successfully achieves this integration.

Related work. In this paper, we relax the assumption of i.i.d. arm perturbations, thereby generalizing traditional FTPL methods. The i.i.d. assumption, which underlies most FTPL algorithms, is also relaxed in [Melo and Müller, 2023] by interpreting the arm-sampling probabilities as choice probabilities in a nested logit model, a concept commonly studied in discrete choice theory. In this work, however, the noise must follow a generalized extreme-value distribution, and the resulting algorithm does not achieve BOBW regret bounds. In contrast, we work with a whole family of distributions and leverage ideas from discrete choice theory and distributionally robust optimization to develop an efficient bisection algorithm for computing arm-sampling probabilities under the most advantageous noise distribution. This general framework encompasses several algorithms that achieve BOBW regret bounds. The FTPL algorithm with i.i.d. Fréchet-distributed noise is also known to exhibit BOBW capabilities [Honda et al., 2023], but its regret analysis is specifically tailored to the Fréchet distribution. While this method extends to some other noise distributions [Lee et al., 2024], the underlying regret analysis remains complex. In contrast, our perturbation-based algorithm achieves BOBW regret bounds by leveraging its exact equivalence with FTRL algorithms that possess the BOBW property, leading to a more unified and efficient approach.

Notation. We denote by $[K] = \{1, \ldots, K\}$ 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 $i \in [K]$ to denote the i-th standard basis vector of $\mathbb{R}^K$. The Bregman divergence function induced by a differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$ is defined through $\mathbb{D}_\phi(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \boldsymbol{x} - \boldsymbol{y}, \nabla\phi(\boldsymbol{y})\rangle$.

2 Multi-Armed Bandits

We study a multi-armed bandit (MAB) problem running over $T \in \mathbb{N}$ rounds. In each round the learner must select one of $K \in \mathbb{N}$ arms and earns a random reward that depends on the chosen arm. More precisely, in round $t \in [T]$, the learner selects an arm $a_t \in [K]$, the environment generates a reward vector $\boldsymbol{r}_t = (r_{t,1}, r_{t,2}, \ldots, r_{t,K}) \in [-1, 0]^K$, and the learner receives the reward r_{t,a_t} associated with arm a_t. The crux of MAB problems is that the learner observes only the reward r_{t,a_t} associated with the chosen arm but receives no information about the rewards $r_{t,k}$ of the other arms $k \neq a_t$. Accordingly, we assume throughout the paper that the arm a_t is sampled from a distribution over $[K]$ chosen by the learner that may depend on the history $(a_1, \ldots, a_{t-1})$ of the chosen arms and the history $(r_{1,a_1}, \ldots, r_{t-1,a_{t-1}})$ of the corresponding rewards. Similarly, the reward vector $\boldsymbol{r}_t$ is sampled from a distribution over $[-1, 0]^K$ unknown to the learner. We distinguish two main reward generation regimes. In the (non-oblivious) adversarial regime, the reward distribution may depend on the history $(a_1, \ldots, a_{t-1})$ of the chosen arms as well as the history $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{t-1})$ of the rewards. In the stochastic regime, on the other hand, the reward distribution is kept fixed, and the rewards are sampled independently from this distribution. There are also intermediate reward generation regimes under which the environment has varying levels of adversarial power.

The learner's objective is to minimize the regret

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

which measures the difference between the expected cumulative reward of the best arm under full distributional information and the learner's expected cumulative reward. Here, the expectations are taken with respect to the arm-sampling distributions chosen by the learner and the reward distributions chosen by the environment. We highlight that $\mathcal{R}(T)$ is sometimes termed *pseudo-regret* [Zimmert and Seldin, 2021]. As we do not distinguish different notions of regret in this paper, we simply refer to $\mathcal{R}(T)$ as the *regret* to keep terminology simple. We also emphasize that $\mathcal{R}(T)$ may be negative for certain arm-sampling and reward distributions. Indeed, if the reward distribution changes over time, it may be strictly suboptimal to select the same arm in each round. However, the *worst-case* regret, which is obtained by maximizing $\mathcal{R}(T)$ over all admissible reward distributions, is nonnegative even in the stochastic regime, where the environment has only minimal adversarial power. Indeed, if the reward of each arm follows a fixed Bernoulli distribution independent of t, then the regret is already lower bounded by $\mathcal{O}(\sqrt{KT})$ [Auer et al., 1995].

Algorithm 1 Gradient-based prediction algorithm (GBPA)

Require: Differentiable convex function ϕ with $\nabla_{\boldsymbol{u}} \phi(\boldsymbol{u}) \in \Delta^K$
  Initialize $\hat{\boldsymbol{u}}_0 = \mathbf{0}$
  for round $t = 1, \ldots, T$ **do**
    Environment chooses a reward vector $\boldsymbol{r}_t \in [-1, 0]^K$
    Learner chooses $a_t \sim \boldsymbol{p}_t = \nabla_{\boldsymbol{u}} \phi(\boldsymbol{u})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{t-1}}$ and receives reward r_{t,a_t}
    Learner estimates single-round reward vector $\hat{\boldsymbol{r}}_t = (r_{t,a_t}/p_{t,a_t})\boldsymbol{e}_{a_t}$
    Update $\hat{\boldsymbol{u}}_t \leftarrow \hat{\boldsymbol{u}}_{t-1} + \hat{\boldsymbol{r}}_t$
  end for

In this paper, we use different variants of a gradient-based prediction algorithm (GBPA) [Abernethy et al., 2012, 2014, 2015; Kim and Tewari, 2019] to select the arms; see Algorithm 1. GBPA recursively constructs a statistic $\hat{\boldsymbol{u}}_{t-1} \in \mathbb{R}^K$ whose k-th component estimates the expected cumulative reward achievable by pulling arm k in each of the rounds $1, \ldots, t-1$. In round $t \in [T]$, GBPA 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 $\boldsymbol{p}_t$ and samples an arm a_t from $\boldsymbol{p}_t$. Next, GBPA updates $\hat{\boldsymbol{u}}_{t-1}$ by adding the single-round reward estimate $\hat{\boldsymbol{r}}_t = (r_{t,a_t}/p_{t,a_t})\boldsymbol{e}_{a_t}$. One readily verifies that if $\boldsymbol{p}_t > \boldsymbol{0}$, then $\hat{\boldsymbol{r}}_t$ constitutes an unbiased estimator for $\mathbb{E}[\boldsymbol{r}_t]$. GBPA unifies several MAB algorithms, including those described in [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 with full information (where the learner observes the full reward vector $\boldsymbol{r}_t$ and not only the reward of the chosen arm). These algorithms differ primarily in the choice of the convex potential function ϕ used as an input to GBPA. Below we discuss the policies corresponding to different choices of ϕ.

Follow-the-leader (FTL). GBPA with $\phi(\boldsymbol{u}) = \max_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top \boldsymbol{u}$ is known as the FTL algorithm. In this case we have $\nabla_{\boldsymbol{u}}\phi(\boldsymbol{u}) \in \operatorname{argmax}_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top \boldsymbol{u}$ by Danskin's theorem [Bertsekas, 2016, Proposition B.25], that is, the learner simply chooses an arm with maximal cumulative reward estimate.[1] While FTL is easy to implement, it is well known that the regret of FTL can grow linearly with T even if there are only $K = 2$ arms. For example, if the adversary chooses the reward vectors $\boldsymbol{r}_1 = \{-1/2, 0\}$, $\boldsymbol{r}_t = \{-1, 0\}$ when $t > 1$ is odd and $\boldsymbol{r}_t = \{0, -1\}$ when t is even, then one can show that FTL selects arm 1 whenever t is odd and arm 2 whenever t is even. Thus, the cumulative reward at time T is at most $(-T+1)/2$, and the regret is at least $T/2 - 1$ [Hazan, 2016, Chapter 5].

Follow-the-regularized-leader (FTRL). A popular approach to stabilize the FTL algorithm is to add a convex regularization function $\psi : \mathbb{R}^K \to \mathbb{R}$ to the linear objective function $\boldsymbol{p}^\top \boldsymbol{u}$. In this case, the learner generically constructs a non-degenerate arm-sampling distribution by solving $\max_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top \boldsymbol{u} - \psi(\boldsymbol{p})$. GBPA with potential function $\Phi^R(\boldsymbol{u};\psi) = \max_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top \boldsymbol{u} - \psi(\boldsymbol{p})$ is known as the FTRL algorithm. In the adversarial regime, the FTRL algorithm achieves the minimax optimal regret of $\mathcal{O}(\sqrt{KT})$ if $\psi(\boldsymbol{p}) = \eta\psi_\alpha^{\mathbb{T}}(\boldsymbol{p})$, where $\eta = \sqrt{T(1-\alpha)/(2\alpha)}$ is the learning rate and

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

is the Tsallis entropy with parameter $\alpha \in (0,1)$ [Abernethy et al., 2015, Corollary 3.2]. Similarly, in the stochastic regime, the FTRL algorithm achieves the optimal regret of $\mathcal{O}(\log T)$ if the potential function $\psi(\boldsymbol{p}) = \eta_t\psi_\alpha^{\mathbb{T}}(\boldsymbol{p})$ scales with a time-dependent learning rate $\eta_t = 2\sqrt{t}$ [Ito, 2021, Theorem 2]. If FTRL is equipped with a hybrid regularizer that combines the Shannon entropy with the Tsallis entropy, then it enjoys a BOBW capability, that is, it can be shown to achieve optimal regret both in the adversarial as well as in the stochastic bandit regime [Zimmert et al., 2019]. Thus, the FTRL algorithm comes with strong statistical guarantees. On the flipside, however, it is computationally expensive because it requires the solution of a different convex optimization problem in each round.

[1] In this informal discussion we disregard technical complications arising when the maximizer $\boldsymbol{p}$ is not unique for a given reward estimate $\boldsymbol{u}$, in which case the potential function $\phi(\boldsymbol{u})$ fails to be differentiable at $\boldsymbol{u}$.

Follow-the-perturbed-leader (FTPL). As an alternative to the computationally expensive FTRL method, it has been proposed to inject stochastic noise $\boldsymbol{z} \sim \mathbb{Q} \in \mathcal{P}(\mathbb{R}^K)$ into the cumulative reward estimate $\boldsymbol{u}$ and to sample arms from $\boldsymbol{p} = \mathbb{E}_{\boldsymbol{z}\sim\mathbb{Q}}[\boldsymbol{e}_{k^\star(\boldsymbol{z})}]$, where $k^\star(\boldsymbol{z}) \in \operatorname{argmax}_{k\in[K]} u_k + z_k$. Using the dominated convergence theorem in conjunction with Danskin's theorem [Bertsekas, 2016, Proposition B.25], one can show that the arm-sampling distribution $\boldsymbol{p}$ coincides with the gradient of the potential function $\Phi^P(\boldsymbol{u};\mathbb{Q}) = \mathbb{E}_{\boldsymbol{z}\sim\mathbb{Q}}[\max_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top(\boldsymbol{u}+\boldsymbol{z})]$. GBPA with $\phi(\boldsymbol{u}) = \Phi^P(\boldsymbol{u};\mathbb{Q})$ is known as the FTPL algorithm. Existing FTPL algorithms assume that the disturbances associated with different arms (that is, the components of $\boldsymbol{z}$) are mutually independent under $\mathbb{Q}$.

