A General Approach to Multi-Armed Bandits Under Risk Criteria

Asaf Cassel SASAFCA@CAMPUS.TECHNION.AC.IL
Faculty of Electrical Engineering, Technion, Israel Institute of Technology.

Shie Mannor SHIE@EE.TECHNION.AC.IL
Faculty of Electrical Engineering, Technion, Israel Institute of Technology.

Assaf Zeevi ASSAF@GSB.COLUMBIA.EDU
Graduate School of Business, Columbia University

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

Different risk-related criteria have received recent interest in learning problems, where typically each case is treated in a customized manner. In this paper we provide a more systematic approach to analyzing such risk criteria within a stochastic multi-armed bandit (MAB) formulation. We identify a set of general conditions that yield a simple characterization of the oracle rule (which serves as the regret benchmark), and facilitate the design of upper confidence bound (UCB) learning policies. The conditions are derived from problem primitives, primarily focusing on the relation between the arm reward distributions and the (risk criteria) performance metric. Among other things, the work highlights some (possibly non-intuitive) subtleties that differentiate various criteria in conjunction with statistical properties of the arms. Our main findings are illustrated on several widely used objectives such as conditional value-at-risk, mean-variance, Sharpe-ratio, and more.

Keywords: Multi-Armed Bandit, risk, planning, reinforcement learning, Upper Confidence Bound

1. Introduction

Background and motivation. 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 $F^{(i)}$, $i = 1, \dots, K$ and is independent of t. (To ease notation, 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 $(X_1, \dots, X_t)$. In our setting, performance criteria are defined by a function $\tilde{U}$ that maps the reward vector to a real-valued number. As $\tilde{U}(X_1, \dots, X_t)$ is a random quantity, we consider the accepted notion of expected performance, i.e., $\mathbb{E}\tilde{U}(X_1, \dots, X_t)$. 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 represents the usual notion of regret in the learning problem.

The most widely used 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

[1] Extended abstract. Full version appears as [arXiv reference, 1806.01380]

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 Robbins (1952) and were extensively studied subsequent to that. In particular, the seminal work of Lai and Robbins (1985) establishes that the regret in this problem cannot be made smaller than $\mathcal{O}(\log T)$ and there exist learning algorithms that achieve this regret by maximizing a confidence bound modification of the empirical mean (since then, this class of policies has been come to known as UCB, or upper confidence bound policies); some strands of literature that have emerged from this include Auer et al. (2002) (non-asymptotic analysis of UCB-policies), Maillard et al. (2011) (empirical confidence bounds or KL-UCB), Agrawal and Goyal (2012) (Thompson sampling based algorithms), and various works which consider an adversarial formulation (see, e.g., Auer et al. (1995)).

In this paper we are interested in studying the above problem for more general *path dependent* criteria that are of interest beyond the 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, such as viewed through the lens of the observations collected from the arms, 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. Artzner et al. (1999) 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 the expected value. A typical example where such criteria may be of interest is that of medical trials. 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 and more precise discussion below), the analysis tends to be driven by very case-specific properties of the criterion in question. Unlike the standard mean criterion, various subtleties may arise. To see this, consider the $CVaR_\alpha$ example, which we will reference repeatedly to communicate salient features of our analysis. In terms of $\tilde{U}$, it is given by $\tilde{U}^{CVaR_\alpha}(X_1, \dots, X_t) = \frac{1}{\lceil t\alpha \rceil}\sum_{s=1}^{\lceil t\alpha \rceil} X_s^*$, where X_s^* is the s^{th} order statistic of $(X_1, \dots, X_t)$. Now, for horizon $t = 2$ and $\alpha < 0.5$, an oracle will at first select the arm that maximizes the mean value, just as it would under the traditional mean-criterion. But in step 2 it would seek the arm that maximizes the expected value of the minimum of the first two observations, namely, $\mathbb{E}\min\{X_1, X_2\}$. It is easy to see that this results in a rule that need not select the same arm throughout the horizon of the problem. This presents a further obstacle in characterizing a learning policy that seeks to minimize regret by mimicking the oracle rule. However, as our analysis will flesh out, the oracle policy can be approximated asymptotically by a *simple policy*, that is, one that does select a single arm throughout the horizon. This simplification can be leveraged to address the *learning problem* which becomes much more tractable. It is therefore of interest to understand 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 of the type outlined above. We identify the aforementioned examples, as well as others, as part of a wider class that we term *Empirical Distribution Performance Measures* (EDPM). In particular, let $\hat{F}$ be the *empirical distribution* of the vector $(X_1, \ldots, X_t)$, 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., $U(\hat{F}) = \tilde{U}(X_1, \ldots, X_t)$. Alternatively, U may also serve to evaluate the distributions of the random variables X_s $(s = 1, \ldots, t)$. These evaluations may be aggregated to form a different type of performance criteria that we term proxy regret and consider as an intermediate learning goal. The construct U plays a central role in the framework we develop, and while it may seem somewhat vague at this stage, it will be illustrated shortly by revisiting the $CVaR_\alpha$ example.

