Submitted to *Mathematics of Operations Research*
manuscript (Please, provide the manuccript number!)

A General Framework for Bandit Problems Beyond Cumulative Objectives

Asaf Cassel
School of Computer Science, Tel Aviv University, acassel@mail.tau.ac.il

Shie Mannor
Faculty of Electrical Engineering, Technion, Israel Institute of Technology, shie@technion.ac.il
Nvidia Research, smannor@nvidia.com

Assaf Zeevi
Graudate School of Business, Columbia University, assaf@gsb.columbia.edu

The stochastic multi-armed bandit (MAB) problem is a common model for sequential decision problems. In the standard setup, a decision maker has to choose at every instant between several competing arms, each of them provides a scalar random variable, referred to as a "reward." Nearly all research on this topic considers the total cumulative reward as the criterion of interest. This work focuses on other natural objectives that cannot be cast as a sum over rewards, but rather more involved functions of the reward stream. Unlike the case of cumulative criteria, in the problems we study here the oracle policy, that knows the problem parameters a priori and is used to "center" the regret, is not trivial. We provide a systematic approach to such problems, and derive general conditions under which the oracle policy is sufficiently tractable to facilitate the design of optimism-based (upper confidence bound) learning policies. These conditions elucidate an interesting interplay between the arm reward distributions and the performance metric. Our main findings are illustrated for several commonly used objectives such as conditional value-at-risk, mean-variance trade-offs, Sharpe-ratio, and more.

Key words: Multi-armed Bandit, risk, planning, reinforcement learning, Upper Confidence Bound, optimism principle.

1. Introduction

Consider a sequential decision making problem where at each stage one of K independent alternatives is to be selected. When choosing alternative i at stage t (also referred to as time t), the decision maker receives a reward X_t that is distributed according to some *unknown* distribution $F^{(i)}$, $i=1,\ldots,K$ and is independent of t. (Where unambiguous, we avoid indexing X_t with i, and leave that implicit; the information will be encoded in the policy that governs said choices, which will be detailed in what follows.) At time t, the decision maker has accumulated a vector of rewards $(X_1,\ldots,X_t)$. In our setting, performance criteria are defined by a function $\tilde{U}$ that maps the reward vector to a real-valued number. As $\tilde{U}(X_1,\ldots,X_t)$ is a random quantity, we consider the accepted notion of expected performance, i.e., $\mathbf{E}\tilde{U}(X_1,\ldots,X_t)$, assuming this expectation exists and is finite. An oracle, with full knowledge of the arms' distributions, will make a sequence of selections based on this information so as to maximize the expected performance criterion. This serves as a benchmark for any other policy which does not have such information a priori, and hence needs to learn it on the fly. The gap between the former (performance of the oracle) and the latter (performance of the policy) represents the usual notion of regret in the learning problem.

The most ubiquitous performance criterion in the literature concerns the long run average reward, which involves the empirical mean, $\tilde{U}^{ave}(X_1,\dots,X_t)=\frac{1}{t}\sum_{s=1}^{t}X_s$. In this case, the oracle rule, that maximizes the expected value of the above, just samples from the distribution with the highest mean value, namely, it selects $i^*\in\arg\max\{\int x dF^{(i)}(x)\}$. Learning algorithms for such problems date back to Robbins' paper [20] and were extensively studied subsequent to that. In particular, the seminal work of [16] establishes that the normalized (per epoch) regret in this problem, when the arms are "well separated," cannot be made smaller than $\mathcal{O}(\log T/T)$, and there exist learning algorithms that achieve this regret by maximizing a confidence bound modification of the empirical mean. (When the arms are not well separated, equivalent statements hold with order-$1/\sqrt{T}$.) Since then, this class of policies has come to be known as UCB, or upper confidence bound policies. Some strands of literature that have emerged from this include [4] (non-asymptotic analysis of UCB-policies), [18] (empirical confidence bounds or KL-UCB), [2] (Thompson sampling based algorithms), and various works which consider an adversarial formulation (see, e.g., [5]).

Main research questions. In this paper we are interested in studying the above problem for more general *path dependent* objectives that are of interest beyond just the vanilla average. Many of these objectives bear an interpretation as "risk criteria" insofar as they focus on a finer probabilistic nature of the primitive distributions than the mean, and typically relate to the spread or tail behavior. Examples include: the so-called *Sharpe ratio*, which is the ratio between the mean and standard deviation; *value-at-risk* (VaR_α) which focuses on the α percentile of the distribution (with α small); or a close counterpart that integrates (averages) the values out in the tail beyond that point known as the *expected shortfall* (or conditional value at risk; $CVaR_\alpha$). The last example is of further interest as it belongs to the class of *coherent* risk measures which has various attractive properties from the risk theory perspective. A discussion thereof is beyond the scope of this paper; cf. [3] for further details. In our problem setting, the above criteria are applied via the function $\tilde{U}$ to the empirical observations, and then the decision maker seeks, as before, to optimize its expected value. A typical example where such criteria may be of interest is that of clinical trials (one of the original motivations for the development of the MAB framework). More specifically, suppose several new drugs are sequentially tested on individuals who share similar characteristics. If we consider average performance, we may conclude that the best choice is a drug with a non-negligible fatality rate but a high success rate. If we wish to control the fatality rate then using $CVaR_\alpha$ for example may be appropriate.

While some of the above mentioned criteria have been examined in the decision making and learning literature (see references below), the analysis tends to be driven by specific properties of the criterion in question and is very much done on a case-by-case basis. One of the purposes of this paper is to present a more unified approach to a large set of such problems. One of the main obstacles that arises under the non-cumulative criteria is the more complicated structure of the oracle rule. In particular, unlike the case of the mean objective, here the oracle rule need not select the same arm throughout the horizon of the problem. This presents further obstacles in identifying and characterizing a learning policy, as most such blueprints call for minimizing regret by mimicking the oracle rule. To that end, as our analysis will flesh out, under suitable conditions the oracle policy can be approximated (asymptotically) by a *simple policy*, that is, one that statically selects a single arm. This simplification can be leveraged to address the *learning problem* which becomes more tractable and amenable to optimism-based design principles. It is therefore of interest to understand and characterize in what instances does this simplified structure exist. This is one of the main thrusts of the paper.

Main contributions of this paper. In this paper we consider a general approach to the analysis of performance criteria where the oracle policy is a simple policy. We identify a class of criteria that we term *Empirical Distribution Performance Measures* (EDPM). In particular, let $\hat{F}$ be the *empirical distribution* of the vector $(X_1, \dots, X_t)$, i.e., $\hat{F}(y)$ is the fraction of rewards less or equal to real valued y. An EDPM evaluates performance by means of a function U, which maps $\hat{F}$ to $\mathbb{R}$, i.e., $U(\hat{F}) = \tilde{U}(X_1, \dots, X_t)$. Alternatively, U may also serve to evaluate the distributions of the random variables X_s $(s = 1, \dots, t)$. These evaluations may be aggregated to form a different type of performance criteria that we term "pseudo regret," and consider as an intermediate learning goal. Our main results provide easily verifiable explicit conditions that characterize the asymptotic behavior of the oracle rule, and culminate in a UCB-type learning algorithm with either $\mathcal{O}(\log T/T)$ or $\mathcal{O}(1/\sqrt{T})$ normalized regret (depending on the properties of U and the arm distributions).

Previous works on bandits that concern path-dependent and risk criteria. Sequential performance measures of the type considered here were previously studied in [21], which considered the Mean-Variance of the sequence and presented the MV-UCB, and MV-DSEE algorithms, and [27, 26], which complete the regret analysis of said algorithms and also consider performance under Value at Risk. [30] also consider the Mean-Variance and give a Thompson sampling-based method. Additionally, [1] consider a constrained bandit problem with a concave objective. This may, for example, capture the Mean-Variance setting, however, it is not the main focus of the paper and the results are restricted to order $\mathcal{O}(\sqrt{T})$ regret (or order-$1/\sqrt{T}$ normalized regret in our setting).

Other works consider simpler performance measures that are more closely related to our notion of pseudo regret. [11] present the MaRaB algorithm which uses $CVaR_\alpha$ in its implementation, however, they analyze the average reward performance, and do so under the assumptions that $\alpha = 0$, and the assumption that the $CVaR_\alpha$ and average optimal arms coincide. [23] also consider $CVaR_\alpha$ and give a sample-wise optimistic algorithm. [6] consider the concentration of risk measures and apply it exclusively to bound $CVaR_\alpha$ pseudo regret. It should be noted that some of their findings, and in particular the pseudo regret bound, where previously established in [8], which is an antecedent to the present paper. [31] consider criteria based on the mean and variance of distributions, and present and analyze the $\varphi - LCB$ algorithm. We note that these criteria correspond to a much narrower class of problems than the ones considered here. [17] presents and analyzes the RA-UCB algorithm which considers the measure of *entropic risk* with a parameter λ.

Slightly farther afield, other works consider the best arm identification or simple regret settings. [13] propose distribution independent algorithms for a linear combination of the mean and $CVaR_\alpha$ measures. [15] consider the estimation of $CVaR_\alpha$ for both light and heavy tailed distributions, and provide an algorithm for best $CVaR_\alpha$ arm identification. [25] consider a general functional of the arm distributions and demonstrate results on Mean-Variance, VaR_α, and $CVaR_\alpha$ best arm and simple regret. [29] consider a Mean-Variance best arm identification. [9] consider a quantile risk constrained setting for best arm identification.

Finally, we mention two alternative settings that consider quantile-based sequential performance measures such as VaR_α and $CVaR_\alpha$. [24] consider the convergence of approximate dynamic programming in Markov Decision Processes (MDP), and [12] consider a stochastic optimization setting with bandit feedback. The approach in our paper is quite distinct from these.

Organization. Throughout, all proofs are provided as a sketch that communicates their key ideas, with the full details deferred to the Appendix. In Section 2 we give initial motivation for our suggested class of performance criteria. In Section 3 we formulate the problem setting, oracle rule, and regret metric under non-cumulative criteria. In Section 4 we provide the main results, and in Section 5 we demonstrate them on well-known risk criteria. We also include some negative examples, which illustrate the implications of violation of the proposed conditions, indicating in some way the necessity of such conditions in achieving the unifying theme in our proposed framework.

2. A Motivating Example

In the standard bandit setting, for a sequence of integrable random variables we are interested in designing a policy π that maximizes the average reward

$U^{\text{ave}}_\pi = \mathbf{E}\big[\sum_{t=1}^T X_{\pi,t}\big]$, or, equivalently, minimizes regret compared to an oracle strategy, which is known to pull a single arm throughout the horizon. It is well known that optimistic strategies achieve optimal performance in this setting. Now, suppose we seek to design a policy π that maximizes $U^{CVaR_\alpha}_\pi = \mathbf{E}\big[\frac{1}{\lceil t\alpha\rceil}\sum_{s=1}^{\lceil t\alpha\rceil} X^*_{\pi,s}\big]$, where $X^*_{\pi,s}$ is the s^{th} order statistic of $(X_{\pi,1},\dots,X_{\pi,t})$. This is known as Conditional Value at Risk ($CVaR_\alpha$) or Expected Shortfall, at percentile level $\alpha\in(0,1)$, and is a widely accepted performance measure from the risk literature.

Question: Can we minimize regret using an optimistic strategy?

To answer this, we first need to understand what constitutes an optimistic strategy, which in turn requires that we further understand the oracle rule, which is aware of the true distributions of the arms. The current formulation of $U^{CVaR_\alpha}_\pi$ presents it as a direct function of the reward sequence. While this is is very intuitive, it is in fact a sequence of mappings (one for each sample size) that do not naturally share a domain; studying this sequence can be challenging. In lieu of that, we first observe that $U^{CVaR_\alpha}_\pi$ may be reformulated in terms of a single function that evaluates the sequence of empirical reward distributions $\hat{F}^\pi_t$, where $\hat{F}^\pi_t(x)$ is the fraction of rewards less than or equal to x. Formally, $U^{CVaR_\alpha}_\pi = \mathbf{E}U^{CVaR_\alpha}\big(\hat{F}^\pi_t\big)$ where

$$U^{CVaR_\alpha}\big(F\big) = \max_{z\in\mathbb{R}} z - \frac{1}{\alpha}\int_{-\infty}^{z} F(x)dx,$$

is the accepted notion for measuring $CVaR_\alpha$ for a random variable X with cumulative distribution function F. In Appendix A we show that essentially any performance measure that is invariant to the order of the reward sequence may be reformulated as $\mathbf{E}U\big(\hat{F}^\pi_t\big)$ for some function U. Importantly, this captures many well-known performance measures such as Mean-Variance, Entropic Risk, Sharpe Ratio, and more.

Having restricted ourselves to the class of performance measures that can be expressed this way, we have obtained a consistent way to evaluate performance, which is independent of the time t. Our study now turns to investigating the properties of the function U, in conjunction with the arm distributions, that allow for an optimistic strategy. In the case of $CVaR_\alpha$, if we only assume sub-Gaussian arm distributions, we find that U^{CVaR_α} may be non-smooth, and our framework can only guarantee a regret of $\mathcal{O}(1/\sqrt{T})$ using an optimistic strategy. However, if for example the arm distributions have positive density around their α percentile then we show that the same strategy only incurs $\mathcal{O}(\log T/T)$ regret. The remainder of this paper will flesh out these ideas and provide a set of easy to verify conditions that yield the results discussed in this section.

3. Problem Formulation

Model and admissible policies. Consider a standard MAB with $\mathbb{K}=\{1,\dots,K\}$, the set of arms. Arm $i\in\mathbb{K}$ is associated with a sequence $X_{i,t}$ $(t\geq 1)$ of *i.i.d* random variables with distribution $F^{(i)}\in\mathcal{D}$, the set of all distributions on the real line. When pulling arm i for the t^{th} time, the decision maker receives reward $X_{i,t}$, which is independent of the remaining arms, i.e., the variables $X_{i,t}$ (for all $i\in\mathbb{K}, t\geq 1$) are mutually independent.

We define the set of *admissible* policies (strategies) of the decision maker in the following way. Let $\tau_i(t)$ be the number of times arm i was pulled up to time t. Let V be a random variable over a probability space $(\mathbb{V},\mathcal{V},P_v)$ which is independent of the rewards. An *admissible* policy $\pi=(\pi_1,\pi_2,\dots)$ is a random process recursively defined by

$$\pi_t := \pi_t\big(V,\pi_1,\dots,\pi_{t-1},X_{\pi,1},\dots,X_{\pi,t-1}\big) \tag{1}$$

$$\tau_i(t) = \sum_{s=1}^{t}\mathbb{1}\{\pi_s = i\} \tag{2}$$

$$X_{\pi,t} := X_{i,\tau_i(t)},\text{ given the event } \{\pi_t = i\}. \tag{3}$$

We denote the set of *admissible* policies by Π, and note that *admissible* policies π are non anticipating, i.e., depend only on the past history of actions and observations, and allow for randomized strategies via their dependence on V. Formally, let $\{\mathcal{H}_t\}_{t=0}^{\infty}$ be the filtration defined by $\mathcal{H}_t = \sigma(V, \pi_1, X_{\pi,1}, \ldots, \pi_t, X_{\pi,t})$, then π_t is $\mathcal{H}_{t-1}$ measurable.

Empirical Distribution Performance Measures (EDPM). The classical bandit optimization criterion centers on the *empirical mean*, i.e., $\frac{1}{t}\sum_{s=1}^{t} X_{\pi,s}$. We generalize this by considering criteria that are based on the *empirical distribution*. Formally, the *empirical distribution* of a real number sequence $x_1, \ldots, x_t$ is obtained through the mapping $\hat{F}_t : \mathbb{R}^t \to \mathcal{D}$, given by,

$$\hat{F}_t(x_1, \ldots, x_t; \cdot) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}_{[x_s,\infty]}(\cdot), \tag{4}$$

where $\mathbb{I}_{[a,b]}(\cdot)$ is the indicator function of the interval $[a,b]$ defined on the extended real line, i.e.

$$\mathbb{I}_{[a,b]}(y) = \begin{cases} 1 & , y \in [a,b] \\ 0 & , y \notin [a,b]. \end{cases}$$

Of particular interest to this work are the empirical distributions of the reward sequence under policy π, and of arm i. We denote these respectively by,

$$\hat{F}_t^{\pi}(\cdot) := \hat{F}_t(X_{\pi,1}, \ldots, X_{\pi,t}; \cdot) \tag{5}$$

$$\hat{F}_t^{(i)}(\cdot) := \hat{F}_t(X_{i,1}, \ldots, X_{i,t}; \cdot). \tag{6}$$

The decision maker possesses a function $U : \mathcal{D} \to \mathbb{R}$, which measures the “quality” of a distribution. The resulting criterion is called EDPM, and the decision maker aims to maximize $\mathbf{E}U\left(\hat{F}_T^{\pi}\right)$. (Throughout it is assumed implicitly that the class of distributions and performance functions is such that this expectation exists and is finite valued.)

Oracle and regret. For given horizon T, the oracle policy $\pi^*(T) = (\pi_1^*(T), \pi_2^*(T), \ldots)$ is one that achieves optimal performance given full knowledge of the arm distributions $F^{(i)}$ $(i \in \mathbb{K})$. Formally, it satisfies

$$\pi^*(T) \in \arg\max_{\pi \in \Pi} \mathbf{E}\left[U\left(\hat{F}_T^{\pi}\right)\right]. \tag{7}$$

Similarly to the classic bandit setting, we define a notion of regret that compares the performance of policy π to that of $\pi^*(T)$. The expected (normalized) regret of policy $\pi \in \Pi$ at time T is given by,

$$R_{\pi}(T) := \mathbf{E}\left[U\left(\hat{F}_T^{\pi^*(T)}\right) - U\left(\hat{F}_T^{\pi}\right)\right]. \tag{8}$$

We note that this definition is normalized with respect to the horizon T, thus transforming familiar regret bounds such as $\mathcal{O}(\log T)$ into $\mathcal{O}(\frac{\log T}{T})$. With that convention we smply refer to the above as the “regret” without added qualifiers.

Assumptions. Beyond the existence and finiteness of expectations, flagged earlier, we make the following assumption throughout. Let the simplex in $\mathbb{R}^K$ be

$$\Delta = \left\{ p = (p_1, \ldots, p_K) \in \mathbb{R}^K \;\middle|\; \textstyle\sum_{i=1}^{K} p_i = 1,\, p_i \geq 0\, \forall i \in \mathbb{K} \right\},$$

and define the set of all convex combinations of the arms’ reward distributions by

$$\mathcal{D}^{\Delta} = \left\{ F_p = \textstyle\sum_{i=1}^{K} p_i F^{(i)} \;\middle|\; p \in \Delta \right\}. \tag{9}$$

Let $i^* \in \arg\max U\left(F^{(i)}\right)$ be an "optimal" arm. We assume that

$$U\left(F\right) \leq U\left(F^{(i^*)}\right) \quad ,\forall F \in \mathcal{D}^{\Delta}.$$

While there is a potential loss of generality here, we could not find any interesting performance measure that violates this inequality. In particular, it provably holds when U is quasiconvex, which will always be the case in our examples.

4. Main Results

When defining an objective, it was sufficient to consider U as a mapping from $\mathcal{D}$ (a *set*) to $\mathbb{R}$. Moving forward, our analysis relies on properties such as continuity and differentiability, which require that we consider U as a mapping between seminormed spaces. To that end $\mathcal{D}$ is a subset of an infinite dimensional vector space for which norm equivalence does not hold. This hints at the importance of using the "correct" (semi)norm for each U. As a result, our analysis is done with respect to a general seminorm $\|\cdot\|$ and its matching seminormed space $L_{\|\cdot\|}$. We therefore consider EDPMs as mappings $U : L_{\|\cdot\|} \to \mathbb{R}$.

The goal of this work is to provide a generic analysis of the regret, similar to that of the classical bandit setting, and which culminates in the following result.

Theorem (Informal meta-result). *There exists an efficient algorithm such that:*

1. *Under suitable regularity conditions obtains regret* $R_\pi(T) = \mathcal{O}(\frac{\log T}{\sqrt{T}})$;
2. *Under an additional smoothness condition obtains regret* $R_\pi(T) = \mathcal{O}(\frac{\log T}{T})$.

In what follows, we introduce the technical details required to make this statement rigorous. This culminates in Section 4.5, where Theorem 4 gives the desired statement, and where we also explain how the standard bandit setting fits into our framework. Unlike the classical bandit setting, the oracle policy $\pi^*(T)$, defined in (7), need not choose a single arm. Since the typical learning algorithms are structured to emulate the oracle rule, we need to first understand the structure of the oracle policy before we can analyze $R_\pi(T)$.

4.1. Insights From the Infinite Horizon Oracle

The oracle problem in (7) does not admit a tractable solution, in the absence of further structural assumptions. In this section we consider a *relaxation* of the oracle problem which examines asymptotic behavior. We provide conditions under which this behavior is "simple" thus suggesting it as a proxy for the finite time performance. More concretely, let $U_\pi = \liminf_{t\to\infty} U\left(\hat{F}_t^\pi\right)$ be the *worst case* asymptotic performance of policy π, then the infinite horizon oracle $\pi^*(\infty) = (\pi_1^*(\infty), \pi_2^*(\infty), \ldots)$ satisfies

$$\pi^*(\infty) \in \arg\max_{\pi\in\Pi} \mathbf{E}[U_\pi]. \tag{10}$$

Note that U_π is well defined as the limit inferior of a sequence of random variables, however (as indicated earlier) we require that its expectation exist for (10) to be well defined.