FTPL algorithms are computationally efficient because they simply select the arm with the maximum perturbed reward and because this arm can be identified by searching. This is significantly cheaper than solving a convex optimization problem. However, due to their stochastic nature, the analysis of FTPL algorithms is more cumbersome compared to the straightforward and mature analysis of FTRL algorithms. Even though FTPL algorithms have been shown to enjoy a BOBW capability [Honda et al., 2023; Lee et al., 2024], it is unclear whether there exists an algorithm that is as efficient as an FTPL method yet admits a streamlined analysis like an FTRL algorithm.

A promising approach to achieving this goal is to establish a correspondence between FTRL and FTPL algorithms. It is well known that essentially any FTPL algorithm can be represented as an FTRL algorithm [Hofbauer and Sandholm, 2002, Theorem 2.1]. The reverse problem of framing a given FTRL algorithm as an FTPL algorithm, however, is perceived as challenging [Abernethy et al., 2017; Honda et al., 2023]. Accordingly, finding a bridge between regularization- and perturbation-based algorithms constitutes indeed an unresolved open problem.

Open Problem: *Given a convex regularization function $\psi : \mathbb{R}^K \to \mathbb{R}$ and a reward estimate $\boldsymbol{u}$, construct a perturbation distribution $\mathbb{Q}$ on $\mathbb{R}^K$ such that $\nabla_{\boldsymbol{u}}\Phi^P(\boldsymbol{u};\mathbb{Q}) = \nabla_{\boldsymbol{u}}\Phi^R(\boldsymbol{u};\psi)$.*

Since FTRL with Tsallis entropy regularizer achieves the minimax optimal regret rate in adversarial bandits, a simpler but still interesting version of the above open problem is to seek an FTPL algorithm with the same arm-sampling distribution as the special instance of the FTRL algorithm with Tsallis entropy regularizer. But to date, even for this simpler problem, only a negative result is available. Kim and Tewari [2019, Theorem 8] show that FTRL with Tsallis entropy regularizer cannot be recovered by any FTPL algorithm with mutually independent disturbances z_k, $k \in [K]$. Another negative result is due to Hofbauer and Sandholm [2002, Proposition 2.2], who identify generalized FTRL algorithms that correspond to *extended real-valued* regularization functions and that cannot be matched by any FTPL algorithm. However, we assume here that ψ is real-valued.

In the next section, we describe a general framework for mapping regularization functions in FTRL to disturbance distributions in FTPL, thus providing a systematic solution to the open problem mentioned above. To circumvent the impossibility result by Kim and Tewari [2019], we will study *ambiguous* noise-sampling distributions that allow for *correlations* across the arms.

3 Distributionally Optimistic Perturbations

We now introduce a new class of smooth potential functions that can be viewed as best-case expected utilities of the type studied in semi-parametric discrete choice theory. That is, we define

$$\Phi(\boldsymbol{u};\mathcal{B}) = \sup_{\mathbb{Q}\in\mathcal{B}} \mathbb{E}_{\boldsymbol{z}\sim\mathbb{Q}}\left[\max_{k\in[K]}(u_k+z_k)\right], \tag{2}$$

where $\boldsymbol{z}$ represents a random vector of perturbations governed by a distribution $\mathbb{Q}$ 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 distribution at the origin of $\mathbb{R}^K$, then Algorithm 1 with potential function $\phi(\boldsymbol{u})=\Phi(\boldsymbol{u};\mathcal{B})$ reduces to FTL. In addition, $\Phi(\boldsymbol{u};\{\mathbb{Q}\})$ trivially coincides with $\Phi^P(\boldsymbol{u};\mathbb{Q})$. Hence, GBPA with potential function $\phi(\boldsymbol{u})=\Phi(\boldsymbol{u};\mathcal{B})$ generalizes traditional FTPL, which injects i.i.d. noise into the cumulative reward estimates.

The family of GBPA algorithms with potential function $\phi(\boldsymbol{u})=\Phi(\boldsymbol{u};\mathcal{B})$ also includes the Exp3 algorithm by Auer et al. [1995], which is arguably one of the most popular FTPL algorithms.

Remark 1 (Exp3 algorithm)**.** *If $\mathcal{B}=\{\mathbb{Q}\}$ is a singleton with $\mathbb{Q}=\otimes_{k=1}^K\mathbb{Q}_k$ and if $\mathbb{Q}_k\in\mathcal{P}(\mathbb{R})$ is a Gumbel distribution with zero mean and variance $\pi^2\eta^2/6$ for some $\eta>0$, then one can show that $\Phi(\boldsymbol{u};\mathcal{B})=\eta\log(\sum_{k=1}^K\exp(u_k/\eta))$. In this case, the arm-sampling probabilities are available in closed form and are equivalent to the choice probabilities in the celebrated multinomial logit model in discrete choice theory, that is, $p_k(\boldsymbol{u})=(\nabla_{\boldsymbol{u}}\Phi(\boldsymbol{u},\mathcal{B}))_k=\exp(u_k/\eta)/(\sum_{j=1}^K\exp(u_j/\eta))$, see [McFadden, 1981, Theorem 5.2]. This reveals that GBPA with potential function $\phi(\boldsymbol{u})=\Phi(\boldsymbol{u};\mathcal{B})$ reduces indeed to the celebrated Exp3 algorithm by Auer et al. [1995].*

From now on we focus on *marginal* ambiguity sets, which specify the marginal distributions of the components of $\boldsymbol{z}$ but do not impose any constraints on their dependence structure.

Definition 1 (Marginal ambiguity set)**.** *The marginal ambiguity set induced by K cumulative distribution functions $F_k:\mathbb{R}\to[0,1]$, $k\in[K]$, is given by*

$$\mathcal{B}=\left\{\mathbb{Q}\in\mathcal{P}(\mathbb{R}^K):\ \mathbb{Q}[z_k\le s]=F_k(s)\ \ \forall s\in\mathbb{R},\ \forall k\in[K]\right\}. \tag{3}$$

We henceforth refer to GBPA with potential function $\phi(\boldsymbol{u})=\Phi(\boldsymbol{u};\mathcal{B})$ induced by a marginal ambiguity set $\mathcal{B}$ as the *distributionally optimistic perturbation algorithm* (DOPA). DOPA establishes a bridge between many commonly used FTRL and FTPL methods. To see this, we first recall an important property of marginal ambiguity sets, which was initially discovered in the context of semi-parametric discrete choice theory. Below we denote by $F_k^{-1}:[0,1]\to\mathbb{R}$ the (left) quantile function corresponding to the cumulative distribution function F_k. It is defined via

$$F_k^{-1}(s)=\inf\{t:F_k(t)\ge s\}\quad \forall s\in\mathbb{R}.$$

Lemma 3.1. *[Natarajan et al., 2009, Theorem 1] If $\mathcal{B}$ is a marginal ambiguity set of the form (3) and if the cumulative distribution functions $F_k, k\in[K]$, are continuous and strictly increasing in s whenever $F_k(s)\in(0,1)$, then the potential function (2) is convex and differentiable in $\boldsymbol{u}$ and satisfies*

$$\Phi(\boldsymbol{u};\mathcal{B})=\max_{\boldsymbol{p}\in\Delta^K}\sum_{k=1}^K u_kp_k+\sum_{k=1}^K\int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t\quad\forall\boldsymbol{u}\in\mathbb{R}^K. \tag{4}$$

In addition, the unique maximizer of the convex program (4) is given by $\boldsymbol{p}(\boldsymbol{u})=\nabla_{\boldsymbol{u}}\Phi(\boldsymbol{u};\mathcal{B})$.

Lemma 3.1 reveals that if $\mathcal{B}$ is any marginal ambiguity set, then $\Phi(\boldsymbol{u};\mathcal{B})$ can be expressed as the optimal value of a convex maximization problem over the probability simplex. Besides its relevance for semi-parametric discrete choice theory, Lemma 3.1 also has interesting ramifications for semi-discrete optimal transport [Taşkesen et al., 2023, Proposition 3.6]. Note that the objective function of the convex program in (4) represents a sum of K strictly concave and differentiable functions $\varphi_k(p_k) = u_k p_k + \int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t$ on $(0,1)$, $k \in [K]$, provided that $F_k(s)$ is strictly increasing in s whenever $F_k(s) \in (0,1)$. Indeed, the derivative of φ_k satisfies $\varphi_k'(p_k) = u_k + F_k^{-1}(1-p_k)$, which is strictly decreasing in p_k because $F_k(s)$ is strictly increasing in s and $1-p_k$ is strictly decreasing in p_k. Moreover, if $F_k(s)$ is strictly increasing at *every* $s \in \mathbb{R}$, then $\lim_{p_k \to 0} \varphi_k'(p_k) = \infty$. In this case, the optimal arm-sampling distribution $\boldsymbol{p}(\boldsymbol{u}) = \nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u};\mathcal{B})$ that solves (4) satisfies $\boldsymbol{p}(\boldsymbol{u}) > \mathbf{0}$.

Lemma 3.1 implies that DOPA can be viewed as an FTRL algorithm with convex regularization function $\psi(\boldsymbol{p}) = -\sum_{k=1}^K \int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t$. Conversely, the following proposition shows that FTRL algorithms with separable regularization functions can also be interpreted as instances of DOPA.

Proposition 3.2 (FTRL vs. DOPA)**.** *Define $\psi : [0,1]^K \to \mathbb{R}$ through $\psi(\boldsymbol{p}) = \sum_{k=1}^K \psi_k(p_k)$, where $\psi_k : [0,1] \to \mathbb{R}$ is strictly convex and differentiable for every $k \in [K]$. If $\mathcal{B}$ is a marginal ambiguity set of the form (3) induced by cumulative distribution functions $F_k : \mathbb{R} \to [0,1]$ satisfying $F_k(s) = \min\{1, \max\{0, -(\psi_k')^{-1}(1-s)\}\}$ for all $s \in \mathbb{R}$ and $k \in [K]$, then $\nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u};\mathcal{B}) = \nabla_{\boldsymbol{u}} \Phi^R(\boldsymbol{u};\psi)$.*

Proof. We may assume without loss of generality that $\psi_k(0) = 0$. Otherwise, we can simply shift ψ_k by $-\psi_k(0)$ without affecting $\nabla_{\boldsymbol{u}} \Phi^R(\boldsymbol{u};\psi)$. Note also that ψ_k' is strictly increasing because ψ_k is strictly convex. Hence, the function $G_k : [0,1] \to \mathbb{R}$ with $G_k(s) = -\psi_k'(1-s)$ is also strictly increasing. The fundamental theorem of calculus thus implies that $-\int_{1-p_k}^1 G_k(t)\mathrm{d}t = \psi_k(p_k)$. By the defining properties of F_k and G_k, it is now clear that $F_k(s) = \min\{1, \max\{0, G_k^{-1}(s)\}\}$, which implies that $F_k^{-1}(s) = G_k(s)$ for all $s \in (0,1)$. Integrating both sides of this identity then yields $-\int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t = -\int_{1-p_k}^1 G_k(t)\mathrm{d}t = \psi_k(p_k)$. By Lemma 3.1, we may thus conclude that

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

The claim then follows by taking gradient with respect to $\boldsymbol{u}$ on both sides. □

Next, we show that there is also a close connection between FTPL and DOPA.