Our main results provide easy to verify explicit conditions which characterize the asymptotic behavior of the oracle rule, and culminate in a UCB-type learning algorithm with $\mathcal{O}(\log T)$ regret. To make matters more concrete, we summarize our results for $CVaR_\alpha$. First, its form as an EDPM is essentially given by $U^{CVaR_\alpha}(F) \approx \frac{1}{\alpha}\int_{-\infty}^{F^{-1}(\alpha)} x dF(x)$ (see (7) for exact definition). Our framework will establish that for arm distributions with integrable lower tails, choosing a single arm (simple policies) is asymptotically optimal. This, together with the above characterization of $CVaR_\alpha$ yield the desired simplification in identifying its oracle rule, and subsequently this is leveraged and incorporated in a UCB-type learning algorithm that emulates the oracle policy. More concretely, if $c_{i,t} \propto \sqrt{\frac{\log t}{\tau_i(T)}}$ is the typical UCB upper confidence bound, then a $CVaR_\alpha$ version of UCB requires $\max\{c_{i,t}, c_{i,t}^2\}$ upper confidence bounds for $i = 1, \ldots, K$ and all t, where the power of 2 is a criterion dependent parameter. The implication for learning is that more exploration is required in the initial problem stages. Assuming sub-Gaussian arm distributions, the algorithm obtains $\mathcal{O}(\sqrt{T})$ regret, and under a further mild assumption yields the familiar $\mathcal{O}(\log T)$ regret which, in the traditional MAB objective, corresponds to the case where the means of the arms are "well separated." Our framework allows for this analysis, and the results just mentioned for $CVaR_\alpha$, to be easily derived for any admissible EDPM.

Previous works on bandits that concern path-dependent and risk criteria. To the best of our knowledge, the only works that consider path dependent criteria of the form presented here are Sani et al. (2012), which consider the mean-variance criterion and present the MV-UCB, and MV-DSEE algorithms, and Vakili and Zhao (2016), which complete the regret analysis of said algorithms. Other works consider criteria which are more in line with our intermediate learning goal (proxy regret), and lead to a different notion of regret. Galichet et al. (2013) 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 $CVaR_\alpha$ and average optimal arms coincide. Maillard (2013) presents and analyzes the RA-UCB algorithm which considers the measure of *entropic risk* with a parameter λ. Zimin et al. (2014) 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.

Paper structure. For brevity, all proofs are deferred to the full version of the paper. In Section 2 we formulate the problem setting, oracle, and regret. In Section 3 characterize the asymptotic behavior of the oracle rule. In Sections 4 we provide the flavor main results deferring rigorous statements to the full version of the paper, and in Section 5 we demonstrate them on well-known risk criteria.

2. Problem Formulation

Model and admissible policies. Consider a standard MAB with $\mathbb{K} = \{1, \ldots, K\}$, the set of arms. Arm $i \in \mathbb{K}$ is associated with a sequence $X_t^{(i)}$ $(t \geq 1)$ 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 $X_t^{(i)}$, which is independent of the remaining arms, i.e., the variables $X_t^{(i)}$ (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 $\pi = (\pi_1, \pi_2, \ldots)$ is a random process recursively defined by

$$\pi_t := \pi_t\left(V, \pi_1, \ldots, \pi_{t-1}, X_1^\pi, \ldots, X_{t-1}^\pi\right) \tag{1}$$

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

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

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 $\mathcal{H}_t = \sigma\left(V, \pi_1, X_1^\pi, \ldots, \pi_t, X_t^\pi\right)$, 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. $\frac{1}{t}\sum_{s=1}^{t} X_s^\pi$. 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,

$$\hat{F}_t^\pi(\cdot) := \hat{F}_t(X_1^\pi, \ldots, X_t^\pi; \cdot) \tag{5}$$

$$\hat{F}_t^{(i)}(\cdot) := \hat{F}_t\left(X_1^{(i)}, \ldots, X_t^{(i)}; \cdot\right). \tag{6}$$

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 $\mathbb{E}U\left(\hat{F}_T^\pi\right)$. In section 5 we provide further examples (including the classic empirical mean), but for now, we continue to consider the $CVaR_\alpha$ (Rockafellar and Uryasev (2000)) as our canonical example. This criterion measures the average reward below percentile level $\alpha \in (0,1)$, and for distribution F is given by

$$U^{CVaR_\alpha}(F) = U^{VaR_\alpha}(F) - \frac{1}{\alpha}\int_{-\infty}^{U^{VaR_\alpha}(F)} F(y)dy, \tag{7}$$

where $U^{VaR_\alpha}(F) = \inf_{y\in\mathbb{R}}\left\{y \mid F(y) \geq \alpha\right\}$, is the reward at percentile level $\alpha \in (0,1)$, which is also known as Value at Risk. For further motivation regarding EDPMs and their relation to permutation invariant criteria we refer the reader to the full version of the paper.