Simple oracle. In the traditional Multi-Armed Bandit problem, the oracle policy, which selects a single arm throughout the horizon, is clearly simple. It may seem intuitive that EDPMs always admit a such a simple infinite horizon oracle policy. However, in (F.3.4) we give counter examples, which arise from the "bad behavior" that is still allowed by this objective. The following result gives sufficient conditions for EDPMs to be "well behaved."

Theorem 1 (EDPM admits a simple oracle policy). *Suppose an EDPM, $U : L_{\|\cdot\|} \to \mathbb{R}$, is continuous on $\mathcal{D}^{\Delta}$, and that $\lim_{t\to\infty} \|\hat{F}_t^{(i)} - F^{(i)}\| = 0$ almost surely for all $i \in \mathbb{K}$. Then the single arm policy that always chooses i^* is an infinite horizon oracle policy, as defined in (10).*

Proof sketch. (see full details in Appendix B) First, define that arm pulling ratio $\hat{p}_i(t) = \tau_i(t)/t$, and notice that the empirical distribution may be written as $\hat{F}_t^\pi = \sum_{i=1}^K \hat{p}_i(t)\hat{F}_t^{(i)}$. Since $p(t) \in \Delta$, which is closed and compact, we have that any subsequence of t has a further subsquence, t_l, such that $p(t_l) \to p \in \Delta$. Next, since we assumed that $\lim_{t\to\infty}\|\hat{F}_t^{(i)} - F^{(i)}\| = 0$ almost surely, we conclude that $\hat{F}_{t_l}^\pi \to F_p$. Applying the continuity assumption we conclude that $U_\pi = \liminf_{t\to\infty} U\left(\hat{F}_t^\pi\right) \le \lim_{l\to\infty} U\left(\hat{F}_{t_l}^\pi\right) = U\left(F_p\right) \le U\left(F^{(i^*)}\right)$. We conclude the proof by showing that similar arguments imply that the proposed simple oracle policy achieves this upper bound.

Remark. Theorem 1 depends not only on U but also on the given distributions $F^{(i)}$. Meaning, it may hold for a given U only for some distributions, and thus the choice of a seminorm is important in order to get sharp conditions on the viable reward distributions. For example, consider the supremum norm given by $\|f\|_\infty = \sup_{x\in\mathbb{R}}|f(x)|$. By the Glivenko-Cantelli theorem ([28]), it satisfies the convergence condition for any given distributions $F^{(i)}$, $i \in \mathbb{K}$. However, in most cases, continuity holds only if the distributions have bounded support.

4.2. Regret Decomposition

Having gained some understanding of the infinite horizon oracle, we consider a regret decomposition that uses the infinite horizon performance as a benchmark. Let

$$F_T^\pi = \frac{1}{T}\sum_{t=1}^T F^{(\pi_t)} = \frac{1}{T}\sum_{i=1}^K \tau_i(T)F^{(i)}, \tag{11}$$

be the pseudo empirical distribution, where we recall that $F^{(i)}$ is the distribution associated with arm $i \in \mathbb{K}$, and i^* is such that $U\left(F\right) \le U\left(F^{(i^*)}\right)$, for all $F \in \mathcal{D}^\Delta$. The regret may now be decomposed as

$$\begin{aligned} R_\pi(T) = &\underbrace{\mathbf{E}\Big[U\left(\hat{F}_T^{\pi^*(T)}\right) - U\left(F_T^{\pi^*(T)}\right)\Big]}_{J_1(T)} \\ &+ \underbrace{\mathbf{E}\Big[U\left(F_T^{\pi^*(T)}\right) - U\left(F^{(i^*)}\right)\Big]}_{J_2(T)} \\ &+ \underbrace{\mathbf{E}\Big[U\left(F^{(i^*)}\right) - U\left(F_T^\pi\right)\Big]}_{\bar{R}_\pi(T)} \\ &+ \underbrace{\mathbf{E}\Big[U\left(F_T^\pi\right) - U\left(\hat{F}_T^\pi\right)\Big]}_{J_3(T)}. \end{aligned} \tag{12}$$

The term $\bar{R}_\pi(T)$ represents what we believe to be the correct notion of pseudo regret in our setting. Unlike the standard bandit pseudo regret $\frac{1}{T}\mathbf{E}\left[\sum_{t=1}^T U\left(F^{(i^*)}\right) - U\left(F^{(\pi_t)}\right)\right]$, which aggregates a policy's decisions in reward space, here aggregation occurs in distribution space and then evaluates to a reward via U. We note that in the standard average reward setting U is linear and both notions coincide. However, in the general non-linear setting, the previous notion may underestimate the regret and is thus unsatisfactory.

The remaining terms in (12) may be viewed as a decomposition into error terms that measure discrepancy between distributions via the criterion U. As hinted at by our notation, these may be bounded essentially independently of the learning algorithm's policy π. The term $J_2(T)$ measures the difference between the optimal infinite horizon arm pulling mixture, which is a single arm, and that of the finite horizon oracle. Notice that by definition of i^* we always have that $J_2(T) \le 0$. The terms $J_1(T)$ and $J_3(T)$, which we refer to as horizon gaps, measure the convergence of empirical distributions to their appropriate infinite horizon counterparts, which are given by pseudo empirical distributions. While bounding these proves to be the crux of our problem, we begin by proposing a learning algorithm that minimizes the pseudo regret.

4.3. Learning Algorithm

Theorem 1 presented conditions for understanding the asymptotic behavior of performance. As we now seek a finite time analysis (of the pseudo regret), it stands to reason to employ the following stronger conditions, which quantify the rate of convergence.

Definition 1 (**Stable EDPM**). We say that U is stable with respect to a seminorm $\|\cdot\|$ if there exist $\upsilon, b > 0, q \geq 1$ such that:

1. U admits $\omega(x) = b(x + x^q)$ as a local modulus of continuity for all $F \in \mathcal{D}^\Delta$, i.e.,

$$|U(F) - U(G)| \leq \omega(\|F - G\|), \qquad \forall F \in \mathcal{D}^\Delta, G \in L_{\|\cdot\|}.$$

2. Recalling $\hat{F}_t^{(i)}$ from (6), we have that for all $i \in \mathbb{K}$, $t \geq 1$

$$\mathbb{P}\left(\|\hat{F}_t^{(i)} - F^{(i)}\| \geq x\right) \leq 2\exp\left(-\upsilon t x^2\right), \qquad \forall x > 0.$$

Pseudo regret decomposition. In the traditional bandit setting, which considers the average reward, the analysis of the regret is well understood. The same analysis extends to any linear EDPM, i.e., when U is linear. This follows straightforwardly as such rewards can be formulated as the usual average criterion with augmented arm distributions. Linearity facilitates the regret analysis by providing a decomposition of contributions from each sub-optimal arm. Define the standard single arm sub-optimality gap

$$\Delta_i = U\left(F^{(i^*)}\right) - U\left(F^{(i)}\right),$$

where we recall that $i^* \in \arg\max U\left(F^{(i)}\right)$ is the optimal arm. The regret of a linear EDPM is given by, $R_\pi(T) = \frac{1}{T}\sum_{i \neq i^*} \Delta_i \mathbf{E}\tau_i(T)$. Departing from the simple realm of linearity, we seek a similar decomposition of the pseudo regret. To that end, denote the diameter of $\mathcal{D}^\Delta$ and the maximum gap ratio respectively as

$$D = \max_{i,j \in \mathbb{K}} \|F^{(i)} - F^{(j)}\| \qquad \rho = \max_{i \in \mathbb{K}} \|F^{(i^*)} - F^{(i)}\| / \Delta_i, \tag{13}$$

where the latter essentially measures how well the chosen seminorm captures the sub-optimality gaps. We provide the following result, which is proved in Appendix C.

Lemma 1 (**Pseudo regret decomposition**). *Let U be a stable EDPM, then U is L-Lipschitz over $\mathcal{D}^\Delta$ with $L = b(1 + D^{q-1})$, and we have that*

$$\bar{R}_\pi(T) \leq \frac{L\rho}{T} \sum_{i \neq i^*} \Delta_i \mathbf{E}\tau_i(T).$$

Learning algorithm. We present $U - UCB$, a natural adaptation of $(\alpha, \psi) - UCB$ (see [7]) to a stable EDPM. Let,

$$\phi(y) = \min\left\{\upsilon\left(\frac{y}{2b}\right)^2, \upsilon\left(\frac{y}{2b}\right)^{2/q}\right\}$$
$$\phi^{-1}(x) = \max\left\{2b\left(\frac{x}{\upsilon}\right)^{1/2}, 2b\left(\frac{x}{\upsilon}\right)^{q/2}\right\},$$

where υ, b, q are the parameters of Definition 1. The $U - UCB$ policy is given by,

$$\pi_t^{U-UCB} \in \arg\max_{i \in \mathbb{K}} \left[U\left(\hat{F}^{(i)}_{\tau_i(t-1)}\right) + \phi^{-1}\left(\frac{\alpha \log t}{\tau_i(t-1)}\right)\right], \quad t \geq K + 1, \tag{14}$$

where for $1 \leq t \leq K$, it samples each arm once as initialization.

Theorem 2 (**U − UCB Pseudo Regret**). *Let U be a stable EDPM. If $\Delta_i > 0$ for all $i \neq i^*$, then for L defined in Lemma 1 and $\alpha > 2$ we have that*

$$\bar{R}_{U-UCB}(T) \leq \frac{L\rho}{T} \sum_{i \neq i^*} \left(\frac{\alpha \Delta_i \log T}{\phi(\Delta_i/2)} + \frac{\alpha+6}{\alpha-2}\Delta_i \right).$$

Proof sketch. (see full details in Appendix C) The proof uses standard techniques from the UCB literature and consists of the following steps. First, we show that if at round t, the algorithm chooses sub-optimal arm i that was pulled more than order-$\log T$ times, then we have significantly overestimated this arm and underestimated the optimal arm. Second, we bound the probability of this estimation failure using standard concentration arguments that are deduced from stability (Definition 1). We conclude that after choosing sub-optimal arm i for order-$\log T$ times, the probability of choosing it again is very small, and thus the expected number of sub-optimal arm pulls is order-$\log T$. Finally, the proof is concluded by plugging this result into the pseudo regret decomposition, given in Lemma 1.

4.4. Bounding the Horizon Gaps

Recall that the horizon gap of a policy $\pi \in \Pi$ is given by $\left|\mathbf{E}\left[U\left(\hat{F}_T^\pi\right) - U\left(F_T^\pi\right)\right]\right|$. At their core, our bounds on the horizon gaps follow from the convergence of the empirical to pseudo empirical distribution. This convergence is quantified by the following lemma.

Lemma 2 (**Empirical Distribution Tail Bound**). *Suppose that Requirement 2 of stability holds, then*

$$\mathbb{P}\left(\|\hat{F}_T^\pi - F_T^\pi\| > x\right) \leq 2KT \exp\left(-\upsilon \frac{Tx^2}{K^2}\right) \qquad , \forall T \geq 1, x \geq 0.$$

The proof decomposes the deviation into its contributions from each arm and uses Requirement 2 of stability, single arm concentration of the empirical distribution, and several union bounds to conclude the desired result. See full details in Section D.1. We now present our first bound on the horizon gap, which is uniform over all policies $\pi \in \Pi$.

Proposition 1 (**Uniform horizon gap bound**). *Suppose that U is a stable EDPM. Then we have that for all $T \geq \frac{4qK^2 \log KT}{\upsilon}$*

$$\left|\mathbf{E}\left[U\left(\hat{F}_T^\pi\right) - U\left(F_T^\pi\right)\right]\right| \leq 4b\left(\frac{K^2 \log KT}{\upsilon T}\right)^{1/2} \qquad , \forall \pi \in \Pi.$$

The proof uses Requirement 1 of stability, local modulus of continuity, to get that

$$\left|\mathbf{E}\left[U\left(\hat{F}_T^\pi\right) - U\left(F_T^\pi\right)\right]\right| \leq \mathbf{E}\left[\omega\left(\|\hat{F}_T^\pi - F_T^\pi\|\right)\right] = b\mathbf{E}\left[\|\hat{F}_T^\pi - F_T^\pi\| + \|\hat{F}_T^\pi - F_T^\pi\|^q\right], \tag{15}$$

and then uses the tail sum formula together with the tail bound in Lemma 2 to obtain the final bound. See full details in Appendix D.

The result of Proposition 1 exhibits a dependence on the time horizon T which may be quite loose. To see this, consider a linear U. It is relatively easy verify that the left hand side of (15) is zero, while its right hand side behaves as $1/\sqrt{T}$ even when $K = 1$.

In order to obtain improved bounds, we require a notion of smoothness. Formally, let $L\left(L_{\|\cdot\|}, \mathbb{R}\right)$ be the space of bounded linear functionals on $L_{\|\cdot\|}$. Assuming U is differentiable on $F \in \mathcal{D}^\Delta$, then its differential $\partial U\left(F\right) \in L\left(L_{\|\cdot\|}, \mathbb{R}\right)$ is well defined, and we denote its (linear) operation on $G \in L_{\|\cdot\|}$ by $\partial U\left(F\right) \cdot G$.

Definition 2 (Smooth EDPM). An EDPM, $U : L_{\|\cdot\|} \to \mathbb{R}$, is smooth if it is differentiable on $\mathcal{D}^{\Delta}$, and there exist $\beta \geq 0, M_0 > 0$ such that for any $F \in \mathcal{D}^{\Delta}, G \in L_{\|\cdot\|}$ satisfying $\|G - F\| \leq M_0$ we have that

$$|U(G) - U(F) - \partial U(F) \cdot (G - F)| \leq \frac{1}{2}\beta \|G - F\|^2$$

This definition is a standard notion from optimization, stated for our infinite dimensional function space. The following result shows that "reasonable" policies enjoy a smaller horizon gap under the smoothness assumption.

Theorem 3 (Improved horizon gap bound). *Suppose that U is a stable and smooth EDPM with $M_0 = \infty$. Letting $J_T^2 = \frac{4K^2 \log KT}{\upsilon T}$, we have that for any policy $\pi \in \Pi$ and fixed $F \in \mathcal{D}^{\Delta}$*

$$\left|\mathbf{E}\left[U\left(\hat{F}_T^{\pi}\right) - U\left(F_T^{\pi}\right)\right]\right| \leq \beta J_T^2 + \beta J_T \mathbf{E}\|F_T^{\pi} - F\|,$$

for all $T \geq T_0$, which is polynomial in problem parameters.

The proof may be found in Appendix D. It explicitly identifies the parameter T_0, and also addresses the case of $M_0 < \infty$, which gives an additional low order term. Notice that $J_T^2 = \mathcal{O}(\log T/T)$ and thus any policy whose arm pull frequencies converge in expectation at a rate of $\mathcal{O}(\sqrt{\log T/T})$ has horizon gap of $\mathcal{O}(\log T/T)$. In particular, this clearly holds for $U - UCB$.

Proof (sketch). First, a simple calculation shows that $\mathbf{E}\hat{F}_t^{\pi} = \mathbf{E}F_t^{\pi}$. Next, since ∂U is a linear operator, we conclude that $\mathbf{E}\left[\partial U(F) \cdot (\hat{F}_T^{\pi} - F_T^{\pi})\right] = 0$ for all $F \in \mathcal{D}^{\Delta}$. We thus obtain the following decomposition

$$\begin{aligned}\left|\mathbf{E}\left[U\left(\hat{F}_T^{\pi}\right) - U\left(F_T^{\pi}\right)\right]\right| \leq\ & \mathbf{E}\Big|\underbrace{U\left(\hat{F}_T^{\pi}\right) - U\left(F_T^{\pi}\right) - \partial U\left(F_T^{\pi}\right) \cdot \left(\hat{F}_T^{\pi} - F_T^{\pi}\right)}_{\delta_1}\Big| \\ & + \mathbf{E}\Big|\underbrace{\left(\partial U\left(F_T^{\pi}\right) - \partial U(F)\right) \cdot \left(\hat{F}_t^{\pi} - F_T^{\pi}\right)}_{\delta_2}\Big|,\end{aligned}$$

To bound $\mathbf{E}|\delta_1|$ we use smoothness to get that

$$\mathbf{E}|\delta_1| \leq \frac{1}{2}\beta \mathbf{E}\left[\|\hat{F}_T^{\pi} - F_T^{\pi}\|^2\right],$$

and using the tail sum formula together with the tail bound in Lemma 2 bounds $\mathbf{E}|\delta_1|$.

Finally, to bound $\mathbf{E}|\delta_2|$ we first use the Cauchy–Schwarz inequality together with smoothness to get that

$$\mathbf{E}|\delta_2| \leq \mathbf{E}\left[\|\partial U\left(F_T^{\pi}\right) - \partial U\left(F_{\gamma}\right)\| \|\hat{F}_T^{\pi} - F_T^{\pi}\|\right] \leq \beta \mathbf{E}\left[\|F_T^{\pi} - F\| \|\hat{F}_T^{\pi} - F_T^{\pi}\|\right].$$

Now, for small deviations we have that

$$\mathbf{E}\left[|\delta_2| \mathbb{1}\left\{\|\hat{F}_T^{\pi} - F_T^{\pi}\| \leq J_T\right\}\right] \leq \beta J_T \mathbf{E}\|F_T^{\pi} - F\|.$$

On the other hand, recalling that D is the diameter of $\mathcal{D}^{\Delta}$, we get that

$$\mathbf{E}\left[|\delta_2| \mathbb{1}\left\{\|\hat{F}_T^{\pi} - F_T^{\pi}\| > J_T\right\}\right] \leq \beta D \mathbf{E}\left[\|\hat{F}_T^{\pi} - F_T^{\pi}\| \mathbb{1}\left\{\|\hat{F}_T^{\pi} - F_T^{\pi}\| > J_T\right\}\right],$$

and using the tail sum formula together with Lemma 2, and summing the two inequalities bounds $\mathbf{E}|\delta_2|$, and concludes the proof. ■

4.5. Regret Bound In order to conclude our regret bounds, we require the following definition.

Definition 3 (**Linear gap**). We say U has a linear gap if there exists $\eta > 0$ such that

$$U\left(F\right) \le U\left(F^{(i^*)}\right) - \frac{1}{\eta}\|F - F^{(i^*)}\| \qquad , \forall F \in \mathcal{D}^{\Delta}.$$

This assumption will be seen to hold for our examples, in particular, the following result gives mild sufficient conditions. See proof in Appendix E.

Proposition 2. *Suppose U is convex over $\mathcal{D}^{\Delta}$ and $\Delta_i > 0$ for all $i \ne i^*$, then U has a linear gap with $\eta = \rho$, where ρ is defined in (13).*

Summarizing our observations thus far, the informal statement of our main findings is made concrete by the following result.

Theorem 4 (**$U-$UCB regret**). *Suppose an EDPM, U, is stable, smooth with $M_0 = \infty$, and has a linear gap. Letting $L = b(1 + D^{q-1})$, the regret of running $U-UCB$ is bounded as*

$$R_{U-UCB}(T) \le \frac{L\rho}{T}\sum_{i \ne i^*}\left(\frac{\alpha\Delta_i \log T}{\phi(\Delta_i/2)} + \frac{\alpha+6}{\alpha-2}\Delta_i\right) + \frac{12\beta K^2 \log KT}{\upsilon T},$$

for all $T \ge T_1$, which is polynomial in problem parameters.

For a detailed proof, which includes the exact dependence of T_1 on the problem parameters, and an additional low order term when $M_0 < \infty$, see Appendix E.

Proof (sketch). First, using Theorem 3 with $F = F^{(i^*)}$ and large enough T we get that

$$J_1(T) = \mathbf{E}\left[U\left(\hat{F}_T^{\pi^*(T)}\right) - U\left(F_T^{\pi^*(T)}\right)\right] \le \frac{4\beta K^2 \log KT}{\upsilon T} + \frac{1}{\eta}\mathbf{E}\|F_T^{\pi^*(T)} - F^{(i^*)}\|,$$
$$J_3(T) = \mathbf{E}\left[U\left(F_T^{U-UCB}\right) - U\left(\hat{F}_T^{U-UCB}\right)\right] \le \frac{6\beta K^2 \log KT}{\upsilon T},$$

where for $J_3(T)$ we bounded the second term of Theorem 3 using the fact that Theorem 2 actually bounds $L\mathbf{E}\|F_T^{U-UCB} - F^{(i^*)}\|$. Second, using the linear gap assumption we get that

$$J_2(T) = \mathbf{E}\left[U\left(F_T^{\pi^*(T)}\right) - U\left(F^{(i^*)}\right)\right] \le -\frac{1}{\eta}\mathbf{E}\|F_T^{\pi^*(T)} - F^{(i^*)}\|.$$

Finally, recall that in (12) we decompose the regret as $R_\pi(T) = \bar{R}_\pi(T) + J_1(T) + J_2(T) + J_3(T)$. Combining the above and using Theorem 2 to bound $\bar{R}_{U-UCB}(T)$ concludes the proof. ■

Remark 1. Notice that even in the absence of the smoothness and linear gap assumptions, we may still apply Proposition 1 to obtain a weaker regret bound in which the last two terms of Theorem 4 are replaced by $8b\sqrt{K^2 \log KT/\upsilon T}$. In Section 5 we will show that typical examples satisfy Theorem 4, however, we also give two cases where this weaker bound is the best that can be achieved within our framework.