Proposition 3.3 (FTPL vs. DOPA)**.** *Suppose that $\mathcal{B}$ is a marginal ambiguity set of the form (3) and that the underlying cumulative distribution functions $F_k, k \in [K]$, are continuous and strictly increasing in s whenever $F_k(s) \in (0,1)$. Then, for every fixed $\boldsymbol{u} \in \mathbb{R}^K$ there exists $\mathbb{Q} \in \mathcal{P}(\mathbb{R}^K)$ that satisfies $\nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u};\mathcal{B}) = \nabla_{\boldsymbol{u}} \Phi^P(\boldsymbol{u};\mathbb{Q})$.*

Proof. Throughout the proof we use $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{u}) = \nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u};\mathcal{B})$ as shorthand for the unique solution of the convex program (4) at the fixed reward estimate $\boldsymbol{u}$. In addition, we use $\mathbb{Q}_k \in \mathcal{P}(\mathbb{R})$ to denote the unique probability distribution satisfying $\mathbb{Q}_k(z_k \le s) = F_k(s)$ for all $s \in \mathbb{R}$, and we define

$$\mathbb{Q} = \sum_{k=1}^K p_k \cdot \left(\otimes_{\ell=1}^{k-1} \mathbb{Q}_\ell^-\right) \otimes \mathbb{Q}_k^+ \otimes \left(\otimes_{\ell=k+1}^K \mathbb{Q}_\ell^-\right),$$

where the truncated distributions $\mathbb{Q}_k^+, \mathbb{Q}_k^- \in \mathcal{P}(\mathbb{R})$ are defined through

$$\mathbb{Q}_k^+(z_k \in B) = \mathbb{Q}_k\left(z_k \in B \,\middle|\, z_k > F_k^{-1}(1-p_k)\right) \quad \text{and} \quad \mathbb{Q}_k^-(z_k \in B) = \mathbb{Q}_k\left(z_k \in B \,\middle|\, z_k \leq F_k^{-1}(1-p_k)\right)$$

for all Borel sets $B \subseteq \mathbb{R}$, respectively. From the proof of [Natarajan et al., 2009, Theorem 1] we know that $\mathbb{Q}$ solves the optimization problem in (2); see also [Taşkesen et al., 2023, Proposition 3.6] for an alternative proof using our notation. The optimality conditions of problem (4) further imply that

$$u_k + F_k^{-1}(1-p_k) = u_\ell + F_\ell^{-1}(1-p_\ell) \quad \forall k, \ell \in [K]. \tag{5}$$

Next, fix an arbitrary $k \in [K]$, and note that for every fixed $z_k > F_k^{-1}(1-p_k)$ and $\ell \neq k$ we have

$$\mathbb{Q}_\ell^-\left(z_\ell < z_k + u_k - u_\ell\right) \geq \mathbb{Q}_\ell^-\left(z_\ell \leq F_k^{-1}(1-p_k) + u_k - u_\ell\right) = \mathbb{Q}_\ell^-\left(z_\ell \leq F_\ell^{-1}(1-p_\ell)\right) = 1, \tag{6}$$

where the first equality follows from (5), and the second equality holds because $\mathbb{Q}_\ell^-$ is supported on the interval $(-\infty, F_\ell^{-1}(1-p_\ell)]$. Similarly, for any fixed $z_k \leq F_k^{-1}(1-p_k)$ and $\ell \neq k$ we have

$$\mathbb{Q}_\ell^+(z_\ell < z_k + u_k - u_\ell) \leq \mathbb{Q}_\ell^+\left(z_\ell < F_k^{-1}(1-p_k) + u_k - u_\ell\right) = \mathbb{Q}_\ell^+\left(z_\ell < F_\ell^{-1}(1-p_\ell)\right) = 0 \tag{7}$$

where the first equality follows from (5), and the second equality holds because $\mathbb{Q}_\ell^+$ is supported on the interval $(F_\ell^{-1}(1-p_\ell), \infty)$. For any fixed $k \in [K]$, we may thus conclude that

$$\begin{aligned}
\mathbb{Q}\left(k \in \operatorname*{argmax}_{\ell \in [K]} u_\ell + z_\ell\right) &= \mathbb{Q}(z_\ell < u_k + z_k - u_\ell \;\forall \ell \neq k) \\
&= p_k \mathbb{E}_{z_k \sim \mathbb{Q}_k^+}\left[\prod_{\ell \neq k} \mathbb{Q}_\ell^-\left(z_\ell < u_k + z_k - u_\ell\right)\right] \\
&\quad + \sum_{\ell \neq k} p_\ell \mathbb{E}_{z_k \sim \mathbb{Q}_k^-}\left[\mathbb{Q}_\ell^+(z_\ell < u_k + z_k - u_\ell) \prod_{j \neq k, \ell} \mathbb{Q}_j^-(z_j < u_k + z_k - u_j)\right] = p_k.
\end{aligned}$$

Here, the first equality follows from the assumption that the marginal distribution functions F_k, $k \in [K]$, are continuous. This implies that $\mathbb{Q}_\ell^+$ and $\mathbb{Q}_\ell^-$, $\ell \in [K]$, are absolutely continuous to the Lebesgue measure on $\mathbb{R}$, which in turn implies that $\mathbb{Q}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^K$. Hence, the event $z_\ell = u_k + z_k - u_\ell$ has zero probability under $\mathbb{Q}$. The second equality exploits the construction of $\mathbb{Q}$, and the third equality follows from (6) and (7).

Finally, the definition of FTPL potential function $\Phi^P(\boldsymbol{u}; \mathbb{Q})$ implies that

$$\begin{aligned}
\frac{\partial}{\partial u_k} \Phi^P(\boldsymbol{u}; \mathbb{Q}) &= \frac{\partial}{\partial u_k} \mathbb{E}_{\boldsymbol{z} \sim \mathbb{Q}}\left[\max_{\ell \in [K]}(u_\ell + z_\ell)\right] = \mathbb{E}_{\boldsymbol{z} \sim \mathbb{Q}}\left[\frac{\partial}{\partial u_k} \max_{\ell \in [K]}(u_\ell + z_\ell)\right] \\
&= \mathbb{E}_{\boldsymbol{z} \sim \mathbb{Q}}\left[\mathbb{1}_{\{k \in \operatorname{argmax}_{\ell \in [K]}(u_\ell + z_\ell)\}}\right] = \mathbb{Q}\left(k \in \operatorname*{argmax}_{\ell \in [K]}(u_\ell + z_\ell)\right) = p_k \quad \forall k \in [K],
\end{aligned}$$

where the first equality holds because $\max_{\boldsymbol{p} \in \Delta^K} \boldsymbol{p}^\top(\boldsymbol{u} + \boldsymbol{z}) = \max_{\ell \in [K]}(u_\ell + z_\ell)$, the second equality follows from the dominated convergence theorem, which applies because $\max_{\ell \in [K]}(u_\ell + z_\ell)$ is Lipschitz continuous in $\boldsymbol{u}$, and the third equality exploits Danskin's theorem [Bertsekas, 2016, Proposition B.25] together with the observation that the optimal solution of $\max_{\ell \in [K]}(u_\ell + z_\ell)$ is $\mathbb{Q}$-almost surely unique. In summary, we have thus shown that $\nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u}; \mathcal{B}) = \boldsymbol{p} = \nabla_{\boldsymbol{u}} \Phi^P(\boldsymbol{u}; \mathbb{Q})$. $\square$

We are now ready to address the open problem posed in Section 2. The following main theorem bridges the gap between regularization-based and perturbation-based algorithms for MAB problems. It shows that any FTRL algorithm with a convex, smooth and additively separable regularization function ψ is equivalent to an FTPL algorithm with some noise-sampling distribution $\mathbb{Q}$. This insight is consistent with the impossibility result by Kim and Tewari [2019] because the noise terms corresponding to different arms may be correlated under $\mathbb{Q}$. The conditions on ψ (especially the additively separability) are restrictive. However, to our best knowledge, the regularization functions of all commonly used FTRL algorithms (such as Tsallis-INF [Zimmert and Seldin, 2021], Exp3 [Auer et al., 1995] or FTRL with hybrid regularizers [Zimmert et al., 2019]) satisfy these properties.

Theorem 3.4 (FTRL vs. FTPL)**.** *Consider a regularization function $\psi : [0,1]^K \to \mathbb{R}$ defined via $\psi(\boldsymbol{p}) = \sum_{k=1}^K \psi_k(p_k)$, where $\psi_k : [0,1] \to \mathbb{R}$ is strictly convex and differentiable for every $k \in [K]$. Then, for every $\boldsymbol{u} \in \mathbb{R}^K$ there exists a distribution $\mathbb{Q} \in \mathcal{P}(\mathbb{R}^K)$ with $\nabla_{\boldsymbol{u}} \Phi^R(\boldsymbol{u}; \psi) = \nabla_{\boldsymbol{u}} \Phi^P(\boldsymbol{u}; \mathbb{Q})$.*

Proof. Fix an arbitrary $\boldsymbol{u} \in \mathbb{R}^K$. By Proposition 3.2, there exists a marginal ambiguity set $\mathcal{B}$ of the form (3) such that $\nabla_{\boldsymbol{u}} \Phi^R(\boldsymbol{u}; \psi) = \nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u}; \mathcal{B})$. Proposition 3.3 further implies that there exists a noise distribution $\mathbb{Q} \in \mathcal{P}(\mathbb{R}^K)$ with $\nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u}; \mathcal{B}) = \nabla_{\boldsymbol{u}} \Phi^P(\boldsymbol{u}; \mathbb{Q})$. Thus, the claim follows. $\square$

We emphasize that the distribution $\mathbb{Q}$ corresponding to a given regularization function ψ generically depends on the current reward estimate $\boldsymbol{u}$. In contrast, classical FTPL algorithms use a single noise distribution independent of $\boldsymbol{u}$. We also emphasize that the proofs of Propositions 3.2 and 3.3 and Theorem 3.4 are constructive and not merely existential, that is, we provide explicit formulas for the ambiguity set $\mathcal{B}$ as well as the noise distribution $\mathbb{Q}$ corresponding to ψ.