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 Banach 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" norm for each U. As a result, our analysis is done with respect to a general norm $\|\cdot\|$ and its matching Banach space $L_{\|\cdot\|}$, which will always be a subspace of L_∞, the space of all bounded functions $f:\mathbb{R}\to\mathbb{R}$, (i.e., $\sup_{x\in\mathbb{R}}|f(x)| < \infty$). We therefore consider EDPMs as mappings $U : L_{\|\cdot\|} \to \mathbb{R}$.

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

$$\pi^*(T) \in \arg\max_{\pi\in\Pi} \mathbb{E}\left[U\left(\hat{F}_T^\pi\right)\right]. \tag{8}$$

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 regret of policy $\pi \in \Pi$ at time T is given by,

$$R_\pi(T) := \mathbb{E}\left[U\left(\hat{F}_T^{\pi^*(T)}\right) - U\left(\hat{F}_T^\pi\right)\right], \tag{9}$$

where 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 $\mathcal{O}(\frac{\log T}{T})$. The goal of this work is to provide a generic analysis of this regret, similar to that of the classic bandit setting. However, unlike the latter, the oracle policy $\pi^*(T)$ here 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 $R_\pi(T)$.

3. The Infinite Horizon Oracle

Infinite horizon oracle. The oracle problem in (8) 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

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

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

Simple policies. In the traditional Multi-Armed Bandit problem, the oracle policy, which selects a single arm throughout the horizon, is clearly simple. In this work, we consider "simple" to mean stationary policies whose actions are mutually independent and independent of the observed rewards. Such policies may differ from the single arm policy in that they allow for a specific type of randomization. The following defines this notion formally.

Definition 1 (Simple policy) *A policy $\pi \in \Pi$ is simple if $(\pi_1, \pi_2, \ldots)$ are $\sigma(V)$ measurable $i.i.d$ random variables. Such policies satisfy*

$$\mathbb{P}(\pi_t = i) = \mathbb{P}(\pi_1 = i), \quad \forall t \geq 1, i \in \mathbb{K}.$$

A deterministic simple policy further satisfies that $\mathbb{P}(\pi_1 = i) = 1$ for some $i \in \mathbb{K}$.

Denote the set of all simple policies by $\Pi^s \subset \Pi$, and the $K-1$ dimensional simplex by,

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

Note that there is a one to one correspondence between Π^s and Δ_{K-1}, we thus associate each $p \in \Delta_{K-1}$ with the simple policy π^p defined by, $\mathbb{P}(\pi_1^p = i) = p_i$, for $i = 1, \ldots, K$.

Stability. It may seem intuitive that EDPMs always admit a simple infinite horizon oracle policy. In the full version of the paper we provide counter examples to this claim, however, these are fabricated edge cases that exploit certain forms of discontinuity that are still allowed by this objective. The following condition is sufficient for capturing the "good behavior" exhibited by typical EDPMs. We denote the convex combinations of the arms' reward distributions by

$$\mathcal{D}^{\Delta} = \left\{ F_p = \sum_{i=1}^{K} p_i F^{(i)} \;\middle|\; p \in \Delta_{K-1} \right\}, \tag{11}$$

and use this in the following definition.

Definition 2 (Stable EDPM) *We say that $U : L_{\|\cdot\|} \to \mathbb{R}$ is a stable EDPM if:*

1. *U is continuous on $\mathcal{D}^{\Delta}$;*
2. $\lim_{t\to\infty} \left\| \hat{F}_t^{(i)} - F^{(i)} \right\| = 0$ *almost surely for all $i \in \mathbb{K}$.*

Note that stability depends not only on U but also on the given distributions $F^{(i)}$. Meaning, a given U could possibly be stable for some distributions and not stable for others. Moreover, the choice of a norm 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 (Van der Vaart (2000)), it satisfies requirement 2 for any given distributions $F^{(i)}$, $i \in \mathbb{K}$. However, in most cases, requirement 1 holds only if the distributions have bounded support.