Example: Average Reward We summarize our approach for the familiar bandit average reward setting that, in our EDPM formulation, is given by $U^{\text{ave}}\left(F\right) = \int_{\mathbb{R}} x dF(x)$. Notice that U^{ave} is linear and so the regret decomposition in (12) becomes trivial, i.e., $J_1 = J_2 = J_3 = 0$ and $R_\pi(T) = \bar{R}_\pi(T)$. For simplicity, suppose that the rewards are constrained to the interval $[0,1]$, and consider the seminorm $\|F\| = |U^{\text{ave}}(F)|$. Notice that $\Delta_i = \|F^{(i^*)} - F^{(i)}\|$, and thus $\rho = 1$. Using Hoeffding's inequality we get that U^{ave} is stable with $\upsilon = 2, b = 1/2, q = 1$, and thus $L = 1$. Since U^{ave} is linear, it is clearly smooth with $\beta = 0$ and $M_0 = \infty$. Plugging all parameters into Theorem 4 we recover the standard regret bound for average reward bandits

$$R_{U-UCB}(T) \leq \frac{1}{T} \sum_{i \neq i^*} \left(\frac{2\alpha \log T}{\Delta_i} + \Delta_i \frac{\alpha + 6}{\alpha - 2} \right). \tag{16}$$

Discussion Our main result demonstrates that the EDPM formulation allows us to convert difficult questions in learning under sequential performance criteria to, essentially, simple questions in functional analysis. Roughly speaking, any quasiconvex function, U, that, for an appropriate seminorm, is twice differentiable and grows at most polynomially, can be be accommodated by our framework (at least for some arm distributions), i.e., may be learned efficiently. We note that our results focused solely on the time horizon parameter T, and we suspect that the dependence on the number of arms K can be improved. The main issue there is the squared dependence on K in the tail bound of the empirical distribution (Lemma 2), and subsequently in the horizon gap (Theorem 3). It is not clear whether this could be improved uniformly over all policies $\pi \in \Pi$, or whether this should be done only for near optimal policies. As a motivating example, it is clear that a single arm policy has horizon gap that does not depend on K. As the optimal policy is close to a single arm policy, we expect that its dependence on K should be weak, perhaps even sub-linear. This would make the horizon gap a low order term compared to the pseudo regret and establish an equivalence between the two notions of regret. We leave this as an open question for future work. So far, we demonstrated how the standard average reward setting fits into our framework. In the following section we use our framework to analyze various well-known performance measures from the risk literature.

5. Illustrative Examples

The purpose of this section is, first and foremost, to show the relative ease with which various performance criteria can be analyzed within the framework developed in the previous sections. To make the exposition more accessible, we forego detailed introductions of the various criteria as well as various other technical details. Our main focus is to show the use of Theorem 4, after which we give some edge cases that demonstrate the subtleties and limitations of our framework. We refer the interested reader to Appendix F for the complete details. We make the following assumption throughout this section.

Assumption 1. The rewards are restricted to the interval $[0,1]$, i.e., the support of $F^{(i)}$ is in $[0,1]$ for all $i \in \mathbb{K}$.

This assumption is intended to simplify the exposition and can always be replaced by an appropriate "light tailed" condition.

5.1. Linear EDPMs

We begin with with a few examples of linear EDPMs, which are essentially standard stochastic multi-armed bandit settings with augmented arm distributions.

Average reward is the classic bandit performance criterion, which is given by $U^{\text{ave}}(F) = \int_{\mathbb{R}} x dF(x)$. For a given reward sequence, it is explicitly stated as

$$U^{\text{ave}}(\hat{F}_T^{\pi}) = \frac{1}{T} \sum_{t=1}^{T} X_{\pi,t}.$$

Squared reward is a less typical performance measure on its own but will serve us in what follows. It is given by $U^{\text{sqr}}(F)=\int_{\mathbb{R}} x^2 dF(x)$, and in terms of the reward sequence as

$$U^{\text{sqr}}(\hat{F}_T^\pi)=\frac{1}{T}\sum_{t=1}^{T}(X_{\pi,t})^2.$$

Below Target Semi-Variance (TSV) measures the negative variation from a threshold parameter $r\in\mathbb{R}$. The goal here is to minimize the variation and since our setting is expressed in terms of maximization, we state its negation $U^{\text{TSV}}(F)=-\int_{\mathbb{R}}(x-r)^2\mathbb{1}\{x\le r\}dF(x)$. In terms of the reward sequence this stated as

$$U^{\text{TSV}}(\hat{F}_T^\pi)=\frac{1}{T}\sum_{t=1}^{T}(X_{\pi,t}-r)^2\mathbb{1}\{X_{\pi,t}\le r\}.$$

The Analysis for all linear EDPMs follows in a similar fashion to the average reward demonstrated in Section 4.5, with the only potential change being the value of υ. More formally, let U^{lin} be any linear EDPM. We define the seminorm $\|F\|=|U^{\text{lin}}(F)|$, which implies that $\Delta_i=\|F^{(i^*)}-F^{(i)}\|$, and consequently $\rho=1$. Requirement 1 of stability clearly holds with $b=1/2, q=1$, and thus $L=1$. Recalling the empirical distribution and indicator functions from (4), Hoeffding's inequality implies Requirement 2 of stability holds with $\upsilon=2/\vartheta_{\text{lin}}$ where

$$\vartheta_{\text{lin}}=\max_{x,y\in[0,1]}\left[U^{\text{lin}}(\mathbb{I}_{[x,\infty]})-U^{\text{lin}}(\mathbb{I}_{[y,\infty]})\right]^2,$$

is the squared length of the reward interval under U^{lin}, which in the examples above is at most 1. Since U^{lin} is linear, it is clearly smooth with $\beta=0$ and $M_0=\infty$. Plugging all parameters into Theorem 4 we recover the standard bandit regret bound given in (16).

5.2. Composite EDPMs Moving on to more complex performance criteria, we consider compositions of linear EDPMs. Such criteria are often used to state a trade-off between multiple objectives. A partial list of widely used risk metrics which we consider here consists of: Entropic Risk, Variance, Mean-Variance (Markowitz), Sharpe ratio, and Sortino ratio. Formally, we say an EDPM U^h is composite if there exist linear EDPMs $U^{(1)},\dots,U^{(n)}$ and $h:\mathbb{R}^n\to\mathbb{R}$ such that

$$U^h(F)=h(U^{(1)}(F),\dots,U^{(n)}(F)).$$

Considering this class under the seminorm $\|F\|=\|(U^{(1)}(F),\dots,U^{(n)}(F))\|_2$, where $\|\cdot\|_2$ is the ℓ^2 norm in $\mathbb{R}^n$, the verification of our framework becomes easy, as seen in the following result.

Lemma 3 (Informal). *Suppose $U^{(1)},\dots,U^{(n)}$ are linear, and stable, then:*

1. *If h admits a polynomial local modulus of continuity, then U^h is stable;*
2. *If h is locally smooth, then U^h is smooth;*
3. *If h is convex then so is U^h.*

The formal statement along with its proof may be found in Section F.2. Verifying Lemma 3 is typically very easy, often amounting to bounding the gradient and hessian of h, and yields all the needed properties to invoke Theorem 4 and obtain an $\mathcal{O}(\log T/T)$ regret bound for $U-UCB$.

Entropic risk is a risk assessment measure that uses an exponential utility function with risk aversion parameter $\theta>0$. It is given by

$$U^{\text{ent}}(F)=-\frac{1}{\theta}\log\left(\int_{\mathbb{R}}\exp(-\theta x)dF(x)\right)=-\frac{1}{\theta}\log(U^{\exp}(F)),$$

where $U^{\exp}(F)=\int_{\mathbb{R}}\exp(-\theta x)dF(x)$ is a linear EDPM and thus U^{ent} is composite with $h(x)=-\frac{1}{\theta}\log x$, which is convex. Since the rewards are in $[0,1]$, we can bound the derivatives of h to conclude that it satisfies Lemma 8 with $b=\frac{1}{2\theta}\exp(\theta), q=1, \upsilon=2, \beta=\frac{1}{\theta}\exp(2\theta)$, and $M_0=\infty$.

Variance measures the empirical squared deviation from the mean reward. As we seek to minimize this deviation, it is given by

$$U^{\text{var}}\big(F\big) = -\Big[U^{\text{sqr}}\big(F\big) - \big[U^{\text{ave}}\big(F\big)\big]^2\Big],$$

and thus $h\big(x_1, x_2\big) = x_1^2 - x_2$, which is convex. It is then easy to verify that Lemma 8 holds with $b = \sqrt{5}, q = 2, \upsilon = 1/2, \beta = 2$, and $M_0 = \infty$.

Mean-variance (Markowitz) measures performance as an additive trade-off between the empirical mean and variance. For $\rho \geq 0$ it is given by

$$U^{\text{MV}}\big(F\big) = U^{\text{ave}}\big(F\big) + \rho U^{\text{var}}\big(F\big),$$

and thus $h\big(x_1, x_2\big) = x + \rho(x_1^2 - x_2)$, which is convex. A simple calculation shows that Lemma 8 holds with $b = 2(1+\rho), q = 2, \upsilon = 1/2, \beta = 2\rho$, and $M_0 = \infty$.

Sharpe ratio measures performance as a ratio between the empirical mean and variance. For $r \in \mathbb{R}$ and $\varepsilon_0 > 0$ it is given by

$$U^{\text{Sh}}\big(F\big) = \frac{U^{\text{ave}}\big(F\big) - r}{\sqrt{\varepsilon_0 - U^{\text{var}}\big(F\big)}},$$

and thus $h\big(x_1, x_2\big) = (x_1 - r)/\sqrt{\varepsilon_0 - x_1^2 + x_2}$. The parameter r is essentially a threshold for the average reward, while ε_0 is a regularization parameter. Unlike the previous examples, where h was convex, here it is only quasiconvex. While a quasiconvex function could potentially have no linear gap, we show that Sharpe ratio has a linear gap with a slightly worse constant. The calculation is technical and deferred to Appendix F.

Sortino ratio is Sharpe ratio with variance replaced by below target semi-variance. As such, for $r \in \mathbb{R}$ and $\varepsilon_0 > 0$ it is given by

$$U^{\text{So}}\big(F\big) = \frac{U^{\text{ave}}\big(F\big) - r}{\sqrt{\varepsilon_0 - U^{\text{TSV}}\big(F\big)}},$$

and thus $h\big(x_1, x_2\big) = (x_1 - r)/\sqrt{\varepsilon_0 - x_2}$. Similar to Sharpe ratio, here h is also quasiconvex. The resulting analysis is thus similar, if perhaps a bit simpler, and we defer it to Appendix F.

5.3. Non-composite EDPMs We now consider two examples of non-composite criteria. The first, $CVaR_\alpha$, is found to be smooth and stable under appropriate conditions. The second, VaR_α, is stable but appears to be non-smooth. In both cases the resulting conditions possess a more particular nature than those presented for composite EDPMs.

Conditional Value at Risk (CVaR$_\alpha$) is the average reward below percentile level $\alpha \in (0,1)$, which is given by

$$U^{CVaR_\alpha}\big(F\big) = \max_{z \in \mathbb{R}} z - \frac{1}{\alpha}\int_{-\infty}^{z} F(x)dx. \tag{17}$$

We note that a more explicit expression can be obtained by plugging in the maximizer $z^* = U^{VaR_\alpha}\big(F\big)$, which is the Value at Risk of F, defined in (20). Now, in order to invoke Theorem 4 we need to show that $CVaR_\alpha$ is convex, stable and smooth.

Convexity is immediate since the expression in (17) is a maximum over linear functions, which is convex. For stability, we use the norm

$$\|F\| = \max\left\{\|F\|_\infty, \left|\int_{-\infty}^{0} x dF(x)\right|, \left|\int_{0}^{\infty} x dF(x)\right|\right\}, \tag{18}$$

where $\|F\|_\infty = \max_{x\in\mathbb{R}}|F(x)|$ is the ℓ^∞ norm. The two additional terms may be foregone when the rewards are constrained to $[0,1]$. The concentration of $\|F\|_\infty$ follows from the Dvoretzky-Kiefer-Wolfowitz inequality [19], while the other two follow from Hoeffding's inequality. A further technical calculation yields the desired modulus of continuity and thus stability is concluded with parameters $b = 4/\alpha \min\{\alpha, 1-\alpha\}, q = 2, \upsilon = 2/3$. Recalling Remark 1, we can now conclude that $U - UCB$ obtains regret $\mathcal{O}(\sqrt{\log T/T})$ for any bounded reward distributions.

We would like to show that $CVaR_\alpha$ is smooth and thus conclude the requirements for Theorem 4. However, we find that even for bounded distributions, $CVaR_\alpha$ may be non-smooth (see Figure 1). To overcome this limitation, we require that the arm distributions have a positive density around their α percentile. Formally, we require that there exist $b_\alpha > 0, M_\alpha \geq D$ such that

$$|F\big(U^{VaR_\alpha}(F) + b_\alpha y\big) - \alpha| \geq |y| \qquad , \forall F \in \mathcal{D}^\Delta, y \in [-M_\alpha, M_\alpha], \tag{19}$$

where $U^{VaR_\alpha}(F)$ is the Value at Risk of F, defined in (20). This condition is one of the subtleties that arise from our framework. With it, we show that $CVaR_\alpha$ is smooth with $\beta = 2b_\alpha/\alpha$ and $M_0 = M_\alpha$ and thus obtain the desired $\mathcal{O}(\log T/T)$ regret bound. Without it, we provide a numerical experiment (see Figure 1), suggesting that the horizon gap truly behaves as $\Omega(\sqrt{1/T})$ as suggested by our framework.

Value at Risk (*VaR$_\alpha$*) is the reward at percentile $\alpha \in (0,1)$, which is given by

$$U^{VaR_\alpha}(F) = \inf_{x\in\mathbb{R}}\{x \mid F(x) \geq \alpha\}. \tag{20}$$

We show that VaR_α is quasiconvex, however, as previously mentioned, we could not find any set of conditions that ensure the smoothness of VaR_α, and thus we cannot invoke Theorem 4. On a positive note, we show that for distributions satisfying (19), stability holds under the semi-norm in ((18)) and with parameters $b = \max\{b_\alpha, (M_\alpha + 2)/\min\{\alpha, 1-\alpha\}M_\alpha\}, q = 1, \upsilon = 2/3$. We conclude that $\mathcal{O}(\sqrt{\log T/T})$ regret is obtainable by our framework (by means of Remark 1). We conjecture that improved regret is possible for VaR_α when the arm distributions are also twice differentiable, but it is not clear whether this could be achieved through the smoothness condition.

Notice that the stability issues of VaR_α are such that even Theorem 1 (single arm infinite horizon oracle) is not satisfied without further assumptions on the arm distributions. Concretely, denote the α level set of a function $F \in \mathcal{D}$ by $L_\alpha(F) = \{x \in \mathbb{R} \big| F(x) = \alpha\}$, then Theorem 1 holds if $|L_\alpha(\alpha)| \leq 1$ for all $F \in \mathcal{D}^\Delta$. Intuitively, this condition ensures that the arm distributions do not have a flat region at their α percentile. If such a flat region exists, then arbitrarily small perturbations to the distribution may change the percentile by a constant, thus causing the instability issue. Interestingly, even in the presence of this instability, we have the following result.

Proposition 3 (VaR$_\alpha$ oracle policy). *For $\alpha \in (0,1)$, VaR_α always admits a simple oracle policy $\pi^*(\infty)$, i.e., choosing a single arm throughout the horizon is asymptotically optimal.*

5.4. A numerical illustration of an edge case When considering the existence of simple oracle policies, Proposition 3 essentially means that stability, while a sufficient condition, is not necessary. However, for the purpose of regret analysis, we highlight the importance of stability by means of a simulation. Note that Theorem 4 relies on a suitably fast diminishing horizon gap (see $J_1(T), J_3(T)$ in (12)). We calculate this gap in a simple simulation with $K = 1$ arms. This is done for three different distributions, each not satisfying a different subset of the previously discussed conditions for stability and or smoothness. Figure 1 displays the simulation results, which show that the obtained rate is slower than the desired $\frac{\log T}{T}$ which is achieved in Theorem 3.

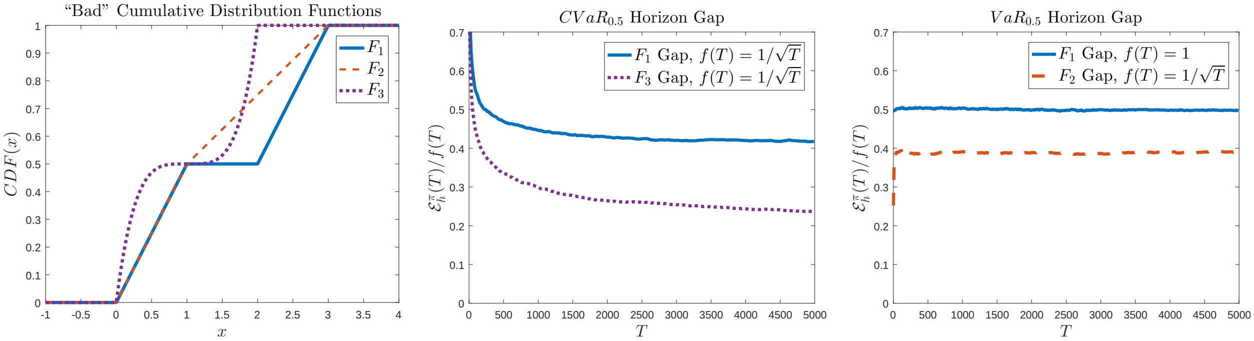

FIGURE 1. VaR_α and $CVaR_\alpha$ horizon gap for "bad" distributions. (F_1) has $|L_{0.5}(F)| > 1$ and has no density to the right of the percentile; (F_2) has no density around percentile $\alpha = 0.5$; and (F_3) is not differentiable. The figures essentially show that $\lim_{T\to\infty} J_1(T)/f(T) = c > 0$ thus claiming that $J_1(T)$ behaves as $f(T)$, which is slower than the desired $\mathcal{O}(\frac{\log T}{T})$.

6. Open Problems and Future Directions

One main question that we leave open is the dependence of the regret on the number of arms K. We conjecture that a finer analysis of the horizon gap may reduce it from our $K^2 \log K$ to either K or $K \log K$. The subject of lower bounds remains open as well. Future directions may include a more complete taxonomy of performance criteria, or an extension of this framework to different settings (e.g., adversarial or contextual). Additionally, we note that the majority of our proof techniques also apply to non-quasiconvex criteria. If such criteria are found to be of interest then extending the framework to this case may be appealing.

A future direction of great interest is to consider a Markov decision model for the dynamics. The same criteria of interest are still relevant, but now it is unclear whether a simple (Markov) policy could approximate the oracle and if so, at what rate.

Acknowledgments. A preliminary version of this work appeared at the Conference on Learning Theory, 2018. We thank Ron Amit, Guy Tennenholtz, Nir Baram and Nadav Merlis for helpful discussions of this work, and the anonymous reviewers for their helpful comments. This work was partially funded by the Israel Science Foundation under contract 1380/16 and by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 306638 (SUPREL).

References

[1] Agrawal S, Devanur NR (2019) Bandits with global convex constraints and objective. *Operations Research* 67(5):1486–1502.

[2] Agrawal S, Goyal N (2012) Analysis of thompson sampling for the multi-armed bandit problem. *COLT*, 39–1.

[3] Artzner P, Delbaen F, Eber JM, Heath D (1999) Coherent measures of risk. *Mathematical finance* 9(3):203–228.

[4] Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3):235–256.

[5] Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (1995) Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, 322–331 (IEEE).

[6] Bhat SP, Prashanth L (2019) Concentration of risk measures: A wasserstein distance approach. *Advances in Neural Information Processing Systems*, 11762–11771.

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

[8] Cassel A, Mannor S, Zeevi A (2018) A general approach to multi-armed bandits under risk criteria. *Conference On Learning Theory*, 1295–1306.

[9] David Y, Szörényi B, Ghavamzadeh M, Mannor S, Shimkin N (2018) Pac bandits with risk constraints. *ISAIM*.

[10] Fisher E (1992) On the law of the iterated logarithm for martingales. *The Annals of Probability* 675–680.

[11] Galichet N, Sebag M, Teytaud O (2013) Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. *ACML*, 245–260.

[12] Jiang DR, Powell WB (2018) Risk-averse approximate dynamic programming with quantile-based risk measures. *Mathematics of Operations Research* 43(2):554–579.

[13] Kagrecha A, Nair J, Jagannathan KP (2019) Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. *NeurIPS*, 11269–11278.

[14] Klenke A (2014) *Law of the Iterated Logarithm*, 509–519 (London: Springer London), ISBN 978-1-4471-5361-0, URL http://dx.doi.org/10.1007/978-1-4471-5361-0_22.

[15] LA P, Jagannathan K, Kolla R (2020) Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. *Proceedings of Machine Learning and Systems 2020*, 3657–3666 (PMLR).

[16] Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6(1):4–22.