4 Regret Analysis of DOPA

A fundamental desideratum in algorithm design is stability. That is, small perturbations in the input of an algorithm should not dramatically alter its output. For example, GBPA with a convex differentiable potential function ϕ is stable if the arm-sampling distribution $\boldsymbol{p}(\boldsymbol{u}) = \nabla_{\boldsymbol{u}} \phi(\boldsymbol{u})$ is Lipschitz-continuous in the cumulative reward estimate $\boldsymbol{u}$. It is well known that adding a convex regularizer to the objective function of a parametric minimization problem improves the stability of its optimal solution [Bousquet and Elisseeff, 2002]. Improving the stability of FTRL, for instance, is tantamount to reducing the Lipschitz modulus of $\nabla_{\boldsymbol{u}} \boldsymbol{p}(\boldsymbol{u})$, which can be achieved by increasing the strong convexity constant of the underlying regularization function ψ. This is a direct consequence of the relation $\nabla_{\boldsymbol{u}} \boldsymbol{p}(\boldsymbol{u}) = \mathrm{diag}(\nabla^2_{\boldsymbol{u}} \phi(\boldsymbol{u})) \leq \mathrm{diag}((\nabla^2_{\boldsymbol{p}} \psi(\boldsymbol{p}))^{-1})$ [Penot, 1994; Abernethy et al., 2015].

Stability is a prerequisite for establishing sublinear regret bounds for FTRL algorithms in the adversarial regime [Abernethy et al., 2014, Section 3.1]. We will now leverage these results to identify conditions under which DOPA enjoys sublinear regret. Our analysis will reveal that the regret of DOPA critically depends on the choice of the marginal distribution functions F_k, $k \in [K]$.

Theorem 4.1 (Regret analysis of DOPA)**.** *Suppose that $\mathcal{B}$ is a marginal ambiguity set of the form (3). Assume also that the k-th marginal distribution function F_k is differentiable and strictly increasing whenever $F_k(s) \in (0,1)$ and that $\int_{\mathbb{R}} s F_k'(s) \mathrm{d}s = 0$ for all $k \in [K]$. If $\boldsymbol{p}(\boldsymbol{0}) = \nabla_{\boldsymbol{u}} \Phi(\boldsymbol{0}; \mathcal{B})$*

is the initial arm-sampling distribution and if there exist constants $\gamma \in (1,2)$ and $B > 0$ with

$$F_k'\left(F_k^{-1}(1-p)\right) \leq Bp^{\gamma} \quad \forall p \in (0,1),\ \forall k \in [K], \tag{8}$$

then the regret of DOPA satisfies

$$\mathcal{R}(T) \leq \sum_{k=1}^{K} \int_{1-p_k(\mathbf{0})}^{1} F_k^{-1}(t)\mathrm{d}t + \frac{1}{2}BTK^{2-\gamma}$$

under every possible reward distribution of a non-oblivious adversarial environment.

Proof. Let $k^\star \in \operatorname{argmax}_{k\in[K]} \mathbb{E}[\sum_{t=1}^{T} r_{t,k}]$ be the index of an arm with zero regret. Note that $\hat{r}_{t,k}$ as defined in Algorithm 1 is an unbiased estimator for $\mathbb{E}[r_{t,k}]$ for every arm $k \in [K]$ that has a positive probability $p_k(\hat{\boldsymbol{u}}_{t-1})$ of being selected. Also, recall from Lemma 3.1 that DOPA can be viewed as an FTRL algorithm with a convex regularization function $\psi(\boldsymbol{p}) = -\sum_{k=1}^{K}\int_{1-p_k}^{1} F_k^{-1}(t)\mathrm{d}t$. Thus, the FTRL regret decomposition in [Lattimore and Szepesvári, 2020, Theorem 28.10] implies that

$$\mathcal{R}(T) \leq \psi(\boldsymbol{e}_{k^\star}) - \psi(\boldsymbol{p}(\mathbf{0})) + \mathbb{E}\left[\sum_{t=1}^{T}\Big((\boldsymbol{p}_{t+1}-\boldsymbol{p}_t)^\top \hat{\boldsymbol{r}}_t - \mathbb{D}_\psi(\boldsymbol{p}_{t+1},\boldsymbol{p}_t)\Big)\right],$$

where $\hat{\boldsymbol{r}}_t$ and $\boldsymbol{p}_t$ are defined as in Algorithm 1. From the discussion after Lemma 3.1 we know that ψ is strictly convex. We may thus use [Lattimore and Szepesvári, 2020, Theorems 26.12 & 26.13] to bound the round-t term in the above sum by

$$(\boldsymbol{p}_{t+1}-\boldsymbol{p}_t)^\top \hat{\boldsymbol{r}}_t - \mathbb{D}_\psi(\boldsymbol{p}_{t+1},\boldsymbol{p}_t) \leq \sup_{\lambda\in[0,1]} \frac{1}{2}\hat{\boldsymbol{r}}_t^\top \left(\nabla^2_{\boldsymbol{p}}\psi(\lambda\boldsymbol{p}_{t+1}+(1-\lambda)\boldsymbol{p}_t)\right)^{-1}\hat{\boldsymbol{r}}_t.$$

The definition of $\hat{\boldsymbol{r}}_t$ further implies that

$$\hat{\boldsymbol{r}}_t^\top \left(\nabla^2_{\boldsymbol{p}}\psi(\lambda\boldsymbol{p}_{t+1}+(1-\lambda)\boldsymbol{p}_t)\right)^{-1}\hat{\boldsymbol{r}}_t = \frac{r_{t,a_t}^2}{p_{t,a_t}^2}\left(\left(\nabla^2_{\boldsymbol{p}}\psi(\lambda\boldsymbol{p}_{t+1}+(1-\lambda)\boldsymbol{p}_t)\right)^{-1}\right)_{a_t a_t}$$

for all $\lambda \in [0,1]$. Finally, observe that $\psi(\boldsymbol{e}_{k^\star}) = -\int_0^1 F_{k^\star}^{-1}(t)\,\mathrm{d}t = 0 = -\int_{\mathbb{R}} sF_{k^\star}'(s)\mathrm{d}s$ by assumption. Taken together, all of these insights allow us to conclude that

$$\mathcal{R}(T) \leq \sum_{k=1}^{K}\int_{1-p_k(\mathbf{0})}^{1} F_k^{-1}(t)\mathrm{d}t + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sup_{\lambda\in[0,1]}\frac{r_{t,a_t}^2}{p_{t,a_t}^2}\left(\nabla^2_{\boldsymbol{p}}\psi(\lambda\boldsymbol{p}_{t+1}+(1-\lambda)\boldsymbol{p}_t)\right)^{-1}_{a_t a_t}\right]. \tag{9}$$

In the remainder of the proof, we will establish an upper bound on the second term in (9). Recalling from Algorithm 1 that $\boldsymbol{p}_t = \boldsymbol{p}(\hat{\boldsymbol{u}}_{t-1})$, $\boldsymbol{u}_t = \boldsymbol{u}_{t-1} + \hat{\boldsymbol{r}}_t$ and $\hat{\boldsymbol{r}}_t = s\boldsymbol{e}_{a_t}$ with $s = -r_{t,a_t}/p_{t,a_t} \geq 0$, we find

$$p_{t+1,a_t} = p_{a_t}(\hat{\boldsymbol{u}}_t) = p_{a_t}(\hat{\boldsymbol{u}}_{t-1}+\hat{\boldsymbol{r}}_t) = p_{a_t}(\hat{\boldsymbol{u}}_{t-1} - s\boldsymbol{e}_{a_t}) \leq p_{a_t}(\hat{\boldsymbol{u}}_{t-1}) = p_{t,a_t}, \tag{10}$$

where the inequality holds because $\Phi(\boldsymbol{u};\mathcal{B})$ is convex in $\boldsymbol{u}$ such that $p_{a_t}(\boldsymbol{u}) = \partial_{u_{a_t}}\Phi(\boldsymbol{u};\mathcal{B})$ is non-decreasing in u_{a_t} and because $s \geq 0$. If $\boldsymbol{p} = \lambda\boldsymbol{p}_{t+1} + (1-\lambda)\boldsymbol{p}_t$ for some $\lambda \in [0,1]$, then we have

$$(\nabla^2_{\boldsymbol{p}}\psi(\boldsymbol{p}))^{-1}_{a_t a_t} = F_{a_t}'(F_{a_t}^{-1}(1-p_{a_t})) \leq Bp_{a_t}^{\gamma} = B(\lambda p_{t+1,a_t} + (1-\lambda)p_{t,a_t})^{\gamma} \leq Bp_{t,a_t}^{\gamma}, \tag{11}$$

where the first equality follows from the definition of $\psi(\boldsymbol{p})$, which implies via the inverse function theorem that $\partial^2_{p_k^2}\psi(\boldsymbol{p}) = (F_k'(F_k^{-1}(1-p_k)))^{-1}$. The second equality exploits the definition of $\boldsymbol{p}$, and the two inequalities follow from (8) and (10), respectively. The inequality (11) then implies that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\lambda\in[0,1]} \frac{r_{t,a_t}^2}{p_{t,a_t}^2}\left(\nabla^2_{\boldsymbol{p}}\psi(\lambda\boldsymbol{p}_{t+1}+(1-\lambda)\boldsymbol{p}_t)\right)^{-1}_{a_t a_t}\right] \\
&\qquad \le \mathbb{E}\left[\frac{r_{t,a_t}^2}{p_{t,a_t}^2}Bp_{t,a_t}^{\gamma}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{r_{t,a_t}^2}{p_{t,a_t}^2}Bp_{t,a_t}^{\gamma}\,\middle|\,\hat{\boldsymbol{u}}_{t-1},\boldsymbol{r}_t\right]\right] = \mathbb{E}\left[B\sum_{k=1}^K r_{t,k}^2 p_{t,k}^{\gamma-1}\right] \le \mathbb{E}\left[B\sum_{k=1}^K p_{t,k}^{\gamma-1}\right],
\end{aligned}$$

where the first equality exploits the law of iterated conditional expectations, and the second equality holds because $a_t = k$ with probability $p_{t,k}$ conditional on $\hat{\boldsymbol{u}}_{t-1}$ and $\boldsymbol{r}_t$. The second inequality holds because $r_{t,k}^2 \in [0,1]$ for all $k \in [K]$. Next, note that the $1/(2-\gamma)$-norm and the $1/(\gamma-1)$-norm are mutually dual for every $\gamma \in (1,2)$. Hölder's inequality thus implies that

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

This observation completes the proof. □

5 Optimality of DOPA

We now aim to identify marginal ambiguity sets $\mathcal{B}$ for which DOPA achieves optimal regret guarantees across different regimes. We will see that such optimal regret guarantees are available for certain Fréchet ambiguity sets $\mathcal{B}$ that are defined in terms of a marginal generator F.

Definition 2 (Marginal generator)**.** *A marginal generator $F : \mathbb{R} \to \mathbb{R}$ is a strictly increasing differentiable function with $\lim_{s\to-\infty} F(s) \le 0$, $\lim_{s\to+\infty} F(s) \ge 1$ and $\int_0^1 F^{-1}(t)\mathrm{d}t = 0$.*

Definition 3 (Fréchet ambiguity set)**.** *A Fréchet ambiguity set $\mathcal{B}$ is a marginal ambiguity set of the form (3), where the marginal cumulative distribution functions are defined through*

$$F_k(s) = \min\left\{1, \max\{0, 1 - F(-s/\eta_k)\}\right\} \quad \forall k \in [K] \tag{12}$$

for some vector $\boldsymbol{\eta} \in \mathbb{R}^K_{++}$ and some marginal generator F.

Before analyzing the regret of DOPA under Fréchet ambiguity sets, we present an auxiliary result that relates any potential function of the form $\Phi(\boldsymbol{u};\mathcal{B})$ induced by some Fréchet ambiguity set $\mathcal{B}$ to a potential functions of the form $\Phi^R(\boldsymbol{u};\psi)$ induced by some regularization function ψ. We will see that both $\mathcal{B}$ and ψ are uniquely determined by a vector $\boldsymbol{\eta}$ and a marginal generator F.

Theorem 5.1 (Fréchet regularization)**.** *Suppose that $\mathcal{B}$ is a Fréchet ambiguity set in the sense of Definition 3 induced by some $\boldsymbol{\eta} \in \mathbb{R}^K_{++}$ and some marginal generator F. If $f(s) = \int_0^s F^{-1}(t)\mathrm{d}t$ for all $s \in [0,1]$ and $\psi(\boldsymbol{p}) = \sum_{k=1}^K \eta_k f(p_k)$, then we have $\Phi(\boldsymbol{u};\mathcal{B}) = \Phi^R(\boldsymbol{u};\psi)$.*

As the marginal generator F is strictly increasing and as its range covers the open interval $(0,1)$, the inverse function $F^{-1}(t)$ is well-defined for for all $t \in (0,1)$, which in turn implies that $f(s)$ is well-defined for all $s \in [0,1]$. We trivially have $f(0) = 0$. In addition, $f(s)$ is smooth and convex (because F^{-1} inherits the monotonicity of F), and we have $f(1) = 0$ (because $\int_0^1 F^{-1}(t)\mathrm{d}t = 0$).

Proof of Theorem 5.1. As the marginal generator F is strictly increasing, the definition of F_k in (12) implies that $F_k^{-1}(x) = -F^{-1}(1-x)\eta_k$ for all $x \in (0,1)$. By the definition of f, we thus find

$$f(s) = \int_0^s F^{-1}(t)\mathrm{d}t = -\int_1^{1-s} F^{-1}(1-x)\,\mathrm{d}x = -\frac{1}{\eta_k}\int_{1-s}^1 F_k^{-1}(x)\mathrm{d}x \quad \forall s \in [0,1], \tag{13}$$

where the second equality follows from the substitution $x \leftarrow 1-t$. By Lemma 3.1, we thus obtain

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

This observation completes the proof. □

We are now ready to analyze the regret of DOPA under Fréchet ambiguity sets.

Theorem 5.2 (Regret analysis of DOPA with Fréchet ambiguity sets)**.** *Suppose that all conditions of Theorem 5.1 hold and that $\boldsymbol{\eta} = \eta\mathbf{1}$ for some $\eta > 0$. If there exist $\gamma \in (1,2)$ and $C > 0$ with*

$$F'(F^{-1}(p)) \le Cp^\gamma \quad \forall p \in (0,1), \tag{14}$$

then the regret of DOPA satisfies

$$\mathcal{R}(T) \le -\eta K f(1/K) + \frac{CTK^{2-\gamma}}{2\eta} \quad \forall T \in \mathbb{N}$$

under every possible reward distribution of a non-oblivious adversarial environment.

Proof. We first show that all conditions of Theorem 4.1 are satisfied. To this end, select any arm $k \in [K]$. Thanks to the assumed properties of F, the distribution function F_k as defined in (12) is differentiable and strictly increasing whenever $F_k(s) \in (0,1)$. In addition, observe that

$$\int_{\mathbb{R}} sF_k'(s)\mathrm{d}s = \int_0^1 F_k^{-1}(x)\mathrm{d}x = -\eta\int_0^1 F^{-1}(1-x)\mathrm{d}x = -\eta\int_0^1 F^{-1}(t)\mathrm{d}t = 0,$$

where the first equality follows from the substitution $F_k^{-1}(x) \leftarrow s$, the second equality holds because (12) implies that $F_k^{-1}(x) = -\eta F^{-1}(1-x)$ for all $x \in (0,1)$, the third equality exploits the substitution $t \leftarrow 1-x$, and the last equality holds by assumption. Furthermore, we have

$$\begin{aligned} F_k'(F_k^{-1}(1-p_k)) &= \left(\nabla_{\boldsymbol{p}}^2\left(-\sum_{k=1}^K \int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t\right)\right)_{kk}^{-1} \\ &= \left(\nabla_{\boldsymbol{p}}^2\left(\sum_{k=1}^K \eta\int_0^{p_k} F^{-1}(t)\mathrm{d}t\right)\right)_{kk}^{-1} = \frac{1}{\eta}F'(F^{-1}(p)) \le \frac{C}{\eta}p_k^\gamma \quad \forall p_k \in [0,1], \end{aligned}$$

where the first and third equalities follow from the inverse function theorem, and the second equality exploits again the definition of F_k in (12) and a simple variable substitution. The inequality, finally, follows from the assumption (14). Thus, F_k satisfies (8) with $B = C/\eta$. As $k \in [K]$ was chosen arbitrarily, we have now verified all conditions of Theorem 4.1. We may thus conclude that

$$\mathcal{R}(T) \leq \sum_{k=1}^{K} \int_{1-p_k(\mathbf{0})}^{1} F_k^{-1}(t)\mathrm{d}t + \frac{CTK^{2-\gamma}}{2\eta} = -\eta \sum_{k=1}^{K} f(p_k(\mathbf{0})) + \frac{CTK^{2-\gamma}}{2\eta}. \tag{15}$$

It remains to be shown that $\boldsymbol{p}(\mathbf{0}) = \nabla_{\boldsymbol{u}} \Phi(\mathbf{0}; \mathcal{B}) = \mathbf{1}/K$ is the unique maximizer of problem (4) at $\boldsymbol{u} = \mathbf{0}$. By the formula (13) for f in the proof of Theorem 5.1, problem (4) at $\boldsymbol{u} = \mathbf{0}$ has the same maximizers as $\max_{\boldsymbol{p} \in \Delta^K} H(\boldsymbol{p})$, where $H(\boldsymbol{p}) = -\sum_{k=1}^{K} f(p_k)$ is shorthand for the rescaled objective function. This problem is solvable thanks to Weierstrass' maximum theorem, which applies because $H(\boldsymbol{p})$ is continuous (in fact smooth) and Δ^K is compact. In the following, we use Π_K to denote the group of all permutations of $[K]$. Note that both $H(\boldsymbol{p})$ as well as the feasible set Δ^K are permutation symmetric. Hence, if $\boldsymbol{p}^\star$ is a maximizer, then so is $\boldsymbol{p}^\star_\pi = (p^\star_{\pi(1)}, \ldots, p^\star_{\pi(K)})$, for any $\pi \in \Pi_K$. As Δ^K is convex and as $|\Pi_K| = K!$, it is clear that the uniform convex combination $\bar{\boldsymbol{p}}^\star = \frac{1}{K!} \sum_{\pi \in \Pi_K} \boldsymbol{p}^\star_\pi$ belongs to the feasible set Δ^K, too. In addition, the objective function value of $\bar{\boldsymbol{p}}^\star$ satisfies

$$H(\bar{\boldsymbol{p}}^\star) \geq \frac{1}{K!} \sum_{\pi \in \Pi_K} H(\boldsymbol{p}^\star_\pi) = \max_{\boldsymbol{p} \in \Delta^K} H(\boldsymbol{p}).$$

Here, the inequality follows from Jensen's inequality, which applies because f is convex and H is concave. This implies that $\bar{\boldsymbol{p}}^\star$ is also an optimal solution. By construction, $\bar{\boldsymbol{p}}^\star$ is invariant under permutations of its elements, which allows us to conclude that $\bar{\boldsymbol{p}}^\star = \mathbf{1}/K$. We know from Lemma 3.1 that the maximizer of problem (4) is unique. In summary, we have thus verified that $\boldsymbol{p}(\mathbf{0}) = \nabla_{\boldsymbol{u}} \Phi(\mathbf{0}; \mathcal{B}) = \mathbf{1}/K$ is indeed the unique maximizer of (4). The claim then follows from (15). □

It is well known that the optimal regret in the adversarial regime is of the order $\mathcal{O}(\sqrt{KT})$ [Audibert and Bubeck, 2009, Theorem 1] and is achieved by an FTRL algorithm with a Tsallis entropy regularizer [Abernethy et al., 2015, Theorem 3.1]. The next corollary of Theorem 5.2 identifies a Fréchet ambiguity set for which DOPA offers the same optimal regret guarantee.

Corollary 5.3 (Optimality of DOPA)**.** *Suppose that $\mathcal{B}$ is a Fréchet ambiguity set and that the marginal generator is a shifted Pareto distribution of the form $F(s) = (1/\alpha - s(1-\alpha)/\alpha)^{-\frac{1}{1-\alpha}}$ with $\alpha \in (0,1)$. Then, the regret of DOPA with $\boldsymbol{\eta} = \eta \mathbf{1}$ and $\eta = \sqrt{(T(1-\alpha))/(2\alpha)} K^{\alpha - \frac{1}{2}}$ satisfies*

$$\mathcal{R}(T) \leq \sqrt{KT/(\alpha(1-\alpha))}$$

under every possible reward distribution of a non-oblivious adversarial environment.

Corollary 5.3 asserts that if $\mathcal{B}$ is a Fréchet ambiguity set generated by a shifted Pareto distribution, then DOPA with a learning rate η that is adapted to T attains the optimal adversarial regret $\mathcal{O}(\sqrt{KT})$.

Proof of Corollary 5.3. Observe first that F is indeed a marginal generator in the sense of Definition 2. From Theorem 5.1 we thus know that $\Phi(\boldsymbol{u};\mathcal{B}) = \Phi^R(\boldsymbol{u};\psi)$, where $\psi(\boldsymbol{p}) = \sum_{k=1}^K \eta f(p_k) = \eta \sum_{k=1}^K \int_0^{p_k} F^{-1}(t)\mathrm{d}t$. Thanks to our specific choice of F, we have

$$\int_0^{p_k} F^{-1}(t)\mathrm{d}t = \int_0^{p_k} \frac{1-\alpha t^{\alpha-1}}{1-\alpha}\mathrm{d}t = \frac{p_k - p_k^\alpha}{1-\alpha} \quad \forall k \in [K].$$

This implies that $\psi = \eta\psi_\alpha^{\mathbb{T}}$, where $\psi_\alpha^{\mathbb{T}}$ is the Tsallis entropy with parameter α defined in (1). In addition, one readily verifies that $F'(F^{-1}(p)) = p^{2-\alpha}/\alpha$. Thus, all conditions of Theorem 5.2 are satisfied with $C = 1/\alpha$ and $\gamma = 2-\alpha$, and we may conclude that

$$\mathcal{R}(T) \le -\eta K f(1/K) + \frac{CTK^{2-\gamma}}{2\eta} = \eta\frac{K^{1-\alpha}-1}{1-\alpha} + \frac{TK^\alpha}{2\alpha\eta} = \sqrt{\frac{KT}{\alpha(1-\alpha)}} \quad \forall T \in \mathbb{N},$$

where the second equality exploits our specific choice of η. Hence, the claim follows. $\square$

The regret analysis of DOPA developed in Corollary 5.3 is arguably simpler than that of optimal FTPL methods, which require subtle probabilistic arguments [Honda et al., 2023]. As it only uses basic tools from convex analysis instead of non-standard concepts from variational analysis such as sub-Hessians, it is even somewhat simpler than the regret analysis of the optimal FTRL method with Tsallis entropy regularizer in [Abernethy et al., 2015, Theorem 3.1]—despite many similarities.