[17] Maillard OA (2013) Robust risk-averse stochastic multi-armed bandits. *International Conference on Algorithmic Learning Theory*, 218–233 (Springer).

[18] Maillard OA, Munos R, Stoltz G (2011) A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. *COLT*, 497–514.

[19] Massart P (1990) The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability* 18(3):1269–1283.

[20] Robbins H (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5):527–535.

[21] Sani A, Lazaric A, Munos R (2012) Risk-aversion in multi-armed bandits. *Advances in Neural Information Processing Systems*, 3275–3283.

[22] Simonnet M (1996) *The Strong Law of Large Numbers*, 311–325 (New York, NY: Springer New York), ISBN 978-1-4612-4012-9, URL http://dx.doi.org/10.1007/978-1-4612-4012-9_15.

[23] Tamkin A, Keramati R, Dann C, Brunskill E (2019) Distributionally-aware exploration for cvar bandits. *NeurIPS 2019 Workshop on Safety and Robustness in Decision Making; RLDM 2019* .

[24] Torossian L, Garivier A, Picheny V (2019) $\mathcal{X}$-armed bandits: Optimizing quantiles, cvar and other risks. Lee WS, Suzuki T, eds., *Proceedings of Machine Learning Research*, volume 101 of *Proceedings of Machine Learning Research*, 252–267 (Nagoya, Japan: PMLR), URL http://proceedings.mlr.press/v101/torossian19a.html.

[25] Tran-Thanh L, Yu JY (2014) Functional bandits. *arXiv preprint arXiv:1405.2432* .

[26] Vakili S, Zhao Q (2015) Mean-variance and value at risk in multi-armed bandit problems. *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1330–1335 (IEEE).

[27] Vakili S, Zhao Q (2016) Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing* 10(6):1093–1111.

[28] Van der Vaart AW (2000) *Asymptotic statistics*, volume 3 (Cambridge university press).

[29] Yu X, King I, Lyu MR (2017) Risk control of best arm identification in multi-armed bandits via successive rejects. *2017 IEEE International Conference on Data Mining (ICDM)*, 1147–1152 (IEEE).

[30] Zhu Q, Tan V (2020) Thompson sampling algorithms for mean-variance bandits. *Proceedings of Machine Learning and Systems 2020*, 2645–2654 (PMLR).

[31] Zimin A, Ibsen-Jensen R, Chatterjee K (2014) Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833* .

Appendix A: EDPMs as Permutation Invariant Performance Measures The following continues the discussion in Section 2 on the motivation behind EDPMs and their relation to permutation invariant performance measures. Let $\{\tilde{U}_t\}_{t=1}^{\infty}$, where $\tilde{U}_t : \mathbb{R}^t \to \mathbb{R}$ is a function that measures the quality of a given reward sequence of length t. A decision maker may then wish to maximize the expected performance, i.e., $\mathbf{E}\tilde{U}_t(X_{\pi,1}, \ldots, X_{\pi,t})$. It makes sense that the preferences of the decision maker remain fixed over time. This means $\tilde{U}_t$ $(t \geq 1)$ should, in some sense, be time invariant. However, such an invariance is hard to grasp when the functions $\tilde{U}_t$ do not share a domain. One way of addressing this issue is to assume that $\tilde{U}_t$ is permutation invariant, i.e., it maps all the permutations of its reward sequence to the same value. We provide a formal definition in the proof of the following (known) result.

Lemma 4 (Permutation invariant function representation). *$\tilde{U}_t$ is permutation invariant if and only if, there exists $U_t : \mathcal{D} \to \mathbb{R}$ such that, $\tilde{U}_t(x_1, \ldots, x_t) = U_t\Big(\hat{F}_t(x_1, \ldots, x_t)\Big)$, where $\hat{F}_t(\cdot)$ is the empirical distribution mapping defined in (4).*

The representation given in Lemma 4 suggests $\mathcal{D}$ as a shared domain thus making it simple to define time invariance. We conclude that EDPMs describe the objectives that are time and permutation invariant.

Proof of Lemma 4. We start with a few definitions. Let Σ_t denote the set of $t \times t$ permutation matrices (binary and doubly stochastic). $\tilde{U}_t$ is said to be permutation invariant if $\tilde{U}_t(\sigma x_{1:t}) = \tilde{U}_t(x_{1:t})$ for all $x_{1:t} \in \mathbb{R}^t$ and $\sigma \in \Sigma_t$. Let, $\hat{\mathcal{D}}_t = \Big\{\hat{F}_t(x_{1:t}) \;\Big|\; x_{1:t} \in \mathbb{R}^t\Big\}$, be the set of empirical distributions created from t elements (the image of $\hat{F}_t$). Let,

$$\hat{F}_t^{-1}\Big(\hat{F}\Big) = \Big\{x_{1:t} \in \mathbb{R}^t \;\Big|\; \hat{F}_t(x_{1:t}) = \hat{F}\Big\},$$

be the inverse image of $\hat{F}_t$ at $\hat{F} \in \hat{\mathcal{D}}_t$. Let,

$$\Sigma(x_{1:t}) = \Big\{\sigma x_{1:t} \;\Big|\; \sigma \in \Sigma_t\Big\},$$

be the set of all permutations of $x_{1:t}$. We can now begin the proof.

First direction: Suppose $\tilde{U}_t(x_1, \ldots, x_t) = U_t\Big(\hat{F}_t(x_1, \ldots, x_t)\Big)$. Notice that $\hat{F}_t$ is indeed permutation invariant as permuting its input simply reorders its finite sum thus not changing the value. This clearly implies that $\tilde{U}_t$ is permutation invariant.

Second direction: Suppose that $\tilde{U}_t$ is permutation invariant. Furthermore, assume that for any $x_{1:t} \in \mathbb{R}^t$, we have that, $\hat{F}_t^{-1}\Big(\hat{F}_t(x_{1:t}))\Big) = \Sigma(x_{1:t})$. Then, define $g : \hat{\mathcal{D}}_t \to \mathbb{R}^t$ in the following way. For any $\hat{F} \in \hat{\mathcal{D}}_t$ choose arbitrarily $g\Big(\hat{F}\Big) \in \hat{F}_n^{-1}(\hat{F}))$. Further define $U_t : \mathcal{D} \to \mathbb{R}$ by,

$$U_t(F) = \begin{cases} \tilde{U}_t(g(F)) & F \in \hat{\mathcal{D}}_t \\ 0 & \text{otherwise.} \end{cases}$$

Then we have that, $g\Big(\hat{F}_t(x_{1:t})\Big) \in \hat{F}_t^{-1}\Big(\hat{F}_t(x_{1:t})\Big) = \Sigma(x_{1:t})$, and thus there exists $\sigma_{g(x)} \in \Sigma_t$, such that $g\Big(\hat{F}_t(x_{1:t})\Big) = \sigma_{g(x)} x_{1:t}$. We conclude that,

$$U_t\Big(\hat{F}_t(x_{1:t})\Big) = \tilde{U}_t\Big(g(\hat{F}_t(x_{1:t}))\Big) = \tilde{U}_t\big(\sigma_{g(x)} x_{1:t}\big) = \tilde{U}_t(x_{1:t}),$$

where the last step uses the permutation invariance of $\tilde{U}_t$.

Proof of assumption: We show that for any $x_{1:t} \in \mathbb{R}^t$, we have that, $\hat{F}_t^{-1}\Big(\hat{F}_t(x_{1:t})\Big) = \Sigma(x_{1:t})$ thus concluding the proof. Let $y_{1:t} \in \Sigma(x_{1:t})$ then there exists $\sigma \in \Sigma_t$ such that $y_{1:t} = \sigma x_{1:t}$. Since $\hat{F}_t$ is permutation invariant then,

$$\hat{F}_t(y_{1:t}) = \hat{F}_t(\sigma x_{1:t}) = \hat{F}_t(x_{1:t}) \implies y_{1:t} \in \hat{F}_t^{-1}\Big(\hat{F}_t(x_{1:t})\Big),$$

and so $\Sigma(x_{1:t}) \subseteq \hat{F}_t^{-1}\Big(\hat{F}_t(x_{1:t})\Big)$. On the other hand, let $y_{1:t} \in \hat{F}_t^{-1}\Big(\hat{F}_t(x_{1:t})\Big)$, then we have that, $\hat{F}_t(y_{1:t}) = \hat{F}_t(x_{1:t})$. Take $\sigma_x^*, \sigma_y^* \in \Sigma_t$ such that, $x_{1:t}^* = \sigma_x^* x_{1:t}$, $y_{1:t}^* = \sigma_y^* y_{1:t}$ are sorted in ascending order. Suppose in contradiction that $x_{1:t}^* \neq y_{1:t}^*$ and let,

$$s_0 = \min\Big\{ s \in \{1, \dots, t\} \;\Big|\; x_{s_0}^* \neq y_{s_0}^* \Big\}$$

be the first index where $x_{1:t}^*$ and $y_{1:t}^*$ differ. Without loss of generality assume that $x_{s_0}^* < y_{s_0}^*$, then we have that,

$$\hat{F}_t(y_{1:t}^*)(x_{s_0}^*) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}_{[y_s^*,\infty]}\big(x_{s_0}^*\big) = \frac{1}{t}\sum_{s=1}^{s_0-1} \mathbb{I}_{[y_s^*,\infty]}\big(x_{s_0}^*\big) = \frac{1}{t}\sum_{s=1}^{s_0-1} \mathbb{I}_{[x_s^*,\infty]}\big(x_{s_0}^*\big) < \hat{F}_t(x_{1:t}^*)\big(x_{s_0}^*\big),$$

where the strict inequality follows since $\mathbb{I}_{[x_{s_0}^*,\infty]}\big(x_{s_0}^*\big) = 1$, and if $s_0 = 1$, then the empty sum is in fact zero. This contradicts $\hat{F}_t(y_{1:t}) = \hat{F}_t(x_{1:t})$ and so, $x_{1:t}^* = y_{1:t}^*$. Since, permutation matrices are invertible then, $y_{1:t} = \sigma_y^{*-1}\sigma_x^* x_{1:t}$. It is well known that $\sigma_y^{*-1}\sigma_x^*$ is always a permutation matrix. So, $y_{1:t} \in \Sigma(x_{1:t})$ and we conclude that $\hat{F}_t^{-1}\Big(\hat{F}_t(x_{1:t})\Big) = \Sigma(x_{1:t})$, as desired. ■

Appendix B: Proofs of Section 4.1

Denote the fraction of time at which arm i was pulled by

$$\hat{p}_i(T) = \frac{\tau_i(T)}{T}, \tag{21}$$

where $\tau_i(T)$ is defined in (2). Recall the definitions of $\hat{F}_T^\pi$ and $\hat{F}_T^{(i)}$ given in (5) and (6). The following Lemma is the main argument of the proof of Theorem 1.

Lemma 5 ($\hat{F}_T^\pi$ sub-convergence)**.** *Suppose $\lim_{t\to\infty}\|\hat{F}_t^{(i)} - F^{(i)}\| = 0$ almost surely for all $i \in \mathbb{K}$. Let $p = (p_1, \dots, p_K) \in \Delta$ and $\{t_l\}_{l=1}^\infty$ be a random vector and subsequence. If*

$$\lim_{l\to\infty}\|\hat{p}(t_l) - p\| = 0 \qquad \textit{Almost Surely,}$$

Then

$$\lim_{l\to\infty}\|\hat{F}_{t_l}^\pi - F_p\| = 0 \qquad \textit{Almost Surely,}$$

where F_p is defined in (9).

Proof. We rearrange the expression of $\hat{F}_T^\pi$ such that the sum is over actions and instead of time:

$$\hat{F}_T^\pi = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}_{[X_{\pi,t},\infty]} = \sum_{i=1}^{K}\frac{\tau_i(T)}{T}\left[\frac{1}{\tau_i(T)}\sum_{t=1}^{\tau_i(T)}\mathbb{I}_{[X_{\pi,t},\infty]}\right] = \sum_{i=1}^{K}\hat{p}_i(T)\hat{F}_{\tau_i(T)}^{(i)}.$$

Then we have that

$$\begin{aligned}
\|\hat{F}_{t_l}^\pi - F_p\| &= \|\sum_{i=1}^{K}\hat{p}_i(t_l)\hat{F}_{\tau_i(t_l)}^{(i)} - p_i F^{(i)}\| \\
&= \|\left[\sum_{i=1}^{K}\hat{p}_i(t_l)\hat{F}_{\tau_i(t_l)}^{(i)} - \hat{p}_i(t_l)F^{(i)}\right] + \left[\sum_{i=1}^{K}\hat{p}_i(t_l)F^{(i)} - p_i F^{(i)}\right]\| \\
&\le \|\sum_{i=1}^{K}\hat{p}_i(t_l)\Big(\hat{F}_{\tau_i(t_l)}^{(i)} - F^{(i)}\Big)\| + \|\sum_{i=1}^{K}F^{(i)}(\hat{p}_i(t_l) - p_i)\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{K} \hat{p}_i(t_l)\|\hat{F}^{(i)}_{\tau_i(t_l)} - F^{(i)}\| + \sum_{i=1}^{K}\|F^{(i)}\||\hat{p}_i(t_l) - p_i| \\
&\leq \sum_{i=1}^{K} \underbrace{\hat{p}_i(t_l)\|\hat{F}^{(i)}_{\tau_i(t_l)} - F^{(i)}\|}_{(*)} + \underbrace{K \max_{1\leq i\leq K}\|F^{(i)}\|\|\hat{p}(t_l) - p\|_\infty}_{(**)\to 0}.
\end{aligned}
$$

The first and second inequalities follows by the triangle inequality and homogeneity of norms. The third follows by Hölder's inequality. By the Lemma's assumption $(**) \to 0$. We show that the same holds for (*). It is enough to show the convergence of the summands in order to conclude the overall convergence of this finite sum. By the Lemma's assumption we have that

$$
\lim_{l\to\infty}\|\hat{F}^{(i)}_{\tau_i(t_l)} - F^{(i)}\| = \begin{cases} 0 & ,\lim_{l\to\infty}\tau_i(t_l) = \infty \\ \|\hat{F}^{(i)}_{\tau} - F^{(i)}\| & ,\lim_{l\to\infty}\tau_i(t_l) = \tau < \infty \end{cases} \qquad \textit{Almost Surely,}
$$

where we used the fact that $\tau_i(t)$ is non-decreasing and thus always converges. Now since both parts of (*) converge then we have that,

$$
\begin{aligned}
\lim_{l\to\infty}\hat{p}_i(t_l)\|\hat{F}^{(i)}_{\tau_i(t_l)} - F^{(i)}\| &= \lim_{l\to\infty}\hat{p}_i(t_l)\lim_{l\to\infty}\|\hat{F}^{(i)}_{\tau_i(t_l)} - F^{(i)}\| \\
&= \begin{cases} 0 & ,\lim_{l\to\infty}\tau_i(t_l) = \infty \\ p_i\|\hat{F}^{(i)}_{\tau} - F^{(i)}\| & ,\lim_{l\to\infty}\tau_i(t_l) = \tau < \infty \end{cases} \qquad \textit{Almost Surely.}
\end{aligned}
$$

Noticing that

$$
\lim_{l\to\infty}\tau_i(t_l) = \tau < \infty \implies p_i = \lim_{l\to\infty}\hat{p}_i(t_l) = \lim_{l\to\infty}\frac{\tau_i(t_l)}{t_l} = 0,
$$

the proof is concluded. ■

Proof of Theorem 1. The remainder of the proof consists of applying Lemma 5. We begin by proving $\mathbf{E}U_{\pi^p} = U\left(F_p\right)$. Let $p \in \Delta$ define the simple policy π^p. Using the strong law of large numbers ([22]) on each coordinate of $\hat{p}(t)$, we conclude that

$$
\lim_{t\to\infty}\|\hat{p}(t) - p\|_\infty = 0 \qquad \textit{Almost Surely.}
$$

Applying Lemma 5 we get that

$$
\lim_{t\to\infty}\|\hat{F}^{\pi^p}_t - F_p\| = 0 \qquad \textit{Almost Surely.}
$$

Since U is assumed to be continuous, we have that,

$$
U_{\pi^p} = \liminf_{t\to\infty} U\left(\hat{F}^{\pi^p}_t\right) \overset{a.s}{=} \lim_{l\to\infty} U\left(\hat{F}^{\pi^p}_{t_l}\right) \overset{a.s}{=} U\left(F_p\right),
$$

where $\{t_l\}_{l=1}^{\infty}$ is the (random) subsequence that achieves the limit inferior. Taking expectation, we conclude that $\mathbf{E}U_{\pi^p} = U\left(F_p\right)$. Now since Δ is compact and $U\left(F_p\right)$ is continuous, then by the Weierstrass theorem we have that there exists $p^* \in \Delta$ such that,

$$
U\left(F_{p^*}\right) = \max_{p\in\Delta} U\left(F_p\right). \tag{22}
$$

We now show that π^{p^*} is optimal thus concluding the first part of the proof. Let $\{t_m\}_{m=1}^{\infty}$ be a (random) subsequence satisfying the limit inferior. The we have that

$$
U_\pi = \liminf_{t\to\infty} U\left(\hat{F}^{\pi}_t\right) \overset{a.s}{=} \lim_{m\to\infty} U\left(\hat{F}^{\pi}_{t_m}\right) = (**).
$$

Noticing again that Δ is compact, we have that for any policy $\pi \in \Pi$, there exist $p \in \Delta$ and $\{t_l\}_{l=1}^{\infty} \subseteq \{t_m\}_{m=1}^{\infty}$ (both random) satisfying $\lim_{l\to\infty} \|\hat{p}(t_l) - p\|_\infty = 0$ almost surely. Using Lemma 5, (22), and the continuity of U we get,

$$(**) \overset{a.s}{=} \lim_{l\to\infty} U\left(\hat{F}^{\pi}_{t_l}\right) \overset{a.s}{=} U\left(F_p\right) \le U\left(F_{p^*}\right) = \mathbf{E}U_{\pi^{p^*}},$$

and taking expectation we have $\mathbf{E}U_\pi \le \mathbf{E}U_{\pi^{p^*}}$ for all $\pi \in \Pi$, i.e., $\pi^{p^*} = \pi^*(\infty)$.