Recall from Remark 1 that DOPA reduces to the Exp3 algorithm if $\mathcal{B}$ is a singleton containing a product Gumbel distribution. The following remark highlights that the Exp3 algorithm is also recovered from DOPA if $\mathcal{B}$ is a Fréchet ambiguity set with an exponential marginal generator.

Remark 2 (Exp3 algorithm revisited)**.** *Suppose that $\mathcal{B}$ is a Fréchet ambiguity set with $\boldsymbol{\eta} = \eta\mathbf{1}$ for some $\eta > 0$ and with marginal generator $F(s) = \exp(s-1)$. In this case, Theorem 5.1 implies that DOPA is equivalent to FTRL with regularization function*

$$\psi(\boldsymbol{p}) = \eta\sum_{k=1}^K \int_0^{p_k} F^{-1}(t)\mathrm{d}t = \eta\sum_{k=1}^K \int_0^{p_k} (\log(t)+1)\mathrm{d}t = \eta\sum_{k=1}^K p_k\log(p_k),$$

which is in turn known to be equivalent to the Exp3 Algorithm; see [Abernethy et al., 2015, Section 3]. This can also be checked directly. Indeed, by inspecting the optimality conditions of the convex program $\max_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top\boldsymbol{u} - \psi(\boldsymbol{p})$, one readily verifies that the corresponding arm-sampling probabilities are given by $p_k(\boldsymbol{u}) = \exp(u_k/\eta)/(\sum_{j=1}^K \exp(u_j/\eta))$ for all $k \in [K]$. However, these are precisely the arm-sampling probabilities of the Exp3 algorithm by Auer et al. [1995].

Corollary 5.3 relies on the implicit assumption that the learner knows the duration T of the game *ex ante* and is thus able to choose a learning rate η that adapts to T. In the remainder of this section we will show that DOPA can offer optimal regret guarantees even if T is unknown and even if there is uncertainty about the adversarial power of the environment. To this end, we study a generalized *anytime GBPA* algorithm that runs over an indefinite number of rounds; see Algorithm 2.

Algorithm 2 Anytime GBPA

Require: Differentiable convex functions $(\phi_t)_{t\in\mathbb{N}}$ with $\nabla_{\boldsymbol{u}}\phi_t(\boldsymbol{u}) \in \Delta^K$

Initialize $\hat{\boldsymbol{u}}_0 = \boldsymbol{0}$

for round $t \in \mathbb{N}$ **do**

Environment chooses a reward vector $\boldsymbol{r}_t \in [-1, 0]^K$

Learner chooses $a_t \sim \boldsymbol{p}_t = \nabla_{\boldsymbol{u}}\phi_t(\boldsymbol{u})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{t-1}}$ and receives reward 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

Algorithm 2 extends Algorithm 1 in that it runs forever and allows the potential function ϕ_t to change with t. Below, we will thus study a variant of DOPA with a time-dependent ambiguity set $\mathcal{B}_t$.

The optimal regret guarantees of any bandit algorithm depend on the adversarial power of the environment. Hence, an algorithm that attains the optimal regret in the non-oblivious adversarial regime is not guaranteed to attain the optimal regret in the stochastic regime, say. In order to present versions of DOPA that are simultaneously optimal across different learning regimes, we henceforth describe the adversarial power of the environment in a unifying manner via a self-bounding constraint [Zimmert and Seldin, 2021]. Formally, we thus assume that for any reward distribution available to the environment there exist $\boldsymbol{\Delta} \in [0, 1]^K$ and $C \geq 0$ such that the inequality

$$\mathcal{R}(T) \geq \sum_{t=1}^{T}\sum_{k=1}^{K} \Delta_k \mathbb{P}(a_t = k) - C \tag{16}$$

holds for all planning horizons $T \in \mathbb{N}$ and for all admissible arm-sampling distributions. For example, in the stochastic bandit setting, the reward vectors $\boldsymbol{r}_t$ are drawn independently from some fixed distribution on $[K]$. In this case, the regret can be written as

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

and thus satisfies (16) with $\Delta_k = \max_{\ell\in[K]} \mathbb{E}[r_{t,\ell}] - \mathbb{E}[r_{t,k}]$ for all $k \in [K]$ and $C = 0$. Similarly, one can show that (16) holds in the stochastically constrained adversarial [Wei and Luo, 2018] and the adversarially corrupted stochastic [Lykouris et al., 2018] learning regimes. Exploiting the equivalence of FTRL with Tsallis entropy regularization and DOPA with shifted Pareto marginals, we can now show that the anytime version of DOPA inherits the BOBW capability of FTRL.

Theorem 5.4 (BOBW capability of DOPA)**.** *Suppose that $\mathcal{B}_t$ is a time-dependent Fréchet ambiguity set in the sense of Definition 3 with $\boldsymbol{\eta} = \eta_t\boldsymbol{1}$, $\eta_t = 2\sqrt{t}$ and marginal generator $F(s) = (2-s)^{-2}$ for all $t \in \mathbb{N}$. Then, the regret of DOPA satisfies $\mathcal{R}(T) \leq 4\sqrt{KT} + 1$ for all $T \in \mathbb{N}$ and under all reward distributions of a non-oblivious adversarial environment. In addition, the regret of DOPA satisfies*

$$\mathcal{R}(T) \leq \mathcal{O}\left(\sum_{k\in[K]:\Delta_k>0} \log(T)/\Delta_k\right) \quad \forall T \in \mathbb{N} \tag{17}$$

under all reward distributions of an environment constrained by (16).

Note that the marginal generator $F(s) = (2-s)^{-2}$ coincides with the shifted Pareto distribution corresponding to $\alpha = 1/2$ from Corollary 5.3. Theorem 5.4 implies that DOPA achieves the optimal $\mathcal{O}(\log T)$ regret in the stochastic regime and the optimal $\mathcal{O}(\sqrt{KT})$ regret in the adversarial regime. Thanks to the time-dependent learning rate $\eta_t = 2\sqrt{t}$, DOPA also displays the anytime property, that is, it attains optimal regret bounds for every time horizon T without requiring knowledge of T.

Proof. Theorem 5.1 implies that $\Phi(\boldsymbol{u}; \mathcal{B}_t) = \Phi^R(\boldsymbol{u}; \psi_t)$, where $\psi_t(\boldsymbol{p}) = \eta_t \sum_{k=1}^K \int_0^{p_k} F^{-1}(t)\mathrm{d}t$. The proof of Corollary 5.3 further implies that $\psi_t = \eta_t \psi^{\mathrm{T}}_{1/2}$, where $\psi^{\mathrm{T}}_{1/2}$ stands for the Tsallis entropy with parameter 1/2. By [Zimmert and Seldin, 2021, Theorem 1], which applies to Tsallis-regularized FTPL algorithms with adaptive learning rate $\eta_t = 2\sqrt{t}$, we may then conclude that $\mathcal{R}(T) \leq 4\sqrt{KT} + 1$ for every $T \in \mathbb{N}$. If the adversary selects rewards that satisfy (16) and if $\Delta_k > 0$ for some $k \in [K]$, then [Ito, 2021, Theorem 2] further ensures that (17) holds. Hence, the claim follows. □

The following corollary shows that DOPA can even recover FTRL schemes with hybrid regularizers. This is achieved by studying harmonic averages of two different marginal generators.

Corollary 5.5 (Hybrid Fréchet regularizers)**.** *Select any weights $\gamma_1, \gamma_2 > 0$ and any marginal generators G_1 and G_2. Suppose that $\mathcal{B}$ is a Fréchet ambiguity set with $\boldsymbol{\eta} = \mathbf{1}$ and with marginal generator $F(s) = (\gamma_1 G_1^{-1} + \gamma_2 G_2^{-1})^{-1}(s)$. If $g_1(s) = \int_0^s G_1^{-1}(t)\mathrm{d}t$ and $g_2(s) = \int_0^s G_2^{-1}(t)\mathrm{d}t$ for all $s \in \mathbb{R}$, then $\Phi(\boldsymbol{u}; \mathcal{B}) = \Phi^R(\boldsymbol{u}; \psi)$, where $\psi(\boldsymbol{p}) = \sum_{k=1}^K (\gamma_1 g_1(p_k) + \gamma_2 g_2(p_k))$.*

Corollary 5.5 shows that any FTRL method with a hybrid regularizer representable as a sum of two convex functions can be interpreted as an instance of DOPA with a Fréchet ambiguity set induced by a harmonic average of two marginal generators. In conjunction with Theorem 3.4, this result implies that we can systematically construct FTPL algorithms that are equivalent to FTRL algorithms with hybrid regularizers, some of which are known to display attractive BOBW capabilities. Corollary 5.5 thus addresses an open problem posed by Honda et al. [2023], who state that *"it would be a very challenging task to realize the effect of hybrid regularization by FTPL."*

Proof of Corollary 5.5. By Lemma 3.1, we have

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

The definitions of g_1 and g_2 further imply that

$$\begin{aligned}
\gamma_1 g_1(s) + \gamma_2 g_2(s) &= \gamma_1 \int_0^s G_1^{-1}(t)\,\mathrm{d}t + \gamma_2 \int_0^s G_2^{-1}(t)\,\mathrm{d}t \\
&= \int_{1-s}^1 (\gamma_1 G_1^{-1} + \gamma_2 G_2^{-1})(1-x)\,\mathrm{d}x \\
&= \int_{1-s}^1 F^{-1}(1-x)\,\mathrm{d}x = -\int_{1-s}^1 F_k^{-1}(x)\,\mathrm{d}x \quad \forall k \in [K],
\end{aligned}$$

where the second and the third equalities follow from the variable substitution $x \leftarrow 1 - t$ and the definition of F, respectively. The last equality exploits the definition of F_k in (12) and the assumption that $\boldsymbol{\eta} = \mathbf{1}$. Combining the above derivations and recalling the definition of ψ then yields

$$\Phi(\boldsymbol{u}; \mathcal{B}) = \max_{\boldsymbol{p} \in \Delta^K} \sum_{k=1}^K u_k p_k - \sum_{k=1}^K (\gamma_1 g_1(p_k) + \gamma_2 g_2(p_k)) = \Phi^R(\boldsymbol{u}; \psi).$$

Hence, the claim follows. $\square$

All optimal FTRL methods with hybrid regularizers studied to date assume that the regularizer consists of a sum of merely *two* elementary convex functions [Jin et al., 2024; Zimmert et al., 2019]. More versatile FTRL methods can be obtained by generalizing Corollary 5.5 in the obvious way, that is, by setting the regularization function ψ to the sum of the integrals of $N > 2$ inverse marginal generators $G_1^{-1}, \ldots, G_N^{-1}$. The resulting FTRL method can then be interpreted as a version of DOPA whose Fréchet ambiguity set is generated by the harmonic mean of $G_1, \ldots, G_N$.

To close this section, we use Corollary 5.5 to show that DOPA with a Fréchet ambiguity set generated by two marginal generators can achieve theoretically optimal BOBW guarantees.

Corollary 5.6 (Adaptive hybrid Fréchet regularizers)**.** *Suppose that $\mathcal{B}_t$ is a time-dependent Fréchet ambiguity set in the sense of Definition 3 with $\boldsymbol{\eta} = \mathbf{1}$ and marginal generator $F(s) = (\gamma_t G_1^{-1} + \gamma_t G_2^{-1})^{-1}(s)$, where $\gamma_t = \sqrt{t}$, $G_1(s) = 1 - \exp(-(s+1))$ and $G_2(s) = (-2s)^{-2}$ for all $t \in \mathbb{N}$. Then, the regret of DOPA satisfies $\mathcal{R}(T) \leq \mathcal{O}(\sqrt{KT})$ for all $T \in \mathbb{N}$ and under all reward distributions of a non-oblivious adversarial environment. In addition, the regret of DOPA then also satisfies*

$$\mathcal{R}(T) \leq \mathcal{O}\left(\sum_{k \in [K]: \Delta_k > 0} \log T / \Delta_k \right) + \mathcal{O}\left(\sum_{k \in [K]: \Delta_k > 0} (\log K)^2 / \Delta_k \right) \quad \forall T \in \mathbb{N}$$

under all reward distributions of an environment constrained by (16) with $|\{k \in [K] : \Delta_k = 0\}| = 1$.

Proof. Corollary 5.5 readily implies that $\Phi(\boldsymbol{u}; \mathcal{B}_t) = \Phi^R(\boldsymbol{u}; \psi_t)$ for every $t \in [T]$, where $\psi_t(\boldsymbol{p}) = \sum_{k=1}^K (\gamma_t g_1(p_k) + \gamma_t g_2(p_k))$, $g_1(s) = \int_0^s G_1^{-1}(t)\, \mathrm{d}t$ and $g_2(s) = \int_0^s G_2^{-1}(t)\, \mathrm{d}t$. In addition, one readily verifies that $G_1^{-1}(t) = -1 - \log(1-t)$ and $G_2^{-1}(t) = -(2\sqrt{t})^{-1}$, which implies that $\gamma_t g_1(s) + \gamma_t g_2(s) = -\gamma_t(\sqrt{s} + (s-1)\log(1-s))$. Hence, the version of DOPA at hand is equivalent to an FTRL method with regularizer $\psi_t(\boldsymbol{p}) = \sum_{k=1}^K -\sqrt{t}(\sqrt{p_k} + (p_k - 1)\log(1-p_k))$. The claim then follows from general results on FTRL algorithms with hybrid regualrizers [Zimmert et al., 2019, Theorem 3]. $\square$

The results of this section imply via Lemma 3.1 and Proposition 3.3 that there exist FTPL algorithms that are equivalent to FTRL algorithms with hybrid regularizers (in particular optimal ones).

6 Computational Efficiency of DOPA

FTPL algorithms are popular primarily because of their computational efficiency. Indeed, the arm a_t to be pulled in round t is found by sampling $\boldsymbol{z} \sim \mathbb{Q}$ and then identifying the largest component of the perturbed reward estimate $\hat{\boldsymbol{u}}_{t-1} + \boldsymbol{z}$. Recall that the components of $\boldsymbol{z}$ are usually assumed to be i.i.d. under $\mathbb{Q}$. Hence, the per-iteration complexity of FTPL is of the order $\mathcal{O}(K)$. In contrast, FTRL algorithms need to solve a convex optimization problem in each round t, which imposes a significantly higher computational burden. From Theorem 3.4 we know, however, that every FTRL algorithm induced by an additively separable regularization function ψ is equivalent to an FTPL algorithm induced by some disturbance distribution $\mathbb{Q}$. This connection between FTRL and FTPL is mediated by DOPA. Specifically, the noise distribution $\mathbb{Q}$ is a solution of the optimization problem (2) and thus changes with the reward estimate $\boldsymbol{u} = \hat{\boldsymbol{u}}_{t-1}$. The FTPL algorithm corresponding to a given FTRL algorithm thus needs to solve an instance of (2) in each round t in order to compute the current noise distribution. Hence, it appears that all computational advantages of FTPL *vis-à-vis* FTRL are outweighed by the time needed to solve just another optimization problem.

We will now show that this suspicion is unwarranted. Instead of computing $\mathbb{Q}$ by solving problem (2) at $\boldsymbol{u} = \hat{\boldsymbol{u}}_{t-1}$ and then sampling a_t from $\boldsymbol{p} = \nabla_{\boldsymbol{u}} \Phi^P(\boldsymbol{u}; \mathbb{Q})$ by drawing a sample $\boldsymbol{z}$ from $\mathbb{Q}$, we propose here to compute the arm-sampling distribution $\boldsymbol{p}$ *directly*. This can be done highly efficiently by recalling from Proposition 3.3 that $\boldsymbol{p} = \nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u}; \mathcal{B})$ and by leveraging a bisection method inspired by [Taşkesen et al., 2023, Algorithm 2] for computing $\nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u}; \mathcal{B})$; see Algorithm 3 below. This method has its roots in semi-parametric discrete choice theory, which exploits the structure of the marginal ambiguity set $\mathcal{B}$ to compute the vector of optimal choice probabilities. Algorithm 3 relies on the modulus of uniform continuity of the marginal distribution functions F_k, $k \in [K]$, with respect to a prescribed tolerance $\varepsilon \geq 0$, which is defined as

$$\delta(\varepsilon) = \min_{k \in [K]} \max_{\delta > 0} \left\{ \delta : |F_k(t_1) - F_k(t_2)| \leq \varepsilon / (2\sqrt{K}) \ \forall t_1, t_2 \in \mathbb{R} \text{ with } |t_1 - t_2| \leq \delta \right\}.$$

Algorithm 3 Bisection method for approximating the arm-sampling distribution $\boldsymbol{p} = \nabla_{\boldsymbol{u}} \Phi(\boldsymbol{u}; \mathcal{B})$

Require: error tolerance ε, reward estimate $\boldsymbol{u}$, marginal distribution functions F_k, $k \in [K]$
  Set $\bar{\tau} \leftarrow \max_{k \in [K]} \{-u_k - F_k^{-1}(1 - 1/K)\}$
  Set $\underline{\tau} \leftarrow \min_{k \in [K]} \{-u_k - F_k^{-1}(1 - 1/K)\}$
  for $i = 1, 2, \ldots, \lceil \log_2((\bar{\tau} - \underline{\tau})/\delta(\varepsilon)) \rceil$ **do**
    Set $\tau \leftarrow (\tau + \underline{\tau})/2$
    Set $\hat{p}_k \leftarrow 1 - F_k(-u_k - \tau)$ for $k \in [K]$
    if $\sum_{k \in [K]} \hat{p}_k > 1$ **then** $\bar{\tau} \leftarrow \tau$ **else** $\underline{\tau} \leftarrow \tau$
  end for
  return $\hat{\boldsymbol{p}}$ with $\hat{p}_k = (1 + \sum_{\ell=1}^{K} F_\ell(-u_\ell - \underline{\tau}))/K - F_k(-u_k - \underline{\tau})$ for all $k \in [K]$

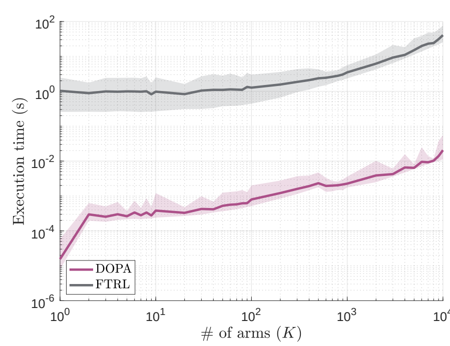

Figure 1: Per-iteration runtime of DOPA (purple) and FTRL (gray) as a function of the number K of arms. The solid lines show the means, and the shaded areas visualize the corridor between the minima and maxima across 10 independent simulations ($\boldsymbol{u}$ is sampled uniformly from $[0, 1]^K$).

The following corollary of [Taşkesen et al., 2023, Theorem 4.9] characterizes the convergence behavior of Algorithm 3.

Theorem 6.1 (Convergence of Algorithm 3)**.** *Suppose that $\mathcal{B}$ is a marginal ambiguity set of the form (3) and that the distribution functions F_k, $k \in [K]$ are continuous and strictly increasing in s whenever $F_k(s) \in (0,1)$. Then, for any $\boldsymbol{u} \in \mathbb{R}^K$ and $\varepsilon > 0$, Algorithm 3 outputs $\hat{\boldsymbol{p}} \in \Delta^K$ with $\|\hat{\boldsymbol{p}} - \nabla_{\boldsymbol{u}}\Phi(\boldsymbol{u}, \mathcal{B})\|_2 \leq \varepsilon$. If $\mathcal{B}$ is additionally a Fréchet ambiguity set with $\boldsymbol{\eta} = \eta\mathbf{1}$ for some $\eta > 0$ and if the marginal generator F is L-Lipschitz continuous whenever $F(s) \in [0,1]$, then Algorithm 3 terminates after at most $\log_2(\varepsilon^{-1}2L\sqrt{K}(\bar{u} - \underline{u})/\eta)$ iterations with $\bar{u} = \max_{k\in[K]} u_k$ and $\underline{u} = \min_{k\in[K]} u_k$.*