Moving on to the second part of the Theorem, notice that Δ is convex, compact and its set of extreme points is also compact (discrete). So, returning to (22) and using the quasiconvexity of U, we notice that a maximizer is attained at an extreme point of Δ. Formally, there exists $i^* \in \mathbb{K}$ such that

$$U\left(F_{e_{i^*}}\right) = \max_{p\in\Delta} U\left(F_p\right),$$

where $\{e_i\}_{i=1}^K$ are the standard unit vectors in $\mathbb{R}^K$. Continuing as before we conclude that $\pi^{e_{i^*}} = \pi^*(\infty)$ as desired. ■

Appendix C: Proofs of Section 4.3

Proof of Lemma 1. We begin by proving the Lipschitz property. Using the local modulus of continuity assumed by stability we get that for any $F_1, F_2 \in \mathcal{D}^\Delta$

$$\begin{aligned}|U\left(F_1\right) - U\left(F_2\right)| &\le b(\|F_1 - F_2\| + \|F_1 - F_2\|^q) \\ &= b\left(1 + \|F_1 - F_2\|^{q-1}\right)\|F_1 - F_2\| \\ &\le b\left(1 + D^{q-1}\right)\|F_1 - F_2\| \\ &= L\|F_1 - F_2\|,\end{aligned}$$

as desired. Next, we show the pseudo regret decomposition. Using quasiconvexity as in the second part of Theorem 1, there exists $i^* \in \mathbb{K}$ such that $F_{p^*} = F^{(i^*)}$. Using the Lipschitz constant L, and the triangle inequality we thus have that,

$$\begin{aligned}\bar{R}_\pi(T) &= \mathbf{E}\left[U\left(F^{(i^*)}\right) - U\left(F^\pi_T\right)\right] \\ &\le L\mathbf{E}\|F^{(i^*)} - F^\pi_T\| \\ &= L\mathbf{E}\|\frac{1}{T}\sum_{i=1}^K \tau_i(T)\left(F^{(i^*)} - F^{(i)}\right)\| \\ &\le \frac{L}{T}\mathbf{E}\left[\sum_{i=1}^K \tau_i(T)\|F^{(i^*)} - F^{(i)}\|\right] \\ &\le \frac{L\rho}{T}\sum_{i\ne i^*} \Delta_i \mathbf{E}\tau_i(T).\end{aligned}$$

■

Proof of Theorem 2. We begin with the following concentration result due to Requirement 2 of stability.

$$\begin{aligned}\mathbb{P}\left(|U\left(\hat{F}^{(i)}_t\right) - U\left(F^{(i)}\right)| \ge x\right) &\le \mathbb{P}\left(b\left(\|\hat{F}^{(i)}_t - F^{(i)}\| + \|\hat{F}^{(i)}_t - F^{(i)}\|^q\right) \ge x\right) \\ &\le \mathbb{P}\left(\|\hat{F}^{(i)}_t - F^{(i)}\| \ge \frac{x}{2b}\right) + \mathbb{P}\left(\|\hat{F}^{(i)}_t - F^{(i)}\| \ge \left(\frac{x}{2b}\right)^{1/q}\right) \\ &\le 2\exp\left(-\upsilon t\left(\frac{x}{2b}\right)^2\right) + 2\exp\left(-\upsilon t\left(\frac{x}{2b}\right)^{2/q}\right) \\ &\le 4\exp(-t\phi(x)).\end{aligned} \tag{23}$$

Now, for all $i \in \mathbb{K}$ and $1 \le t \le T$ denote the events

$$V_i^t = \left\{ U\left(\hat{F}^{(i)}_{\tau_i(t-1)}\right) > U\left(F^{(i)}\right) + \phi^{-1}\left(\frac{\alpha \log t}{\tau_i(t-1)}\right)\right\},$$
$$V_*^t = \left\{ U\left(\hat{F}^{(i^*)}_{\tau_{i^*}(t-1)}\right) + \phi^{-1}\left(\frac{\alpha \log t}{\tau_{i^*}(t-1)}\right) \le U\left(F^{(i^*)}\right)\right\},$$

and their complements by $\overline{V_i^t}$, $\overline{V_*^t}$ respectively. Using the union bound and (23) we have that

$$\begin{aligned}
\mathbb{P}(V_i^t) &= \mathbb{P}\left(U\left(\hat{F}^{(i)}_{\tau_i(t-1)}\right) - \phi^{-1}\left(\frac{\alpha \log t}{\tau_i(t-1)}\right) > U\left(F^{(i)}\right)\right) \\
&\le \mathbb{P}\left(\max_{1\le s\le t}\left\{ U\left(\hat{F}^{(i)}_s\right) - \phi^{-1}\left(\frac{\alpha \log t}{s}\right)\right\} > U\left(F^{(i)}\right)\right) \\
&\le \sum_{s=1}^{t} \mathbb{P}\left(U\left(\hat{F}^{(i)}_s\right) - \phi^{-1}\left(\frac{\alpha \log t}{s}\right) > U\left(F^{(i)}\right)\right) \\
&\le \sum_{s=1}^{t} \mathbb{P}\left(|U\left(\hat{F}^{(i)}_s\right) - U\left(F^{(i)}\right)| \ge \phi^{-1}\left(\frac{\alpha \log t}{s}\right)\right) \\
&\le \sum_{s=1}^{t} 4\exp\left(-s\phi\left(\phi^{-1}\left(\frac{\alpha \log t}{s}\right)\right)\right) \\
&\le \sum_{s=1}^{t} \frac{4}{t^\alpha} = \frac{4}{t^{\alpha-1}}.
\end{aligned}$$

The same holds for $\mathbb{P}(V_*^t)$, and so we obtain

$$\mathbb{P}(V_i^t \cup V_*^t) \le \mathbb{P}(V_i^t) + \mathbb{P}(V_*^t) \le \frac{8}{t^{\alpha-1}}. \tag{24}$$

Next, we denote $u = \frac{\alpha \log T}{\phi(\Delta_i/2)}$, and show that

$$\left\{\pi_t^{U-UCB} = i\right\} \cap \{\tau_i(t-1) \ge u\} \subseteq V_i^t \cup V_*^t. \tag{25}$$

Indeed, assume in contradiction that $\left\{\pi_t^{U-UCB} = i\right\} \cap \{\tau_i(t-1) \ge u\} \cap \overline{V_i^t} \cap \overline{V_*^t} \ne \emptyset$, then noticing that $\{\tau_i(t-1) \ge u\}$ implies $\left\{\Delta_i \ge 2\phi^{-1}\left(\frac{\alpha \log t}{\tau_i(t-1)}\right)\right\}$ we have:

$$\begin{aligned}
U\left(\hat{F}^{(i^*)}_{\tau_{i^*}(t-1)}\right) + \phi^{-1}\left(\frac{\alpha \log t}{\tau_{i^*}(t-1)}\right) &> U\left(F^{(i^*)}\right) \\
&= U\left(F^{(i)}\right) + \Delta_i \\
&\ge U\left(F^{(i)}\right) + 2\phi^{-1}\left(\frac{\alpha \log t}{\tau_i(t-1)}\right) \\
&\ge U\left(\hat{F}^{(i)}_{\tau_i(t-1)}\right) + \phi^{-1}\left(\frac{\alpha \log t}{\tau_i(t-1)}\right),
\end{aligned}$$

which implies that $\pi_T^{U-UCB} \ne i$, thus contradicting our assumption. Finally, denoting

$$t_0 = \max_{1\le t\le T}\left\{ t \;\middle|\; \tau_i(t-1) \le \max\{u, 1\}\right\},$$

and using (24) and (25) we have that

$$\mathbf{E}\tau_i(T) = \mathbf{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\pi_t^{U-UCB} = i\right\}\right]$$

$$\begin{aligned}
&= \mathbf{E}\left[\sum_{t=1}^{t_0} \mathbb{1}\left\{\pi_t^{U-UCB} = i\right\} + \sum_{t=t_0+1}^{T} \mathbb{1}\left\{\pi_t^{U-UCB} = i\right\}\right] \\
&= \mathbf{E}\left[\tau_i(t_0) + \sum_{t=t_0+1}^{T} \mathbb{1}\left\{\pi_t^{U-UCB} = i \bigcap \tau_i(t-1) \geq u\right\}\right] \\
&\leq \mathbf{E}\left[u+1+ \sum_{t=K+1}^{T} \mathbb{1}\left\{\pi_t^{U-UCB} = i \bigcap \tau_i(t-1) \geq u\right\}\right] \\
&= u+1+ \sum_{t=K+1}^{T} \mathbb{P}\left(\pi_t^{U-UCB} = i \bigcap \tau_i(t-1) \geq u\right) \\
&\leq u+1+ \sum_{t=K+1}^{T} \mathbb{P}(V_i^t \cup V_*^t) \\
&\leq u+1+ \sum_{t=K+1}^{T} \frac{8}{t^{\alpha-1}} \leq u+1+\int_K^\infty \frac{8}{t^{\alpha-1}} dt \leq u+1+\frac{8}{(\alpha-2)K^{\alpha-2}} \leq u+\frac{\alpha+6}{\alpha-2}.
\end{aligned}$$

Combining this with the expression for the pseudo regret given in Lemma 1 we obtain the desired. ■

Appendix D: Proofs of Section 4.4 We first need the following technical lemma whose proof may be found in Section D.1.

Lemma 6. *Suppose that Requirement 2 of stability holds. Then for any integer $d \geq 1$, policy $\pi \in \Pi$, and K, T such that $\log KT \geq 3$, we have that:*

1. $\mathbf{E}\|\hat{F}_T^\pi - F_T^\pi\|^d \leq [1+dm!]\left(\frac{K^2 \log KT}{\upsilon T}\right)^{d/2}$, *where* $m = \left\lceil \frac{d}{2} - 1 \right\rceil$;
2. $\mathbf{E}\|\hat{F}_T^\pi - F_T^\pi\|^d \leq 2\left(\frac{K^2 \log KT}{\upsilon T}\right)^{1/2}$ *for all* $T \geq \frac{4dK^2 \log KT}{\upsilon}$;
3. $\mathbf{E}\left[\|\hat{F}_T^\pi - F_T^\pi\|^d \mathbb{1}\left\{\|\hat{F}_T^\pi - F_T^\pi\| > M\right\}\right] \leq \frac{1}{T^2}$ *for all* $T \geq \frac{4dK^2 \log KT}{\upsilon}, M^2 \geq \frac{4dK^2 \log KT}{\upsilon T}$;

Proof of Proposition 1. We use Requirement 1 of stability together with the second part of Lemma 6 to get that

$$\begin{aligned}
\left|\mathbf{E}\left[U\left(\hat{F}_T^\pi\right) - U\left(F_T^\pi\right)\right]\right| &\leq \mathbf{E}\left|U\left(\hat{F}_T^\pi\right) - U\left(F_T^\pi\right)\right| \\
&\leq b\left[\mathbf{E}\|\hat{F}_T^\pi - F_T^\pi\| + \mathbf{E}\|\hat{F}_T^\pi - F_T^\pi\|^q\right] \\
&\leq 4b\left(\frac{K^2 \log KT}{\upsilon T}\right)^{1/2}.
\end{aligned}$$

■

Proof of Theorem 3. First, notice that $\mathbf{E}\hat{F}_T^\pi = \mathbf{E}F_T^\pi$ for any $\pi \in \Pi$. This is easily seen as,

$$\mathbf{E}\hat{F}_T^\pi = \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{[X_{\pi,t},\infty]}\right] = \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\left[\mathbb{I}_{[X_{\pi,t},\infty]} \middle| \pi_t\right]\right] = \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} F^{(\pi_t)}\right] = \mathbf{E}F_T^\pi.$$

Since ∂U is a linear operator, we conclude that

$$\mathbf{E}\left[\partial U\left(F\right) \cdot \left(\hat{F}_T^\pi - F_T^\pi\right)\right] = 0, \qquad \forall F \in \mathcal{D}^\Delta.$$

With this in mind, we have the following decomposition

$$\left|\mathbf{E}\left[U\left(\hat{F}_T^\pi\right) - U\left(F_T^\pi\right)\right]\right| \leq \mathbf{E}\left|\underbrace{U\left(\hat{F}_T^\pi\right) - U\left(F_T^\pi\right) - \partial U\left(F_T^\pi\right) \cdot \left(\hat{F}_T^\pi - F_T^\pi\right)}_{\delta_1}\right|$$

$$+\mathbf{E}\Big|\underbrace{\left(\partial U\left(F_T^\pi\right)-\partial U\left(F_\gamma\right)\right)\cdot\left(\hat{F}_t^\pi-F_T^\pi\right)}_{\delta_2}\Big|,$$

We bound $\mathbf{E}|\delta_1|, \mathbf{E}|\delta_2|$ to conclude the proof. For $\mathbf{E}|\delta_1|$, recalling the parameter M_0 in Definition 2 (smoothness), we use smoothness together with the first part of Lemma 6 to get that

$$\mathbf{E}\Big[|\delta_1|\mathbb{1}\big\{\|\hat{F}_T^\pi-F_T^\pi\|\le M_0\big\}\Big]\le\frac{1}{2}\beta\mathbf{E}\Big[\|\hat{F}_T^\pi-F_T^\pi\|^2\Big]\le\frac{3\beta K^2\log KT}{2\upsilon T}.$$

Next, we use stability (modulus of continuity) together with part 3 of Lemma 6, which holds due to our assumption on M_0, to get that

$$\begin{aligned}\mathbf{E}\Big[|\delta_1|\mathbb{1}\big\{\|\hat{F}_T^\pi-F_T^\pi\|>M_0\big\}\Big]&\le\mathbf{E}\Big[\mathbb{1}\big\{\|\hat{F}_T^\pi-F_T^\pi\|>M_0\big\}\Big(\omega\big(\|\hat{F}_T^\pi-F_T^\pi\|\big)+2b\|\hat{F}_T^\pi-F_T^\pi\|\Big)\Big]\\&=b\mathbf{E}\Big[\mathbb{1}\big\{\|\hat{F}_T^\pi-F_T^\pi\|>M_0\big\}\Big(\|\hat{F}_T^\pi-F_T^\pi\|^q+3\|\hat{F}_T^\pi-F_T^\pi\|\Big)\Big]\\&\le\frac{4b\mathbb{1}\{M_0<\infty\}}{T^2},\end{aligned}$$

where the first step also used the modulus of continuity to bound the Gateaux derivative of U. Summing the two inequalities bounds $\mathbf{E}|\delta_1|$.

Finally, to bound $\mathbf{E}|\delta_2|$ we first use the Cauchy–Schwarz inequality together with smoothness (Definition 2) to get that

$$\mathbf{E}|\delta_2|\le\mathbf{E}\Big[\|\partial U\left(F_T^\pi\right)-\partial U\left(F_\gamma\right)\|\|\hat{F}_T^\pi-F_T^\pi\|\Big]\le\beta\mathbf{E}\Big[\|F_T^\pi-F_\gamma\|\|\hat{F}_T^\pi-F_T^\pi\|\Big].$$

Next, let $J_T^2=\frac{4K^2\log KT}{\upsilon T}$ and use the assumption on F_γ to get that

$$\mathbf{E}\Big[|\delta_2|\mathbb{1}\big\{\|\hat{F}_T^\pi-F_T^\pi\|\le J_T\big\}\Big]\le\beta J_T\mathbf{E}\|F_T^\pi-F\|\le\frac{2\gamma\beta K^2\log KT}{\upsilon T}.$$

On the other hand, recalling that D is the diameter of $\mathcal{D}^\Delta$, we use part 3 of Lemma 6 to get that

$$\mathbf{E}\Big[|\delta_2|\mathbb{1}\big\{\|\hat{F}_T^\pi-F_T^\pi\|>J_T\big\}\Big]\le\beta D\mathbf{E}\Big[\|\hat{F}_T^\pi-F_T^\pi\|\mathbb{1}\big\{\|\hat{F}_T^\pi-F_T^\pi\|>J_T\big\}\Big]\le\frac{\beta D}{T^2}.$$

Summing the two inequalities bounds $\mathbf{E}|\delta_2|$, and concludes the proof. ■

D.1. Technical Side Lemmas

Proof of Lemma 2. We start by using the triangle inequality and the union bound to get,

$$\begin{aligned}\mathbb{P}\Big(\|\hat{F}_T^\pi-F_T^\pi\|>x\Big)&=\mathbb{P}\left(\|\frac{1}{T}\sum_{i=1}^K\tau_i(T)\Big(\hat{F}^{(i)}_{\tau_i(T)}-F^{(i)}\Big)\|>x\right)\\&\le\mathbb{P}\left(\frac{1}{T}\sum_{i=1}^K\tau_i(T)\|\hat{F}^{(i)}_{\tau_i(T)}-F^{(i)}\|>x\right)\\&\le\sum_{i=1}^K\mathbb{P}\Big(\tau_i(T)\|\hat{F}^{(i)}_{\tau_i(T)}-F^{(i)}\|>\frac{T}{K}x\Big).\end{aligned}$$

Now notice that $0\le\tau_i(T)\le T$. So we have that,

$$\tau_i(T)\|\hat{F}^{(i)}_{\tau_i(T)}-F^{(i)}\|\le\max_{1\le s\le T}s\|\hat{F}^{(i)}_s-F^{(i)}\|,$$

where the case of $\tau_i(T)=s=0$ is dropped as it is clearly not the maximizer. Using this expression together with the union bound we get that,

$$\begin{aligned}\mathbb{P}\Big(\|\hat{F}_T^\pi - F_T^\pi\| > x\Big) &\le \sum_{i=1}^K \mathbb{P}\Big(\max_{1\le s\le T} s\|\hat{F}_s^{(i)} - F^{(i)}\| > \frac{T}{K}x\Big)\\ &\le \sum_{i=1}^K\sum_{s=1}^T \mathbb{P}\Big(\|\hat{F}_s^{(i)} - F^{(i)}\| > \frac{Tx}{sK}\Big).\end{aligned}$$

Applying Requirement 2 of stability (concentration), we have that

$$\begin{aligned}\mathbb{P}\Big(\|\hat{F}_T^\pi - F_T^\pi\| > x\Big) &\le \sum_{i=1}^K\sum_{s=1}^T 2\exp\bigg(-\upsilon s\Big(\frac{Tx}{sK}\Big)^2\bigg)\\ &= 2K\sum_{s=1}^T \exp\Big(-\upsilon\frac{T^2x^2}{sK^2}\Big)\\ &\le 2KT\exp\Big(-\upsilon\frac{Tx^2}{K^2}\Big),\end{aligned}$$

where in the last step we use the fact that $s=T$ maximizes the summands. ■

Proof of Lemma 6. For the first claim, recall that $m=\lceil \frac{d}{2}-1\rceil$, and let $x_0\ge 0$ be a constant to be determined later. We begin by using the tail sum formula and exchanging variables to get that

$$\begin{aligned}\mathbf{E}\|\hat{F}_T^\pi - F_T^\pi\|^d &= \int_0^\infty \mathbb{P}\Big(\|\hat{F}_T^\pi - F_T^\pi\|^d > x\Big)dx\\ &\le x_0^d + \int_{x_0^d}^\infty \mathbb{P}\Big(\|\hat{F}_T^\pi - F_T^\pi\|^d > x\Big)dx\\ &= x_0^d + dx_0^d\int_1^\infty x^{d-1}\mathbb{P}\Big(\|\hat{F}_T^\pi - F_T^\pi\| > x_0x\Big)dx && (x=(x_0x')^d)\\ &\le x_0^d\bigg[1+d\int_1^\infty x^{2m+1}\mathbb{P}\Big(\|\hat{F}_T^\pi - F_T^\pi\| > x_0x\Big)dx\bigg],\end{aligned}$$

where the last transition used the fact that $2m+1\ge d-1$. Next, we use the tail bound in Lemma 2 to get that

$$\mathbf{E}\|\hat{F}_T^\pi - F_T^\pi\|^d \le x_0^d\bigg[1+2dKT\int_1^\infty x^{2m+1}\exp\Big(-\frac{\upsilon Tx_0^2x^2}{K^2}\Big)dx.\bigg]$$

Now, choose $x_0^2=\frac{K^2\log KT}{\upsilon T}$, and using Lemma 7 with $a=\log KT\ge 3$ to solve this known integral, we get that

$$\mathbf{E}\|\hat{F}_T^\pi - F_T^\pi\|^d \le [1+dm!]\Big(\frac{K^2\log KT}{\upsilon T}\Big)^{d/2}.$$

This holds for all $\pi\in\Pi$ thus concluding the proof of the first claim.

Next, we prove the second claim by showing that, under the assumption on T, the first claim may be bounded by the desired term. To see this notice that

$$\begin{aligned}\frac{[1+dm!]}{2}\Big(\frac{K^2\log KT}{\upsilon T}\Big)^{\frac{d-1}{2}} &\le d^{d/2}\Big(\frac{K^2\log KT}{\upsilon T}\Big)^{\frac{d-1}{2}}\\ &\le \frac{d^{1/2}}{4^{(d-1)/2}} && (T\ge \tfrac{4dK^2\log KT}{\upsilon})\\ &= \exp\Big(\frac{1}{2}\log d - \frac{d-1}{2}\log 4\Big)\\ &\le \exp\Big(\frac{d-1}{2} - \frac{d-1}{2}\log 4\Big) \le 1. && (\log x\le x-1)\end{aligned}$$

Changing sides gives the desired bound, and concludes the proof of the second claim.