Proof. For simplicity of notation, we define $\boldsymbol{p} = \nabla_{\boldsymbol{u}}\Phi(\boldsymbol{u}; \mathcal{B})$. Note that the output of Algorithm 3 can be expressed as $\hat{\boldsymbol{p}} = \boldsymbol{q} + d\mathbf{1}/K$, where $q_k = 1 - F_k(-u_k - \underline{\tau})$ for all $k \in [K]$ and $d = 1 - \sum_{k=1}^K q_k$ is a nonnegative normalization constant. By [Taşkesen et al., 2023, Theorem 4.9], we know that $|q_k - p_k| \leq \varepsilon/(2\sqrt{K})$ for all $k \in [K]$. This implies that

$$\|\hat{\boldsymbol{p}} - \boldsymbol{p}\|_2 = \left\|\boldsymbol{q} + \frac{d\mathbf{1}}{K} - \boldsymbol{p}\right\|_2 \leq \|\boldsymbol{q} - \boldsymbol{p}\|_2 + \left\|\frac{d\mathbf{1}}{K}\right\|_2 = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

where the second equality holds because $d = 1 - \sum_{k=1}^K q_k \leq 1 - \sum_{k=1}^K p_k + K\varepsilon/(2\sqrt{K}) = \varepsilon\sqrt{K}/2$. As for the second claim, note that the L-Lipschitz continuity of F implies via the definition of F_k in (12) that the uniform continuity parameter $\delta(\varepsilon)$ is bounded below by $\varepsilon\eta/(2L\sqrt{K})$. Also, as all components of $\boldsymbol{\eta}$ are identical, one readily verifies that $\overline{\tau} - \underline{\tau} \leq \overline{u} - \underline{u}$. The number of iterations of Algorithm 3 is therefore bounded above by $\log_2(\varepsilon^{-1}2L\sqrt{K}(\bar{u} - \underline{u})/\eta)$. □

Note that DOPA calls Algorithm 3 with input $\boldsymbol{u} = \hat{\boldsymbol{u}}_{t-1}$ in each iteration $t = 1, \ldots, T$ of Algorithm 1 in order to compute an arm-sampling distribution $\boldsymbol{p}_t$. As $\hat{\boldsymbol{u}}_{t-1} = \sum_{s=1}^{t-1} \hat{\boldsymbol{r}}_s$, the range $\bar{u} - \underline{u}$ of the reward estimates is uncertain and depends on t. In addition, as $\boldsymbol{r}_s \in [-1, 0]^K$ and $\hat{\boldsymbol{r}}_s = (r_{s,a_s}/p_{s,a_s})\boldsymbol{e}_{a_s}$ for all $s = 1, \ldots, t-1$, we have $\hat{\boldsymbol{u}}_{t-1} = \mathcal{O}(t)$ with high probability. This observation implies that the t-th call of Algorithm 3 has $\mathcal{O}(\log(\sqrt{K}t/\eta))$ iterations with high probability. In addition, each iteration runs in time $\mathcal{O}(K)$. Hence, if $\eta = \mathcal{O}(\sqrt{T})$ (which leads to optimal regret guarantees as explained in Corollary 5.3), then the t-th call of Algorithm 3 runs in time at most $\mathcal{O}(K\log(\sqrt{KT})) = \tilde{\mathcal{O}}(K)$ with high probability, where $\tilde{\mathcal{O}}$ hides logarithmic factors. The efficiency of Algorithm 3 used by DOPA is thus comparable to the sampling procedure used by FTPL. We highlight that the marginal generators of all Fréchet ambiguity sets that were examined in Section 5 and lead to optimal regret bounds satisfy the Lipschitz continuity condition of Theorem 6.1.

We now compare the per-iteration complexities of DOPA and FTRL, that is, we measure the times both methods spend on computing the arm-sampling distributions. All experiments are run on a computer with an Apple M1 Pro processor with 16GB RAM, and all optimization problems are modeled in MATLAB using the YALMIP interface [McCormick, 1976]. The code for reproducing Figure 1 is available from `https://anonymous.4open.science/r/bandit-experiments-FB73/`.

As for DOPA, we set $\mathcal{B}$ to a Fréchet ambiguity set in the sense of Definition 3 with marginal generator $F(s) = (2-s)^{-2}$ and $\boldsymbol{\eta} = \mathbf{1}$. As for FTRL, we set ψ to the Tsallis entropy with parameter $\alpha = \frac{1}{2}$. Theorem 5.1 and Corollary 5.3 then imply that $\Phi(\boldsymbol{u}; \mathcal{B}) = \Phi^R(\boldsymbol{u}; \psi)$; see also the proof of Theorem 5.4 for further details. Hence, DOPA and FTRL use the same arm-sampling distributions and are thus equivalent. We compute the arm-sampling distribution $\nabla_{\boldsymbol{u}}\Phi^R(\boldsymbol{u}; \psi) =$

$\operatorname{argmax}_{\boldsymbol{p}\in\Delta^K} \boldsymbol{p}^\top \boldsymbol{u} - \psi(\boldsymbol{p})$ of FTRL by solving the underlying second-order-cone program with MOSEK [MOSEK ApS, 2024]. In addition, we use Algorithm 3 to compute the arm-sampling distribution $\nabla_{\boldsymbol{u}}\Phi(\boldsymbol{u};\mathcal{B})$ of DOPA to within an error tolerance of $\varepsilon = 10^{-8}$, which matches MOSEK's suboptimality tolerance for conic programs. Figure 1 visualizes the per-iteration runtimes of DOPA and FTRL as a function of the number K of arms. We observe that DOPA runs almost 10^4 times faster uniformly across all K.

7 Concluding Remarks and Limitations

We introduce DOPA as a new GBPA algorithm that builds a bridge between FTPL and FTRL methods. DOPA is based on an "*optimism in the face of ambiguity*" principle and implicitly solves optimization problems over marginal ambiguity sets in order to determine FTPL-type noise distributions. DOPA enables us to establish a one-to-one correspondence between FTRL algorithms with additively separable regularization functions and FTPL algorithms. As a result, it circumvents the challenges associated with the regret analysis of FTPL-type algorithms and with the computational complexity of FTRL-type algorithms. Indeed, DOPA provides a unified regret analysis for perturbation-based methods by connecting them to FTRL methods, thus paving the way for new FTPL algorithms with optimal regret guarantees. In addition, the arm-sampling distributions of DOPA can be computed highly efficient with a bisection algorithm inspired by modern discrete choice theory. We show that the per-iteration complexity of DOPA exceeds that of FTPL algorithms only by logarithimc factors in K and T. We see potential in exploring variants of DOPA with new Fréchet ambiguity sets that induce unconventional regularizers (see, *e.g.*, [Taşkesen et al., 2023, Example 3.11]) or with completely different classes of ambiguity sets.

The design principle behind DOPA extends beyond the K-armed bandit setting while preserving BOBW capability. Notable future applications of DOPA include decoupled exploitation-exploration [Jin et al., 2024], where a learner can choose to receive a reward from one arm while simultaneously gathering information about the reward from another. This concept has significant implications for the development of efficient reinforcement learning algorithms [Huang et al., 2022]. Another potential application of our algorithm, where it can achieve a BOBW regret bound, is the dueling bandit problem [Zimmert and Seldin, 2021]. In this setting, the learner selects two arms in each round to "duel" and receives feedback on the arm with the higher reward. Dueling bandit models have practical applications, such as hyperparameter tuning [Kumagai, 2017]. Furthermore, our framework generalizes the hybrid Tsallis entropy regularizers used in an FTRL-type algorithm with BOBW capability [Ito et al., 2024], making it applicable to both K-armed bandit and linear bandit problems.

We also recognize several limitations of our work. First, certain types of regularizers cannot be captured by marginal ambiguity sets of the form (3). A notable example is the log-barrier regularizer considered by Jin et al. [2024]. [Hofbauer and Sandholm, 2002, Proposition 2.2] shows that it is *impossible* to recover an FTRL algorithm with a log-barrier regularizer using any FTPL algorithm with a stochastic perturbation whose distribution is independent of the reward estimates $\boldsymbol{u}$. Second, the bisection method in Algorithm 3 is efficient as long as the marginal cumulative distribution functions F_k and their inverses F_k^{-1} can be computed efficiently. However, for hybrid regularizers, computing F_k can be cumbersome, making the bisection method computationally inefficient for certain choices of the marginal generators G_1 and G_2.

Appendix: Strongly Convex Regularization Functions

The following lemma borrowed from Taşkesen et al. [2023, Proposition 4.8] identifies sufficient conditions on the distribution functions F_k, $k \in [K]$, under which the regularization function $\psi(\boldsymbol{p}) = -\sum_{k=1}^K \int_{1-p_k}^1 F_k^{-1}(t)\, \mathrm{d}t$ is strongly convex. It exploits a natural duality relation between smoothness and strong convexity properties. We sketch the proof of this result for completeness.

Lemma 7.1. *If $\mathcal{B}$ is a marginal ambiguity set of the form (3), and if the cumulative distribution functions $F_k, k \in [K]$, are Lipschitz continuous with Lipschitz constant L, then the regularization function $\psi(\boldsymbol{p}) = -\sum_{k=1}^K \int_{1-p_k}^1 F_k^{-1}(t)\, \mathrm{d}t$ is L-strongly convex on $[0,1]^K$.*

Proof. The claim holds if we can show that $\psi(\boldsymbol{p}) - \|\boldsymbol{p}\|_2^2/(2L)$ is convex in $\boldsymbol{p}$. As F_k is non-decreasing and Lipschitz continuous by assumption, we have

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

where the second inequality follows from restricting s_1 and s_2 to the image of $(0,1)$ under the (left) quantile function F_k^{-1}. Rearranging terms in the above inequality then yields

$$-F_k^{-1}(1-q_k) - q_k/L \leq -F_k^{-1}(1-p_k) - p_k/L \quad \forall p_k, q_k \in (0,1) \text{ with } q_k < p_k.$$

Thus, the function $-F_k^{-1}(1-p_k) - p_k/L$ is non-decreasing in p_k on the open interval $(0,1)$, and its primitive $-\int_{1-p_k}^1 F_k^{-1}(t)\mathrm{d}t - p_k^2/(2L)$ is convex and continuous in p_k on the closed interval $[0,1]$. The claim then follows because convexity is preserved under summation. □

References

J. 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. 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. Fighting bandits with a new kind of smoothness. *Advances in Neural Information Processing Systems*, 2015.

J. Abernethy, C. Lee, and A. Tewari. Perturbation techniques in online learning and optimization. In T. Hazan, G. Papandreou, and D. Tarlow, editors, *Perturbations, Optimization, and Statistics*, chapter 8, pages 233–264. MIT Press, 2017.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Annual IEEE Symposium on 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.

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2016.

O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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.

G. Feng, X. Li, and Z. Wang. On the relation between several discrete choice models. *Operations research*, 65(6):1516–1525, 2017.

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.

J. Huang, L. Zhao, T. Qin, W. Chen, N. Jiang, and T.-Y. Liu. Tiered reinforcement learning: Pessimism in the face of uncertainty and constant regret. *Advances in Neural Information Processing Systems*, 2022.

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. *Conference on Learning Theory*, 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.

W. Kumagai. Regret analysis for continuous dueling bandit. *Advances in Neural Information Processing Systems*, 30, 2017.

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. *Conference on Learning Theory*, 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. In C. Manski and D. McFadden, editors, *Structural Analysis of Discrete Data with Econometric Applications*, chapter 5, pages 198–272. MIT Press, 1981.

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

MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 10.1.*, 2024.

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.

J.-P. Penot. Sub-Hessians, super-Hessians and conjugation. *Nonlinear Analysis: Theory, Methods & Applications*, 23(6):689–702, 1994.

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.