Finally, for the third claim, we begin by repeating the first steps of the first claim to get that

$$\mathbf{E}\Big[\mathbb{1}\big\{\|\hat{F}_T^\pi - F_T^\pi\| > M\big\}\|\hat{F}_T^\pi - F_T^\pi\|^d\Big] \le \mathbf{E}\Big[\mathbb{1}\big\{\|\hat{F}_T^\pi - F_T^\pi\| > x_1\big\}\|\hat{F}_T^\pi - F_T^\pi\|^d\Big] \\ \le 2dx_1^d KT \int_1^\infty x^{2m+1} \exp\left(-\frac{\upsilon T x_1^2 x^2}{K^2}\right) dx,$$

where we choose $x_1^2 = \frac{4dK^2 \log KT}{\upsilon T} \le M$. Notice that our choice of T ensures that $x_1 \le 1$ and thus applying Lemma 7 with $a = 4d \log KT \ge 3$, we get that

$$\mathbf{E}\Big[\mathbb{1}\big\{\|\hat{F}_T^\pi - F_T^\pi\| > M\big\}\|\hat{F}_T^\pi - F_T^\pi\|^d\Big] \le \frac{dm!}{(KT)^{4d-1}} \le \frac{1}{T^2},$$

where the last step also used the fact that $T \ge d$. ■

Lemma 7. *For any real $a > 0$ and integer $m \ge 0$ we have that*

$$\int_1^\infty x^{2m+1} \exp(-ax^2) dx = \frac{\exp(-a)}{2a^{m+1}} \sum_{j=0}^m \frac{m!}{j!} a^j.$$

If $a \ge 3$ then we also have that

$$\int_1^\infty x^{2m+1} \exp(-ax^2) dx \le m! \exp(-a)/2.$$

Proof. Denote $f(m) = \int_1^\infty x^{2m+1} \exp(-ax^2) dx$, and use integration by parts to get that

$$f(m) = \frac{m}{a} f(m-1) + \frac{\exp(-a)}{2a}.$$

Now, plugging $m = 0$ we get that $f(0) = \frac{\exp(-a)}{2a}$. Finally, it is trivial to verify that the suggested solution satisfies the difference equation as well as the initial condition, thus concluding the first part of the proof. For the second part we use the assumption on $a \ge 3$ to upper bound the expression as

$$f(m) \le \frac{m! \exp(-a)}{2a} \sum_{j=0}^m \frac{1}{j!} \le \frac{m! e \exp(-a)}{2a} \le \frac{m! \exp(-a)}{2}$$

■

Appendix E: Proofs of Section 4.5

Proof of Proposition 2. Recall that for any $F \in \mathcal{D}^\Delta$ there exists $p \in \Delta$ such that $F = \sum_{i=1}^K p_i F^{(i)}$. Now, we use convexity to conclude that

$$U\big(\sum_{i=1}^K p_i F^{(i)}\big) - U\big(F^{(i^*)}\big) \le \sum_{i=1}^K p_i \Big(U\big(F^{(i)}\big) - U\big(F^{(i^*)}\big)\Big) \\ = -\sum_{i=1}^K p_i \Delta_i \le -\frac{1}{\rho} \sum_{i=1}^K p_i \|F^{(i)} - F^{(i^*)}\| = -\frac{1}{\rho}\|F - F^{(i^*)}\|,$$

where ρ, which is defined in (13), is finite since $\Delta_i > 0$ for all $i \ne i^*$. ■

Proof of Theorem 4. We begin by stating the explicit condition on the time horizon T. Letting $J_T^2 = \frac{4K^2 \log KT}{\upsilon T}$, we require that T is large enough such that

$$\frac{\rho}{T}\sum_{i \neq i^*}\left(\frac{\alpha \Delta_i \log T}{\phi(\Delta_i/2)} + \frac{\alpha+6}{\alpha-2}\Delta_i\right) \le J_T \le \min\left\{\frac{1}{\sqrt{q}}, \frac{M_0}{\sqrt{q}}, \frac{1}{\eta\beta}\right\}, \tag{26}$$

which is indeed polynomial in the problem parameters. The first two terms in the minimum are the basic requirements of Theorem 3, and the third term was chosen such that applying Theorem 3 with $F = F^{(i^*)}$, we get that

$$J_1(T) = \mathbf{E}\Big[U\big(\hat{F}_T^{\pi^*(T)}\big) - U\big(F_T^{\pi^*(T)}\big)\Big] \le \frac{2\beta K^2 \log KT}{\upsilon T} + \frac{1}{\eta}\mathbf{E}\|F_T^{\pi^*(T)} - F^{(i^*)}\| + \frac{\beta D + 4b\mathbb{1}\{M_0 < \infty\}}{T^2}.$$

Next, notice that the first step in the decomposition of the pseudo regret, which is given in Lemma 1, is $\bar{R}_\pi(T) \le L\mathbf{E}\|F_T^{U-UCB} - F^{(i^*)}\|$, and thus the bound in Theorem 2 together with the left hand side of (26) imply that $\mathbf{E}\|F_T^{U-UCB} - F^{(i^*)}\| \le J_T$. Applying Theorem 3 with $F = F^{(i^*)}$ we obtain that

$$J_3(T) = \mathbf{E}\Big[U\big(F_T^{U-UCB}\big) - U\big(\hat{F}_T^{U-UCB}\big)\Big] \le \underbrace{\frac{6\beta K^2 \log KT}{\upsilon T}}_{3\beta J_T^2/2} + \frac{\beta D + 4b\mathbb{1}\{M_0 < \infty\}}{T^2}.$$

Next, using the linear gap assumption we get that

$$J_2(T) = \mathbf{E}\Big[U\big(F_T^{\pi^*(T)}\big) - U\big(F^{(i^*)}\big)\Big] \le -\frac{1}{\eta}\mathbf{E}\|F_T^{\pi^*(T)} - F^{(i^*)}\|.$$

Finally, recall that in (12) we decompose the regret as $R_\pi(T) = \bar{R}_\pi(T) + J_1(T) + J_2(T) + J_3(T)$. Combining the above and using Theorem 2 to bound $\bar{R}_{U-UCB}(T)$ concludes the proof. ■

Online Companion

Appendix F: Details of Section 5 In this section we provide the missing details from Section 5. For the most part, our goal is to verify stability and smoothness, which are the conditions for Theorem 4. We note that Theorem 1 will typically hold as long as $|U\left(F^{(i)}\right)| < \infty$. The exception to this rule is VaR_α, for which we will require an additional assumption for this to hold. However, we also show that a single arm infinite horizon oracle always exists, even without this assumption.

In what follow we continue to operate under Assumption 1, which bounds the rewards. This is mostly to make the exposition more concise, and we state explicitly the places where it is indeed necessary.

F.1. Linear EDPMs We expand on the application of Hoeffding's inequality for the stability of linear EDPMs, and note that this could easily be replaced by a sub-Gaussian type assumption. Recall that for a linear EDPM U^{lin}, we use the seminorm $\|F\| = |U^{\text{lin}}\left(F\right)|$. We thus have that

$$\|\hat{F}_t^{(i)} - F^{(i)}\| = |U^{\text{lin}}\left(\hat{F}_t^{(i)} - F^{(i)}\right)| = \left|\frac{1}{t}\sum_{s=1}^{t} U^{\text{lin}}\left(\mathbb{I}_{[X_{i,s},\infty]}\right) - U^{\text{lin}}\left(F^{(i)}\right)\right|,$$

where the last transition used the definition of the empirical distribution in (6), and the linearity of U^{lin}. Notice that the linearity of U^{lin} also implies that $U^{\text{lin}}\left(F^{(i)}\right) = \mathbf{E}U^{\text{lin}}\left(\mathbb{I}_{[X_{i,s},\infty]}\right)$. We now have a sum of zero mean *i.i.d* random variables that take values in an interval of squared length

$$\vartheta_{\text{lin}} = \max_{x,y\in[0,1]} \left[U^{\text{lin}}\left(\mathbb{I}_{[x,\infty]}\right) - U^{\text{lin}}\left(\mathbb{I}_{[y,\infty]}\right)\right]^2,$$

and thus invoking Hoeffding's inequality we get that U^{lin} satisfies Requirement 2 of stability with $\upsilon = 2/\vartheta_{\text{lin}}$.

F.2. Composite EDPMs Recall that an EDPM is composite if there exist $U^{(1)}, \ldots U^{(n)}$ and $h : \mathbb{R}^n \to \mathbb{R}$ such that

$$U^h\left(F\right) = h\left(U^{(1)}\left(F\right), \ldots, U^{(n)}\left(F\right)\right).$$

For a set $S \subseteq \mathcal{D}$ let

$$U^h\left(S\right) = \left\{\left(U^{(1)}\left(F\right), \ldots, U^{(n)}\left(F\right)\right) \in \mathbb{R}^n \middle| \forall F \in S\right\}.$$

be its image under the linear mappings that compose U^h. Lemma 3 is made formal in the following result.

Lemma 8 (Composite EDPM). *Suppose $U^{(1)}, \ldots, U^{(n)}$ are linear, and stable with parameter υ_0. Then:*

1. *If h admits a polynomial local modulus of continuity, i.e., there exist $b > 0, q \ge 1$ such that*

$$|h(x) - h(y)| \le b(\|x-y\|_2 + \|x-y\|_2^q) \qquad , \forall x \in U^h\left(\mathcal{D}^\Delta\right), y \in U^h\left(L_{\|\cdot\|}\right),$$

then U^h is stable with the same b, q and $\upsilon = \frac{\log 2}{n \log 2n}\upsilon_0$;

2. *If h is locally smooth, i.e., there exist $\beta \ge 0, M_0 > 0$ such that for any $x \in U^h\left(\mathcal{D}^\Delta\right), y \in U^h\left(L_{\|\cdot\|}\right)$ satisfying $\|x-y\|_2 \le M_0$ we have that*

$$|h(y) - h(x) - \nabla h(x)^T (y-x)| \le \frac{\beta}{2}\|x-y\|_2^2,$$

then U^h is smooth with the same parameters;

3. *If h is convex then so is U^h.*

Proof. Recall that we consider U^h under the norm

$$\|F\| = \|U^{(1)}(F), \ldots, U^{(n)}(F)\|_2,$$

where $\|\cdot\|_2$ is the ℓ^2 norm on $\mathbb{R}^n$. Starting with Requirement 1 of stability, we use the modulus of continuity assumption on h to get that for all $F \in \mathcal{D}^\Delta$ and $G \in L_{\|\cdot\|}$

$$\begin{aligned} |U^h(F) - U^h(G)| &= |h(U^{(1)}(F), \ldots, U^{(n)}(F)) - h(U^{(1)}(G), \ldots, U^{(n)}(G))| \\ &\le b(\|U^{(1)}(F-G), \ldots, U^{(n)}(F-G)\|_2 + \|U^{(1)}(F-G), \ldots, U^{(n)}(F-G)\|_2^q) \\ &= b(\|F-G\| + \|F-G\|^q). \end{aligned}$$

Next, for Requirement 2 we use the stability of the linear EDPMs to get that

$$\begin{aligned} \mathbb{P}\left(\|\hat{F}_t^{(i)} - F^{(i)}\| > x\right) &= \mathbb{P}\left(\|U^{(1)}(\hat{F}_t^{(i)} - F^{(i)}), \ldots, U^{(n)}(\hat{F}_t^{(i)} - F^{(i)})\|_2 > x\right) \\ &\le \min\left\{1, \sum_{j=1}^n \mathbb{P}\left(|U^{(j)}(\hat{F}_t^{(i)} - F^{(i)})| > \frac{x}{\sqrt{n}}\right)\right\} && \text{(union bound)} \\ &\le \min\left\{1, 2n\exp(-\frac{v_0 t x^2}{n})\right\} \\ &\le 2\exp(-\frac{\log 2v_0}{n \log 2n} t x^2), \end{aligned}$$

thus concluding stability of U^h, which is the first claim.

Now, moving on to smoothness, we use to chain rule to get that for any $F \in \mathcal{D}$ and $G \in L_{\|\cdot\|}$

$$\partial U(F) \cdot G = \partial h(U^{(1)}(F), \ldots, U^{(n)}(F)) \cdot G = \nabla h(U^{(1)}(F), \ldots, U^{(n)}(F))^T (U^{(1)}(G), \ldots, U^{(n)}(G)),$$

and applying the assumed smoothness of h we get that if $\|F - G\| \le M_0$ then

$$\begin{aligned} &|U^h(G) - U^h(F) - \partial U(F) \cdot (G - F)| \\ &= \Big| h(U^{(1)}(G), \ldots, U^{(n)}(G)) - h(U^{(1)}(F), \ldots, U^{(n)}(F)) \\ &\quad - \nabla h(U^{(1)}(F), \ldots, U^{(n)}(F))^T (U^{(1)}(G-F), \ldots, U^{(n)}(G-F)) \Big| \\ &\le \frac{1}{2}\beta \|U^{(1)}(G-F), \ldots, U^{(n)}(G-F)\|_2^2 \\ &= \frac{1}{2}\beta \|F - G\|^2, \end{aligned}$$

thus concluding the smoothness of U^h, which is the second claim.

Finally, if h is convex the U^h is a linear variable on a convex function and as such convex. ■

F.2.1. Entropic risk This is the only example where Assumption 1 is indeed necessary for our framework. We note that this could be removed in the future by expanding our analysis to an exponential family of moduli of continuity. Recall that in terms of Lemma 8, we have that $h(x) = -\frac{1}{\theta}\log x$, which is convex, and $x \in [\exp(-\theta), 1]$. Bounding the first derivative, we get that

$$\frac{dh}{dx}(x) = -\frac{1}{\theta x} \implies \left|\frac{dh}{dx}(x)\right| \le \frac{1}{\theta}\exp(\theta),$$

and thus h is Lipschitz with this constant and has a modulus of continuity with parameters $b = \frac{1}{2\theta}\exp(\theta), q = 1$. Next, recalling the second order charachterization of smoothness, we bound the second derivative, to get that

$$\beta \le \max_{x \in [\exp(-\theta), 1]} \left|\frac{d^2 h}{dx^2}(x)\right| = \max_{x \in [\exp(-\theta), 1]} \frac{1}{\theta x^2} = \frac{1}{\theta}\exp(2\theta),$$

thus proving the desired properties for Lemma 8.

F.2.2. Variance Here $h(x_1,x_2)=x_1^2-x_2$, which is convex. Next, for the modulus of continuity we have that

$$\begin{aligned}|h(x_1,x_2)-h(y_1,y_2)| &= |(x_1-y_1)(x_1+y_1)+(y_2-x_2)| \\ &\le (x_1-y_1)^2+|2x_1(x_1-y_1)+(y_2-x_2)| \\ &\le \sqrt{1+4x_1^2}(\|x-y\|_2^2+\|x-y\|_2).\end{aligned}$$

Since $(x_1,x_2)\in U^h(\mathcal{D}^\Delta)$, we can bound $|x_1|\le \max_{i\in\mathbb{K}}|U^{\text{ave}}(F^{(i)})|$. Since the reward is also bounded in $[0,1]$ we further have that $|x_1|\le 1$, giving us the constant $b=\sqrt{5}$. Finally, for smoothness we may bound the hessian as,

$$\beta=\max_{x_1,x_2\in\mathbb{R}}\|\nabla^2 h(x_1,x_2)\| = \left\|\begin{pmatrix}2 & 0\\ 0 & 0\end{pmatrix}\right\| = 2.$$

F.2.3. Mean-variance (Markowitz) Here we have that for $\rho\ge 0$ $h(x,y)=x+\rho(x^2-y)$, which is convex. Next, for the modulus of continuity we have that

$$\begin{aligned}|h(x_1,x_2)-h(y_1,y_2)| &= |\rho(x_1-y_1)(x_1+y_1)+\rho(y_2-x_2)+(x_1-y_1)| \\ &\le \rho(x_1-y_1)^2+|(2\rho x_1+1)(x_1-y_1)+(y_2-x_2)| \\ &\le \sqrt{1+(2\rho|x_1|+1)^2}(\|x-y\|_2^2+\|x-y\|_2).\end{aligned}$$

Since $(x_1,x_2)\in U^h(\mathcal{D}^\Delta)$, we can bound $|x_1|\le \max_{i\in\mathbb{K}}|U^{\text{ave}}(F^{(i)})|$. Since the reward is also bounded in $[0,1]$ we further have that $|x_1|\le 1$, giving us the constant $b=2(1+\rho)$. Finally, for smoothness we may bound the hessian as,

$$\beta=\max_{x_1,x_2\in\mathbb{R}}\|\nabla^2 h(x_1,x_2)\| = \left\|\begin{pmatrix}2\rho & 0\\ 0 & 0\end{pmatrix}\right\| = 2\rho.$$

F.2.4. Sortino ratio Here we have that for $r\in\mathbb{R}$ and $\varepsilon_0>0$

$$h(x_1,x_2)=(x_1-r)/\sqrt{\varepsilon_0-x_2}.$$

Linear gap: Let $\lambda\in\Delta$ and $F=\sum_{i=1}^K\lambda_i F^{(i)}$. We need to show that

$$U^{\text{So}}(F^{(i^*)})-U^{\text{So}}(F)\ge\frac{1}{\eta}\|F^{(i^*)}-F\|.$$

Denote $x_i=U^{\text{ave}}(F^{(i)})$, $y_i=U^{\text{TSV}}(F^{(i)})$, and $z_i=\sqrt{\varepsilon_0-y_i}$. Notice that z_i is concave in x_i,y_i and so we have that

$$\sqrt{\varepsilon_0-\sum_{i=1}^K\lambda_i y_i}\ge\sum_{i=1}^K\lambda_i z_i.$$

We conclude that

$$\begin{aligned}U^{\text{So}}(F^{(i^*)})-U^{\text{So}}(F) &= h(x_{i^*},y_{i^*})-h(\sum_{i=1}^K\lambda_i x_i,\sum_{i=1}^K\lambda_i y_i) \\ &\ge \frac{x_{i^*}-r}{z_{i^*}}-\frac{\sum_{i=1}^K\lambda_i x_i-r}{\sum_{i=1}^K\lambda_i z_i} \\ &= \frac{\sum_{i=1}^K\lambda_i(z_i(x_{i^*}-r)-z_{i^*}(x_i-r))}{z_{i^*}\sum_{j=1}^K\lambda_j z_j} \\ &= \sum_{i=1}^K\lambda_i\frac{z_i}{\sum_{j=1}^K\lambda_i z_i}(\frac{x_{i^*}-r}{z_{i^*}}-\frac{x_i-r}{z_i})\ge\frac{z_{\min}}{z_{\max}}\sum_{i=1}^K\lambda_i\Delta_i,\end{aligned}$$

where $z_{\min} = \min_{i \neq i^*} z_i$ and $z_{\max} = \max_{i \in \mathbb{K}} z_i$. Using the gap ratio defined in (13) we get a linear gap with $\eta = \rho z_{\max}/z_{\min}$.

Stability: For $x \in U^{\text{So}}(\mathcal{D}^{\Delta})$ and $y \in U^{\text{So}}(L_{\|\cdot\|})$ we have that

$$\begin{aligned}
|h(x_1,x_2) - h(y_1,y_2)| &= \left| \frac{x_1 - r}{\sqrt{\varepsilon_0 - x_2}} - \frac{y_1 - r}{\sqrt{\varepsilon_0 - y_2}} \right| \\
&\leq \left| \frac{x_1 - r}{\sqrt{\varepsilon_0 - y_2}} - \frac{y_1 - r}{\sqrt{\varepsilon_0 - y_2}} \right| + \left| \frac{x_1 - r}{\sqrt{\varepsilon_0 - x_2}} - \frac{x_1 - r}{\sqrt{\varepsilon_0 - y_2}} \right| \\
&\leq \frac{|x_1 - y_1|}{\varepsilon_0} + |x_1 - r| \left| \frac{\sqrt{\varepsilon_0 - y_2} - \sqrt{\varepsilon_0 - x_2}}{\sqrt{\varepsilon_0 - x_2}\sqrt{\varepsilon_0 - y_2}} \right| \\
&\leq \frac{|x_1 - y_1|}{\varepsilon_0} + \frac{|x_1 - r|}{2\varepsilon_0^{3/2}} |y_2 - x_2| \\
&\leq \sqrt{\frac{1}{\varepsilon_0^2} + \frac{(x_1 - r)^2}{4\varepsilon_0^3}} \|x - y\|_2 \\
&\leq \frac{|x_1 - r| + 2}{2\min\{\varepsilon_0, \varepsilon_0^{3/2}\}} \|x - y\|_2.
\end{aligned}$$

Since $(x_1, x_2) \in U^{\text{So}}(\mathcal{D}^{\Delta})$, we can bound $|x_1| \leq \max_{i \in \mathbb{K}} |U^{\text{ave}}(F^{(i)})|$. Since the reward is also bounded in $[0,1]$ we further have that $|x_1| \leq 1$, giving us the constants $b = (|r| + 2)/4\min\{\varepsilon_0, \varepsilon_0^{3/2}\}$ and $q = 1$.

Smoothness: First, we calculate the hessian to get that

$$\nabla^2 h(x_1, x_2) = \begin{pmatrix} 0 & \frac{1}{2(\varepsilon_0 - x_2)^{3/2}} \\ \frac{1}{2(\varepsilon_0 - x_2)^{3/2}} & \frac{3(x_1 - r)}{4(\varepsilon_0 - x_2)^{5/2}} \end{pmatrix}.$$

Next, we upper bound its spectral norm to get that. Let $w \in \mathbb{R}^2$ be such that $\|w\|_2 \leq 1$. Then we have that

$$\begin{aligned}
|w^T \nabla h(x_1, x_2) w| &= \left| \frac{w_1 w_2}{(\varepsilon_0 - x_2)^{3/2}} + w_2^2 \frac{3(x_1 - r)}{4(\varepsilon_0 - x_2)^{5/2}} \right| \\
&\leq \frac{1}{2(\varepsilon_0 - x_2)^{3/2}} + \frac{3|x_1 - r|}{4(\varepsilon_0 - x_2)^{5/2}} \\
&\leq \frac{2\varepsilon_0 + |x_1 - r|}{4\varepsilon_0^{5/2}}
\end{aligned}$$

Here we cannot take $M_0 = \infty$. However, for any $M_0 < \infty$ we can use the above bound to conclude that the smoothness assumption of Lemma 8 holds with

$$\beta = \frac{2\varepsilon_0 + D + M_0 + |r|}{4\varepsilon_0^{5/2}},$$

where under the assumption that the reward is in $[0,1]$ we further have that $D = 1$.

F.2.5. Sharpe ratio

Here we have that for $r \in \mathbb{R}$ and $\varepsilon_0 > 0$

$$h(x_1, x_2) = (x_1 - r)/\sqrt{\varepsilon_0 - x_1^2 + x_2}.$$

Linear gap: Let $\lambda \in \Delta$ and $F = \sum_{i=1}^{K} \lambda_i F^{(i)}$. We need to show that

$$U^{\text{Sh}}(F^{(i^*)}) - U^{\text{Sh}}(F) \geq \frac{1}{\eta} \|F^{(i^*)} - F\|.$$

Denote $x_i = U^{\text{ave}}\big(F^{(i)}\big)$, $y_i = U^{\text{sqr}}\big(F^{(i)}\big)$, and $z_i = \sqrt{\varepsilon_0 - x_i^2 + y_i}$. Notice that z_i is concave in x_i, y_i and so we have that

$$\sqrt{\varepsilon_0 - (\sum_{i=1}^K \lambda_i x_i)^2 + \sum_{i=1}^K \lambda_i y_i} \geq \sum_{i=1}^K \lambda_i z_i .$$

We conclude that

$$\begin{aligned}
U^{\text{Sh}}\big(F^{(i^*)}\big) - U^{\text{Sh}}\big(F\big) &= h\big(x_{i^*}, y_{i^*}\big) - h\big(\sum_{i=1}^K \lambda_i x_i, \sum_{i=1}^K \lambda_i y_i\big) \\
&\geq \frac{x_{i^*} - r}{z_{i^*}} - \frac{\sum_{i=1}^K \lambda_i x_i - r}{\sum_{i=1}^K \lambda_i z_i} \\
&= \frac{\sum_{i=1}^K \lambda_i \big(z_i (x_{i^*} - r) - z_{i^*}(x_i - r)\big)}{z_{i^*} \sum_{j=1}^K \lambda_j z_j} \\
&= \sum_{i=1}^K \lambda_i \frac{z_i}{\sum_{j=1}^K \lambda_i z_i} \big(\frac{x_{i^*} - r}{z_{i^*}} - \frac{x_i - r}{z_i}\big) \geq \frac{z_{\min}}{z_{\max}} \sum_{i=1}^K \lambda_i \Delta_i ,
\end{aligned}$$

where $z_{\min} = \min_{i \neq i^*} z_i$ and $z_{\max} = \max_{i \in \mathbb{K}} z_i$. Using the gap ratio defined in (13) we get a linear gap with $\eta = \rho z_{\max}/z_{\min}$.

Stability: For $x \in U^{\text{Sh}}\big(\mathcal{D}^{\Delta}\big)$ and $y \in U^{\text{Sh}}\big(L_{\|\cdot\|}\big)$ we have that

$$\begin{aligned}
|h\big(x_1, x_2\big) - h\big(y_1, y_2\big)| &= \left| \frac{x_1 - r}{\sqrt{\varepsilon_0 - x_1^2 + x_2}} - \frac{y_1 - r}{\sqrt{\varepsilon_0 - y_1^2 + y_2}} \right| \\
&\leq \left| \frac{x_1 - r}{\sqrt{\varepsilon_0 - y_1^2 + y_2}} - \frac{y_1 - r}{\sqrt{\varepsilon_0 - y_1^2 + y_2}} \right| + \left| \frac{x_1 - r}{\sqrt{\varepsilon_0 - x_1^2 + x_2}} - \frac{x_1 - r}{\sqrt{\varepsilon_0 - y_1^2 + y_2}} \right| \\
&\leq \frac{|x_1 - y_1|}{\varepsilon_0} + |x_1 - r| \left| \frac{\sqrt{\varepsilon_0 - y_1^2 + y_2} - \sqrt{\varepsilon_0 - x_1^2 + x_2}}{\sqrt{\varepsilon_0 - x_1^2 + x_2}\sqrt{\varepsilon_0 - y_1^2 + y_2}} \right| \\
&\leq \frac{|x_1 - y_1|}{\varepsilon_0} + \frac{|x_1 - r|}{2\varepsilon_0^{3/2}} \big|(x_1^2 - y_1^2) + (y_2 - x_2)\big| \\
&\leq \frac{|x_1 - y_1|}{\varepsilon_0} + |x_1 - r| \left| \frac{\sqrt{\varepsilon_0 - y_1^2 + y_2} - \sqrt{\varepsilon_0 - x_1^2 + x_2}}{\sqrt{\varepsilon_0 - x_1^2 + x_2}\sqrt{\varepsilon_0 - y_1^2 + y_2}} \right| \\
&\leq \frac{|x_1 - y_1|}{\varepsilon_0} + \frac{|x_1 - r|}{2\varepsilon_0^{3/2}} \big((x_1 - y_1)^2 + 2|x_1||x_1 - y_1| + |y_2 - x_2|\big) \\
&\leq \max\Big\{\varepsilon_0^{-1}, 2\varepsilon_0^{-3/2}|x_1 - r|\Big\} \big((x_1 - y_1)^2 + (2|x_1| + 1)|x_1 - y_1| + |y_2 - x_2|\big) \\
&\leq 2(1 + |x_1|) \max\Big\{\varepsilon_0^{-1}, 2\varepsilon_0^{-3/2}|x_1 - r|\Big\} \Big(\|x - y\|_2^2 + \|x - y\|_2\Big).
\end{aligned}$$

Since $(x_1, x_2) \in U^{\text{Sh}}\big(\mathcal{D}^{\Delta}\big)$, we can bound $|x_1| \leq \max_{i \in \mathbb{K}} |U^{\text{ave}}\big(F^{(i)}\big)|$. Since the reward is also bounded in $[0,1]$ we further have that $|x_1| \leq 1$, giving us the constants $b = 4\max\{\varepsilon_0^{-1}, 2\varepsilon_0^{-3/2}(|r|+1)\}$ and $q = 2$.

Smoothness: The idea here is to bound the spectral norm of the hessian. This is very similar to previous examples and in particular to Sortino ratio.

F.3. Non-composite EDPMs. In this section we show the properties of VaR_α and $CVaR_\alpha$ required by our framework. Unless stated otherwise, we use the norm defined in (18). We recall the definitions of $CVaR_\alpha$ and VaR_α from (17) and (20), and specifically that we have that

$$U^{CVaR_\alpha}\big(F\big) = z^* - \frac{1}{\alpha} \int_{-\infty}^{z^*} F(x) dx,$$

where $z^* = U^{VaR_\alpha}(F)$. Before starting, we require the following technical lemma, proved in Section F.3.3.

Lemma 9 ($CVaR_\alpha$ **and VaR$_\alpha$ bounds**). *For any $F, G \in L_{\|\cdot\|}$ we have that*

$$|U^{VaR_\alpha}(F) - U^{VaR_\alpha}(G)| \le \frac{2\|F\| + \|F-G\|}{\min\{\alpha, 1-\alpha\}},$$

and

$$\begin{aligned}0 \le U^{CVaR_\alpha}(G) - U^{CVaR_\alpha}(F) + \frac{1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(F)} (G(x)-F(x))dx \\ \le \frac{1}{\alpha}\|F-G\| |U^{VaR_\alpha}(F) - U^{VaR_\alpha}(G)|.\end{aligned}$$

If additionally (19) *holds, $F \in \mathcal{D}^\Delta$, and $\|F-G\| < M_\alpha$ then we also have that*

$$|U^{VaR_\alpha}(F) - U^{VaR_\alpha}(G)| \le b_\alpha \|F-G\|_\infty.$$

F.3.1. Conditional Value at Risk (CVaR) The following summarizes the properties of $CVaR_\alpha$ required for applying Theorem 4.

Proposition 4 (**CVaR$_\alpha$ properties**). *We have that:*

1. *$CVaR_\alpha$ is convex;*
2. *$CVaR_\alpha$ is stable with parameters $b = 4/\alpha \min\{\alpha, 1-\alpha\}, q = 2, \upsilon = 2/3$;*
3. *If* (19) *holds then $CVaR_\alpha$ is smooth with parameters $\beta = 2b_\alpha/\alpha$ and $M_0 = M_\alpha$.*

Proof. As previously mentioned, convexity follows from (17), which expresses $CVaR_\alpha$ as a maximum over linear functions.

Stability. Starting with the easier Requirement 2 of stability, the concentration of $\|\hat{F}_t^{(i)} - F^{(i)}\|_\infty$ follows from the Dvoretzky-Kiefer-Wolfowitz inequality [19] with $\upsilon_0 = 2$, and since the other two terms are linear, the same holds for them by Hoeffding's inequality. As in Lemma 8 (but for *max* norm), we conclude that Requirement 2 holds with $\upsilon = 2\log 2/\log 6 \ge 2/3$. Next, for Requirement 1, first notice that

$$\begin{aligned}\left|\frac{1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(F)} (F(x)-G(x))dx\right| &\le \left|\frac{1}{\alpha}\int_{-\infty}^{0} (F(x)-G(x))dx\right| + \left|\frac{1}{\alpha}\int_{0}^{U^{VaR_\alpha}(F)} (F(x)-G(x))dx\right| \\ &\le \frac{1}{\alpha}\left[\|F-G\| + |U^{VaR_\alpha}(F)|\|F-G\|\right] \\ &\le \frac{\|F-G\|}{\alpha}\left[1 + |U^{VaR_\alpha}(F)|\right],\end{aligned}$$

and apply this to the second claim of Lemma 9 to get that

$$|U^{CVaR_\alpha}(G) - U^{CVaR_\alpha}(F)| \le \frac{\|F-G\|}{\alpha}\left[1 + |U^{VaR_\alpha}(F)| + |U^{VaR_\alpha}(F) - U^{VaR_\alpha}(G)|\right].$$

Finally, using the first part of Lemma 9 we get that

$$\begin{aligned}|U^{CVaR_\alpha}(G) - U^{CVaR_\alpha}(F)| &\le \frac{\|F-G\|}{\alpha}\left[1 + |U^{VaR_\alpha}(F)| + \frac{2\|F\| + \|F-G\|}{\min\{\alpha, 1-\alpha\}}\right] \\ &\le \frac{1}{\alpha}\left(1 + |U^{VaR_\alpha}(F)| + \frac{\max\{1, 2\|F\|\}}{\min\{\alpha, 1-\alpha\}}\right)\left[\|F-G\| + \|F-G\|^2\right],\end{aligned}$$

which is Requirement 1 of stability. Further using Assumption 1, we have that $\|F\|, |U^{VaR_\alpha}(F)| \le 1$, and plugging this into the above gives the desired value for b.

Smoothness. Assume that $\partial U^{CVaR_\alpha}(F)\cdot G = -\frac{1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(F)} G(x)dx$. We show that the smoothness condition holds under this assumption, which in turn implies that this assumption must be true (see definition of Frechet derivative). To that end, we use the second and third parts of Lemma 9 to get that for any $F\in\mathcal{D}^\Delta$ and $G\in L_{\|\cdot\|}$ satisfying $\|F-G\|\le M_\alpha$ we have that

$$\begin{aligned}&\left|U^{CVaR_\alpha}(G) - U^{CVaR_\alpha}(F) - \frac{-1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(F)}(G(x)-F(x))dx\right|\\ &\quad\le \frac{1}{\alpha}\|F-G\||U^{VaR_\alpha}(F)-U^{VaR_\alpha}(G)|\\ &\quad\le \frac{1}{2}\frac{2b_\alpha}{\alpha}\|F-G\|^2,\end{aligned}$$

as desired. ■

F.3.2. Value at Risk (VaR) We begin with the following proposition, which proves the needed properties of VaR_α to get $\mathcal{O}(1/\sqrt{T})$ regret, as described in Remark 1.

Proposition 5 (VaR$_\alpha$ properties). *We have that:*

1. VaR_α *is quasiconvex;*
2. *If* (19) *holds then* VaR_α *is stable with parameters* $\upsilon = 2/3, q = 1$, *and*

$$b = \max\left\{b_\alpha, \frac{M_\alpha + 2}{\min\{\alpha, 1-\alpha\}M_\alpha}\right\}.$$

Proof. Starting with quasiconvexity, let $F_1, F_2\in\mathcal{D}$ and $\lambda\in[0,1]$. Denote $F_\lambda = \lambda F_1 + (1-\lambda)F_2$, then by the definition of U^{VaR_α} we have that

$$\begin{aligned}F_\lambda\left(\max\{U^{VaR_\alpha}(F_1), U^{VaR_\alpha}(F_2)\}\right) &= \lambda F_1\left(\max\{U^{VaR_\alpha}(F_1), U^{VaR_\alpha}(F_2)\}\right)\\ &\quad + (1-\lambda)F_2\left(\max\{U^{VaR_\alpha}(F_1), U^{VaR_\alpha}(F_2)\}\right)\ge\alpha.\end{aligned}$$

Using the definition of U^{VaR_α} another time we conclude that

$$U^{VaR_\alpha}(F_\lambda)\le\max\{U^{VaR_\alpha}(F_1), U^{VaR_\alpha}(F_2)\},$$

which is one of the characterizations of quasiconvexity.

Stability. Since we use the same norm as $CVaR_\alpha$, Requirement 2 is proven by Proposition 4. As for Requirement 1, if $\|F-G\| < M_\alpha$ then using the third part of Lemma 9 we have that

$$|U^{VaR_\alpha}(F) - U^{VaR_\alpha}(G)|\le b_\alpha\|F-G\|.$$

On the other hand, if $\|F-G\|\ge M_\alpha$ then using the first part of Lemma 9 we have that

$$\begin{aligned}|U^{VaR_\alpha}(F) - U^{VaR_\alpha}(G)| &\le \frac{2\|F\| + \|F-G\|}{\min\{\alpha,1-\alpha\}}\\ &= \frac{\|F-G\|\left(1+\frac{2\|F\|}{\|F-G\|}\right)}{\min\{\alpha,1-\alpha\}}\\ &\le \frac{1+\frac{2\|F\|}{M_\alpha}}{\min\{\alpha,1-\alpha\}}\|F-G\|,\end{aligned}$$

and combining both results we obtain get that stability holds with $q=1$ and

$$b = \max\left\{b_\alpha, \frac{M_\alpha + 2\|F\|}{\min\{\alpha,1-\alpha\}M_\alpha}\right\}.$$

Further using Assumption 1, we have that $\|F\|\le 1$, which gives the desired value of b. ■

The following result shows when VaR_α satisfies the conditions of Theorem 1, and thus has a single arm infinite horizon oracle policy. We note that this holds regardless as shown Proposition 3, which uses a different approach that is specific to VaR_α. Denote the α level set of a function $F \in \mathcal{D}$ by $L_\alpha(F) = \{x \in \mathbb{R} \big| F(x) = \alpha\}$.

Proposition 6 (VaR$_\alpha$ Theorem 1 conditions). *If $|L_\alpha(\alpha)| \le 1$ for all $F \in \mathcal{D}^\Delta$. Then U^{VaR_α} satisfies then conditions of Theorem 1, and thus has a single arm infinite horizon oracle policy.*

Proof. Unlike the remainder of this section, here we use the norm $\|F\| = \|F\|_\infty$. By the Glivenko-Cantelli theorem [28], the convergence of the empirical distribution required in Theorem 1 is established. It thus remains to show that VaR_α is continuous on $\mathcal{D}^\Delta$. Our condition on the level set can be interpreted in the following way. For any fixed $F \in \mathcal{D}^\Delta$

$$y > U^{VaR_\alpha}\big(F\big) \implies \exists c_y > 0,\ s.t,\ F(y) \ge \alpha + c_y \tag{27}$$
$$y < U^{VaR_\alpha}\big(F\big) \implies \exists c_y > 0,\ s.t,\ F(y) \le \alpha - c_y. \tag{28}$$

Let $g : [-\frac{\alpha}{2}, \frac{1-\alpha}{2}] \to \mathbb{R}$ be given by, $g(\delta) = U^{VaR_{\alpha+\delta}}\big(F\big)$. We show that g is continuous at 0. g is monotone non decreasing and so has left and right limits at 0. Let $\{\delta_n\}_{n=1}^\infty \searrow 0$ and denote,

$$\lim_{n\to\infty} g(\delta_n) = a^+.$$

By the monotonicity of g we have that $a^+ \ge g(0)$. Using (27) we have that, for any $\varepsilon > 0$,

$$F(g(0)+\varepsilon) \ge \alpha + c_\varepsilon,$$

where $c_\varepsilon > 0$. So, by the expression of U^{VaR_α}, we have that,

$$g(0) + \varepsilon \ge g(c_\varepsilon) \ge a^+,$$

where the second inequality follows by the monotonicity of g. So, $g(0) \le a^+ \le g(0) + \varepsilon$ for all $\varepsilon > 0$ and so $a^+ = g(0)$. Now take $\big\{\bar{\delta}_n\big\}_{n=1}^\infty \nearrow 0$ and denote,

$$\lim_{n\to\infty} g(\bar{\delta}_n) = a^-.$$

A similar set of arguments shows that $a^- = g(0)$, and so g is continuous at 0.

By the continuity of g, for any $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that for all $|\beta| \le \delta_\varepsilon$ we have that,

$$|g(0) - g(\beta)| \le \varepsilon.$$

For any $G \in L_{\|\cdot\|}$ satisfying $\|F - G\|_\infty \le \delta_\varepsilon$ we have that,

$$\begin{aligned} U^{VaR_\alpha}\big(F\big) - U^{VaR_\alpha}\big(G\big) &= g(0) - \min\Big\{y \,\Big|\, G(y) \ge \alpha\Big\} \\ &\le g(0) - \min\Big\{y \,\Big|\, F(y) \ge \alpha - \delta_\varepsilon\Big\} = g(0) - g(-\delta_\varepsilon) \le \varepsilon. \end{aligned}$$

We also have,

$$\begin{aligned} U^{VaR_\alpha}\big(G\big) - U^{VaR_\alpha}\big(F\big) &= \min\Big\{y \,\Big|\, G(y) \ge \alpha\Big\} - g(0) \\ &\le \min\Big\{y \,\Big|\, F(y) \ge \alpha + \delta_\varepsilon\Big\} - g(0) = g(\delta_\varepsilon) - g(0) \le \varepsilon. \end{aligned}$$

We conclude that, $|U^{VaR_\alpha}\big(F\big) - U^{VaR_\alpha}\big(G\big)| \le \varepsilon$ thus concluding the continuity of U^{VaR_α} on $\mathcal{D}^\Delta$. ■

Finally, we prove Proposition 3, showing that VaR_α always has a single arm infinite horizon oracle policy.

Proof of Proposition 3. We begin calculating the performance of a single arm policy. We then proceed to show that the performance of any policy is upper bounded by that of the best single arm policy.

Performance of simple policies. Let π^i be the policy that always plays arm i. We claim that, $\mathbf{E}U_{\pi^i}^{VaR_\alpha} = U^{VaR_\alpha}\big(F^{(i)}\big)$. If $F^{(i)}$ represents a degenerate random variable then the expression holds trivially. Otherwise, let $y < U^{VaR_\alpha}\big(F^{(i)}\big) = a_i$, then there exists $\delta_y > 0$, such that, $F^{(i)}(y) \le \alpha - \delta_y$. Using the strong law of large numbers [22] we have that

$$\lim_{t\to\infty} \hat{F}_t^{\pi^i}(y) \overset{a.s}{=} F^{(i)}(y) \le \alpha - \delta_y.$$

Let E be the event on which the convergence occurs. Then $\forall \omega \in E$ there exists $T(\omega)$ such that $\forall t > T(\omega)$, we have that,

$$\hat{F}_t^{\pi^i}(y,\omega) \le F^{(i)}(y) + \delta_y/2 \le \alpha - \delta_y/2 < \alpha.$$

This implies that $U^{VaR_\alpha}\big(\hat{F}_t^{\pi^i}(\omega)\big) > y$ for all $t \ge T(\omega)$. We get that $U_{\pi^i}^{VaR_\alpha} \ge y$ almost surely, and taking the expectation we get $\mathbf{E}U_{\pi^i}^{VaR_\alpha} \ge y$. Since this holds for all $y < U^{VaR_\alpha}\big(F^{(i)}\big)$, then,

$$\mathbf{E}U_{\pi^i}^{VaR_\alpha} \ge U^{VaR_\alpha}\big(F^{(i)}\big). \tag{29}$$

On the other hand, using the Law of the iterated logarithm [14] we get that

$$\limsup_{t\to\infty} \frac{t}{\lambda\sqrt{2t\log\log t}}\Big(\hat{F}_t^{\pi^i}(a_i) - F^{(i)}(a_i)\Big) = 1 \quad a.s,$$

where, $\lambda = F^{(i)}(a_i)\big(1 - F^{(i)}(a_i)\big) \ne 0$ since $F^{(i)}$ is non-degenerate. We conclude that

$$\hat{F}_t^{\pi^i}(a_i) > F^{(i)}(a_i) \ge \alpha \quad i.o,$$

and thus

$$U_{\pi^i}^{VaR_\alpha} = \liminf_{t\to\infty} U^{VaR_\alpha}\big(\hat{F}_t^{\pi^i}\big) \le a_i \quad a.s. \tag{30}$$

Taking expectation on both sides, we conclude that

$$\mathbf{E}U_{\pi^i}^{VaR_\alpha} \le U^{VaR_\alpha}\big(F^{(i)}\big),$$

which together with (29) proves that $\mathbf{E}U_{\pi^i}^{VaR_\alpha} = U^{VaR_\alpha}\big(F^{(i)}\big)$. Now, recall that in Proposition 5 we showed that U^{VaR_α} is quasiconvex. We thus have that there exists $i^* \in \mathbb{K}$ such that for all $F \in \mathcal{D}^\Delta$

$$U^{VaR_\alpha}\big(F\big) \le U^{VaR_\alpha}\big(F^{(i^*)}\big) = \mathbf{E}U_{\pi^{i^*}}^{VaR_\alpha} = a^*. \tag{31}$$

Global optimizer. Our purpose will be to show that

$$\hat{F}_t^{\pi}(a^*) - \alpha > 0 \quad i.o. \tag{32}$$

Similarly to (30), this implies that

$$U_{\pi}^{VaR_\alpha} \le a^* \quad a.s,$$

and taking the expectation we conclude that, $\mathbf{E}U_{\pi}^{VaR_\alpha} \le \mathbf{E}U_{\pi^{i^*}}^{VaR_\alpha}$, thus concluding the proof.

By (31), we have that

$$F_t^\pi(a^*) = \frac{1}{t}\sum_{s=1}^{t} F^{(\pi_s)}(a^*) \geq \alpha.$$

We thus get that,

$$\hat{F}_t^\pi(a^*) - \alpha \geq \hat{F}_t^\pi(a^*) - F_t^\pi(a^*) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}\{X_{\pi,s} \leq a^*\} - F^{(\pi_s)}(a^*) = \frac{1}{t}\sum_{s=1}^{t} Y_s = \frac{1}{t} W_t, \tag{33}$$

where $Y_s = \left[\mathbb{1}\{X_{\pi,s} \leq a^*\} - F^{(\pi_s)}(a^*)\right]$ and $W_t = \sum_{s=1}^{t} Y_s$. We split our remaining analysis into two cases.

The first is when policy π chooses some non-degenerate arm infinitely often (i.o). For this case we use the Law of the iterated logarithm for martingales given in [10]. We use the same notation as in [10] aside for denoting the martingale W_t instead of U_t, and its difference sequence by Y_t instead of X_t (to avoid confusion with existing notation). We start by showing W_t is a martingale with respect to its natural filtration

$$\begin{aligned}
&\mathbf{E}[W_{t+1}|W_1,\ldots,W_t] \\
&\quad = W_t + \mathbf{E}[Y_{t+1}|W_1,\ldots,W_t] \\
&\quad = W_t + \mathbf{E}\left[\mathbf{E}\left[Y_{t+1}|\pi_{t+1}\right]|W_1,\ldots,W_t\right] \\
&\quad = W_t + \mathbf{E}[0|W_1,\ldots,W_t] = W_t,
\end{aligned}$$

where the second equality is the law of total probability in addition to $Y_{t+1}|\pi_{t+1}$ being independent of $W_1,\ldots,W_t$. Furthermore, $\mathbf{E}|W_t| \leq t < \infty$, so W_t is a martingale.

Next, let $s_t^2 = \sum_{s=1}^{t} \mathbf{E}[Y_s^2|W_1,\ldots,W_{s-1}]$. Since π chooses a non-degenerate arm infinitely often then, $s_t^2 \to \infty$.

Finally, let t_0 denote the first time π chooses a non-degenerate arm. So, we can choose K_t in the following way,

$$K_t = \varphi(s_{t_0})/s_{t_0}.$$

Clearly, there exists $K > 0$ such that, $\limsup_{t\to\infty} K_t < K$. Furthermore,

$$|Y_t| \leq \begin{cases} 0, & t < t_0 \\ 1, & t \geq t_0 \end{cases} \leq K_t s_t/\varphi(s_t).$$

So the conditions of Theorem 1 in [10] are met and we conclude that

$$\limsup_{t\to\infty} W_t/s_t\varphi(s_t) > 0 \quad a.s.$$

This means that, $W_t > 0$ infinitely often and substituting into (33) we conclude that (32) holds.

In the second case, any non-degenerate arm is chosen a finite number of times. Let i_b denote the index of the largest degenerate arm and a_b be its value. Clearly,

$$a_b = \mathbf{E}U_{\pi^{i_b}}^{VaR_\alpha} \leq \mathbf{E}U_{\pi^{e_{i^*}}}^{VaR_\alpha} = a^*.$$

Denote, $I_b = \left\{ i \,\middle|\, F^{(i)} \text{ is degenerate} \right\}$. Since non-degenerate arms are pulled a finite number of times then,

$$\lim_{t\to\infty} \hat{p}_i(t) = 0 \quad, \forall i \notin I_b,$$

where $\hat{p}_i(t)$ is defined in (21). Since there are finitely many arms, this implies that

$$\lim_{t\to\infty}\sum_{i\in I_b}\hat{p}_i(t)=1.$$

Now, since a_b is such that $F^{(i)}(a_b)=\hat{F}_t^{(i)}(a_b)=1$ for all $i\in I_b$, we conclude that

$$\begin{aligned}\lim_{t\to\infty}\hat{F}_t^{\pi}(a^*)&\geq\lim_{t\to\infty}\hat{F}_t^{\pi}(a_b)\\&=\lim_{t\to\infty}\sum_{i=1}^{K}\hat{p}_i(t)\hat{F}_t^{(i)}(a_b)\\&\geq\lim_{t\to\infty}\sum_{i\in I_b}\hat{p}_i(t)\hat{F}_t^{(i)}(a_b)\\&=\lim_{t\to\infty}\sum_{i\in I_b}\hat{p}_i(t)=1.\end{aligned}$$

Since $\alpha<1$, we can clearly conclude (32) holds, thus finishing the proof. ■

F.3.3. Proof of Lemma 9 We prove the individual claims of Lemma 9.

First claim. We start by showing that for all $G\in L_{\|\cdot\|}$

$$|U^{VaR_\alpha}(G)|\leq\frac{\|G\|}{\min\{\alpha,1-\alpha\}}.\tag{34}$$

Suppose that $U^{VaR_\alpha}(G)\geq 0$, then we have that

$$\begin{aligned}\|G\|\geq\int_0^{\infty}(1-G(y))dy&\geq\int_0^{U^{VaR_\alpha}(G)}(1-G(y))dy\\&\geq\int_0^{U^{VaR_\alpha}(G)}(1-\alpha)dy=(1-\alpha)U^{VaR_\alpha}(G)\geq 0.\end{aligned}$$

On the other hand, suppose that $U^{VaR_\alpha}(G)\leq 0$, then we have that

$$\begin{aligned}\|G\|\geq\int_{-\infty}^{0}G(y)dy&\geq\int_{U^{VaR_\alpha}(G)}^{0}G(y)dy\\&\geq\int_{U^{VaR_\alpha}(G)}^{0}\alpha dy=-\alpha U^{VaR_\alpha}(G)\geq 0,\end{aligned}$$

thus concluding (34). Using this result, we have that for all $F,G\in L_{\|\cdot\|}$

$$\begin{aligned}|U^{VaR_\alpha}(F)-U^{VaR_\alpha}(G)|&\leq|U^{VaR_\alpha}(F)|+|U^{VaR_\alpha}(G)|\\&\leq\frac{\|F\|+\|G\|}{\min\{\alpha,1-\alpha\}}\\&\leq\frac{2\|F\|+\|F-G\|}{\min\{\alpha,1-\alpha\}},\end{aligned}$$

as desired.

Second claim. We have that

$$U^{CVaR_\alpha}(G)-U^{CVaR_\alpha}(F)$$

$$
\begin{aligned}
&= U^{VaR_\alpha}\big(G\big) - \frac{1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(G)} G(y)dy - U^{VaR_\alpha}\big(F\big) + \frac{1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(F)} F(y)dy \\
&= U^{VaR_\alpha}\big(G\big) - U^{VaR_\alpha}\big(F\big) + \frac{1}{\alpha}\int_{U^{VaR_\alpha}(G)}^{U^{VaR_\alpha}(F)} G(y)dy - \frac{1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(F)} (G(y)-F(y))dy \\
&= \frac{1}{\alpha}\int_{U^{VaR_\alpha}(G)}^{U^{VaR_\alpha}(F)} (G(y)-\alpha)dy - \frac{1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(F)} (G(y)-F(y))dy,
\end{aligned}
$$

and changing sides we get

$$
U^{CVaR_\alpha}\big(G\big) - U^{CVaR_\alpha}\big(F\big) + \frac{1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(F)} (G(y)-F(y))dy = \frac{1}{\alpha}\int_{U^{VaR_\alpha}(G)}^{U^{VaR_\alpha}(F)} (G(y)-\alpha)dy.
$$

It therefore suffices to show that

$$
0 \leq \frac{1}{\alpha}\int_{U^{VaR_\alpha}(G)}^{U^{VaR_\alpha}(F)} (G(y)-\alpha)dy \leq \frac{1}{\alpha}\|F-G\||U^{VaR_\alpha}\big(F\big) - U^{VaR_\alpha}\big(G\big)|.
$$

Beginning with the left inequality, we have that

$$
\begin{aligned}
&\frac{1}{\alpha}\int_{U^{VaR_\alpha}(G)}^{U^{VaR_\alpha}(F)} (G(y)-\alpha)dy \\
&\geq \frac{1}{\alpha}\big(U^{VaR_\alpha}\big(F\big) - U^{VaR_\alpha}\big(G\big)\big)\big(G\big(\min\{U^{VaR_\alpha}\big(G\big), U^{VaR_\alpha}\big(F\big)\}\big) - \alpha\big) \\
&= \begin{cases} \frac{1}{\alpha}\underbrace{\big(U^{VaR_\alpha}\big(F\big) - U^{VaR_\alpha}\big(G\big)\big)}_{\leq 0}\underbrace{\big(G\big(U^{VaR_\alpha}\big(F\big)\big) - \alpha\big)}_{\leq 0}, U^{VaR_\alpha}\big(G\big) \geq U^{VaR_\alpha}\big(F\big) \\ \frac{1}{\alpha}\underbrace{\big(U^{VaR_\alpha}\big(F\big) - U^{VaR_\alpha}\big(G\big)\big)}_{\geq 0}\underbrace{\big(G\big(U^{VaR_\alpha}\big(G\big)\big) - \alpha\big)}_{\geq 0}, U^{VaR_\alpha}\big(G\big) < U^{VaR_\alpha}\big(F\big) \end{cases} \\
&\geq 0,
\end{aligned}
\tag{35}
$$

where the final inequality holds since $G(y) \underset{<}{\geqq} \alpha$ for $y \underset{<}{\geqq} U^{VaR_\alpha}\big(G\big)$. Next, for the right hand side inequality we have that

$$
\begin{aligned}
&\frac{1}{\alpha}\int_{U^{VaR_\alpha}(G)}^{U^{VaR_\alpha}(F)} (G(y)-\alpha)dy \\
&= \frac{1}{\alpha}\int_{U^{VaR_\alpha}(G)}^{U^{VaR_\alpha}(F)} (F(y)-\alpha)dy + \frac{1}{\alpha}\int_{U^{VaR_\alpha}(G)}^{U^{VaR_\alpha}(F)} (G(y)-F(y))dy \\
&\leq \underbrace{\frac{1}{\alpha}\int_{U^{VaR_\alpha}(G)}^{U^{VaR_\alpha}(F)} (F(y)-\alpha)}_{(*)} + \frac{1}{\alpha}\|F-G\|_\infty |U^{VaR_\alpha}\big(F\big) - U^{VaR_\alpha}\big(G\big)| \\
&\leq \frac{1}{\alpha}\|F-G\||U^{VaR_\alpha}\big(F\big) - U^{VaR_\alpha}\big(G\big)|,
\end{aligned}
$$

where $(*) \leq 0$ is obtained by exchanging the roles of F and G in (35).

Third claim. Consider (19) with $y = \|F-G\|_\infty \leq M_\alpha$, and notice that $F\big(U^{VaR_\alpha}\big(F\big) + y\big) \geq \alpha$ for any $y \geq 0$. Then we have that

$$
F\big(U^{VaR_\alpha}\big(F\big) + b_\alpha\|F-G\|_\infty\big) - \alpha \geq \|F-G\|_\infty.
$$

Using this we get

$$G\big(U^{VaR_\alpha}(F)+b_\alpha\|F-G\|_\infty\big)\geq F\big(U^{VaR_\alpha}(F)+b_\alpha\|F-G\|_\infty\big)-\|F-G\|_\infty\geq\alpha,$$

which by the definition of U^{VaR_α} in (20) implies that

$$U^{VaR_\alpha}(G)\leq U^{VaR_\alpha}(F)+b_\alpha\|F-G\|_\infty,$$

and changing sides we get

$$U^{VaR_\alpha}(G)-U^{VaR_\alpha}(F)\leq b_\alpha\|F-G\|_\infty. \tag{36}$$

On the other hand, consider (19) with $y=-c\|F-G\|_\infty\geq -M_\alpha$, where $1<c\leq\frac{M_\alpha}{\|F-G\|_\infty}$. Noticing that $F\big(U^{VaR_\alpha}(F)-y\big)\leq\alpha$ for any $y\geq 0$, we have that

$$-\big(F\big(U^{VaR_\alpha}(F)-cb_\alpha\|F-G\|_\infty\big)-\alpha\big)\geq c\|F-G\|_\infty,$$

and changing sides we get

$$F\big(U^{VaR_\alpha}(F)-cb_\alpha\|F-G\|_\infty\big)\leq\alpha-c\|F-G\|_\infty.$$

Using this we get

$$\begin{aligned}G\big(U^{VaR_\alpha}(F)-cb_\alpha\|F-G\|_\infty\big)&\leq F\big(U^{VaR_\alpha}(F)-cb_\alpha\|F-G\|_\infty\big)+\|F-G\|_\infty\\&\leq\alpha+(1-c)\|F-G\|_\infty\\&<\alpha,\end{aligned}$$

and using the definition of U^{VaR_α} in (20) and changing sides we get

$$U^{VaR_\alpha}(G)-U^{VaR_\alpha}(F)\geq -cb_\alpha\|F-G\|_\infty.$$

Notice that the constant c may be arbitrarily close to 1. We thus conclude that

$$U^{VaR_\alpha}(G)-U^{VaR_\alpha}(F)\geq -b_\alpha\|F-G\|_\infty,$$

which combined with (36) implies the desired.

F.3.4. Counter examples

$\mathbf{U^{bad1}}$ *details.* Examples such as VaR_α are rather uncommon, but, when encountered, their analysis proves challenging. This fact might motivate a more general framework for EDPMs, ideas for which, can be drawn from the proof of Proposition 3. The following examples show the types of problems such frameworks could have or would need to address. Define an EDPM by,

$$U^{bad1}(F)=U^{VaR_{0.1}}(F)+U^{VaR_{0.9}}(F),$$

where the values 0.1, 0.9 were chosen arbitrarily. When the two components of U^{bad1} are stable then it is clear that so is U^{bad1}. We show that when this is not the case, then it is possible that no simple policy is optimal. Consider a problem with two arms having the following distributions,

$$F^{(1)}(y)=\begin{cases}0 & y<0\\ y/10 & 0\leq y<1\\ 0.1 & 1\leq y<5\\ y/50 & 5\leq y<50\\ 1 & y\geq 50\end{cases}\qquad F^{(2)}(y)=\begin{cases}0 & y<5\\ 1 & y\geq 5.\end{cases}$$

Notice that $F^{(1)}$, $F^{(2)}$ satisfy the conditions of Theorem 1. So, using an intermediate result of Theorem 1 we have that $\lim_{t\to\infty} U^{VaR_{0.9}}\left(\hat{F}_t^{\pi^p}\right) = U^{VaR_{0.9}}\left(F_p\right)$ almost surely. Using this convergence we get that

$$\begin{aligned}\liminf_{t\to\infty} U^{bad1}\left(\hat{F}_t^{\pi^p}\right) &\overset{a.s}{=} \liminf_{t\to\infty} U^{VaR_{0.1}}\left(\hat{F}_t^{\pi^p}\right) + \lim_{t\to\infty} U^{VaR_{0.9}}\left(\hat{F}_t^{\pi^p}\right)\\ &\overset{a.s}{=} U^{VaR_{0.1}}\left(F_p\right) + U^{VaR_{0.9}}\left(F_p\right).\end{aligned}$$

where the convergence of $U^{VaR_{0.1}}\left(\hat{F}_t^{\pi^p}\right)$ is an intermediate result in Proposition 3. Evaluating these last terms we conclude the expression for the performance of a simple policy π^p $(p=(p_1,p_2),\ p_1 = 1-p_2)$,

$$\mathbf{E}U^{bad1}_{\pi^p} = \begin{cases} 46 & p_2 = 0\\ 5+(45-50p_2)/(1-p_2) & 0<p_2<8/9\\ 10 & 8/9<p_2\leq 1\end{cases}$$

It does not attain a maximum over the simplex and thus there is no optimizer inside the set of simple policies. However, the following non-simple policy is optimal,

$$\pi_t^{*bad1} = \begin{cases} 2 & t=1 \quad or \quad \frac{(t-1)\hat{F}^{\pi}_{t-1}(1)+1}{t} \geq 0.1\\ 1 & \text{otherwise.}\end{cases}$$

We explain why π^*_{bad1} is an oracle policy. Each of the summands in U^{bad1} has a simple oracle policy with appropriate optimal performance. Summing these performances provides an upper bound on the performance of U^{bad1}. More specifically, U^{bad1} is bounded by 50. We show that, π^*_{bad1} achieves this value and is thus an oracle policy. It is easy to show by induction that for $(t\geq 1)$, $\hat{F}_t^{\pi^{*bad1}}(1) < 0.1$. This implies that $U^{VaR_{0.1}}\left(\hat{F}_t^{\pi^{*bad1}}\right)\geq 5$, but 5 is also an upper bound an so, $\liminf_{t\to\infty} U^{VaR_{0.1}}\left(\hat{F}_t^{\pi^{*bad1}}\right) = 5$. Finally, it is a technical result to show that $\lim_{t\to\infty}\hat{F}_t^{\pi^{*bad1}}(5) = 0.1$ almost surely, which implies $\lim_{t\to\infty}\hat{p}(t)\overset{a.s}{=} e_1 = (1,0)$. Since $U^{VaR_{0.9}}$ is stable then this implies that

$$\lim_{t\to\infty} U^{VaR_{0.9}}\left(\hat{F}_t^{\pi^{*bad1}}\right)\overset{a.s}{=} U^{VaR_{0.9}}\left(F^{(1)}\right) = 45,$$

and taking expectation the result is concluded.

The problem exhibited here is the lack of an optimizer within the set of simple policies. A way of ensuring that this does not occur is to require that the performance of simple policies be upper semi-continuous (with respect to p).

$\mathbf{U^{bad2}}$ ***details.*** The following example shows that when the performance of simple policies is not lower semi continuous, then while an optimizer exists within the set of simple policies, it might not be a global optimizer. Define an EDPM by,

$$U^{bad2}\left(F\right) = U^{VaR^{++}_{0.1}}\left(F\right) + 5\mathbb{1}\left\{F(10^-)-F(1^+)>0 \ or \ F(1^-)>0\right\},$$

where,

$$\begin{aligned} U^+\left(F;x\right) &= \max_{y\in\mathbb{R}}\left\{y\geq x \;\middle|\; F(y)=F(x)\right\}\\ U^{VaR^+_{0.1}}\left(F\right) &= \begin{cases} U^{VaR_{0.1}}\left(F\right) & if\ F(U^{VaR_{0.1}}\left(F\right))=1\\ U^+\left(F;U^{VaR_{0.1}}\left(F\right)\right) & otherwise.\end{cases}\\ U^{VaR^{++}_{0.1}}\left(F\right) &= \begin{cases} U^{VaR^+_{0.1}}\left(F\right) & if\ F(U^{VaR^+_{0.1}}\left(F\right))=1\\ U^+\left(F;U^{VaR^+_{0.1}}\left(F\right)\right) & otherwise.\end{cases}\end{aligned}$$

Consider a problem with two arms having the following distributions,

$$F^{(1)}(y) = \begin{cases} 0 & y<0 \\ 0.9+y/100 & 0 \le y < 10 \\ 1 & y \ge 10 \end{cases}, F^{(2)}(y) = \begin{cases} 0 & y<1 \\ 0.1 & 1 \le y < 10 \\ 1 & y \ge 10. \end{cases}$$

The performance of a simple policy π^p $(p=(p_1,p_2),\ p_1 = 1-p_2)$ is given by,

$$\mathbf{E}U_{\pi^p}^{bad2} = \begin{cases} 5 & p_2 < 8/9 \\ -85+10/(1-p_2) & 8/9 \le p_2 < 81/91 \\ 6 & 81/91 \le p_2 < 1 \\ 10 & p_2 = 1, \end{cases}$$

which attains a maximum for $p_2 = 1$. However, the resulting policy is not an oracle policy. The following non-simple policy is an oracle policy,

$$\pi_t^{*bad2} = \begin{cases} 1 & t=1 \\ 2 & otherwise. \end{cases}$$

To show this, proceed as for U^{bad1}, i.e., see that the individual components are bounded by 10 and 5 respectively, and so the optimal performance is at most 15. It is trivial to verify that π^{*bad2} obtains this reward, thus showing it is an oracle policy. Finally, we describe how $\mathbf{E}U_{\pi^p}^{bad2}$ is obtained. It is easily seen that for all $t \ge 1$,

$$\mathbb{1}\left\{\hat{F}_t^{\pi^p}(10^-) - \hat{F}_t^{\pi^p}(1^+) > 0 \ or\ \hat{F}_t^{\pi^p}(1^-) > 0\right\} = \begin{cases} 1 & p_2 < 1 \\ 0 & p_2 = 1. \end{cases}$$

As for $U^{VaR_{0.1}^{++}}\left(\hat{F}_t^{\pi^p}\right)$, when $p_2 = 1$ then for all $t \ge 1$ we have $U^{VaR_{0.1}^{++}}\left(\hat{F}_t^{\pi^p}\right) = 10$. If $p_2 < 1$ then the fact that $F^{(1)}$ is strictly increasing causes $U^{VaR_{0.1}^{++}}\left(\hat{F}_t^{\pi^p}\right)$ to converge to $U^{VaR_{0.1}}\left(\hat{F}_t^{\pi^p}\right)$. We conclude that,

$$\mathbf{E}U_{\pi^p}^{bad2} = \begin{cases} \mathbf{E}U_{\pi^p}^{VaR_{0.1}} + 5 & p_2 < 1 \\ 10 & p_2 = 1. \end{cases} \tag{37}$$

Calculating $\mathbf{E}U_{\pi^p}^{VaR_{0.1}}$ and substituting it into (37) yields the result.

These examples show that if an EDPM is either not lower or upper semi-continuous then it might not have a simple oracle policy. However, if it is both lower and upper semi-continuous then it is continuous and thus Theorem 1 typically holds.