Remark 1 *Stability has the advantage of being relatively easy to verify. This is due in part to the fact that continuity is preserved by composition. This facilitates the analysis and creation of complicated rewards by representing them as a composition of simpler ones.*

Theorem 1 (Stable EDPM admits a simple oracle policy) *A stable EDPM has a simple infinite horizon oracle policy $\pi^* (\infty)$. Further assuming that U is quasiconvex, a deterministic simple $\pi^* (\infty)$ exists, i.e., choosing a single arm throughout the horizon is asymptotically optimal.*

The main proof idea of Theorem 1 is as follows. We use requirement 2 of stability to show that with probability one and regardless of policy, any subsequence of the empirical distribution has a further subsequence that converges to an element of $\mathcal{D}^{\Delta}$. Applying the continuity of U, we conclude that asymptotic empirical performance is (*almost surely*) equivalent to that of elements in $\mathcal{D}^{\Delta}$. However, similar claims show that such performance can also be achieved by a simple policy. ■

Example ($\mathbf{CVaR}_\alpha$). We summarize how the presented framework applies to $CVaR_\alpha$. First and foremost, we need to define the "correct" norm. We notice that $CVaR_\alpha$, as defined in (7), integrates only the lower tail of the distribution. This leads us to define the following norm

$$\|F\| = \max\left\{ \|F\|_\infty , \left| \int_{-\infty}^{0} x dF \right| \right\}. \tag{12}$$

Verifying requirement 1 (continuity) of stability is a simple technical task. As for requirement 2, using the Glivenko-Cantelli theorem (Van der Vaart (2000)), and the Strong Law of Large Numbers (Simonnet (1996)), it holds when $\left| \int_{-\infty}^{0} x dF^{(i)} \right| < \infty$ $(\forall i \in \mathbb{K})$. Further noticing that U is convex over $\mathcal{D}$, we may use Theorem 1 to conclude that the single arm solution is asymptotically optimal.

4. Proxy Regret and Regret

Preliminaries. Having gained some understanding of the infinite horizon oracle, we consider an intermediate learning goal that uses the infinite horizon performance as a benchmark. We refer to this goal as the *proxy regret* and dedicate this section to the design and analysis of a learning algorithm that seeks to minimize it. Formally, let

$$F_T^{\pi} = \frac{1}{T} \sum_{t=1}^{T} F^{(\pi_t)} = \frac{1}{T} \sum_{i=1}^{K} \tau_i (T) F^{(i)}, \tag{13}$$

be the proxy distribution, where we recall that $F^{(i)}$ is the distribution associated with arm $i \in \mathbb{K}$. The proxy regret is then defined as,

$$\bar{R}_\pi (T) := \mathbb{E} \left[U (F_{p^*}) - U (F_T^{\pi}) \right], \tag{14}$$

where F_p is defined in (11), and $p^* \in \arg\max_{p \in \Delta_{K-1}} U (F_p)$.

Section 3 presented stability as a means of understanding the asymptotic behavior of performance. As we now seek a finite time analysis (of the proxy regret), it stands to reason to employ a stronger notion of stability which quantifies the rate of convergence. For that purpose, denote the set of *empirical distributions* created from sequences of any length $t \geq 1$ by

$$\hat{\mathcal{D}} = \left\{ \hat{F}_t (x_1, \ldots, x_t; \cdot) \;\middle|\; x_1, \ldots, x_t \in \mathbb{R}, \text{for all } t \geq 1 \right\}.$$

Definition 3 (Strongly stable EDPM) *We say that $U : L_{\|\cdot\|} \to \mathbb{R}$ is a strongly stable EDPM if:*

1. *There exist $b > 0, q \geq 1$ such that the restriction of U to $\mathcal{D}^{\Delta} \cup \hat{\mathcal{D}}$ admits $\omega(x) = b(x + x^q)$ as a local modulus of continuity for all $F \in \mathcal{D}^{\Delta}$, i.e.,*

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

2. *There exists a constant $a > 0$ (which depends only on $F^{(i)}$), such that for all $i \in \mathbb{K}$,*

$$\mathbb{P}\left(\left\|\hat{F}_t^{(i)} - F^{(i)}\right\| \geq x\right) \leq 2\exp\left(-atx^2\right), \qquad \forall x > 0, t \geq 1.$$

One can easily verify that a strongly stable EDPM is indeed a stable EDPM. The first requirement quantifies the continuity of U, and the second gives a rate of concentration for $\left\|\hat{F}_t^{(i)} - F^{(i)}\right\|$, thus refining Definition 2.

Proxy 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. Let

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

be the performance gap for arm $i \in \mathbb{K}$. Defining $i^* \in \arg\max U\left(F^{(i)}\right)$, we have that the regret of a linear EDPM is given by, $R_\pi(T) = \frac{1}{T}\sum_{i \neq i^*} \mathbb{E}[\tau_i(T)]\Delta_i$. Departing from the pleasant realm of linearity, we seek a similar decomposition of the proxy regret. Indeed, provided that U is quasiconvex and strongly stable, we have that

$$\bar{R}_\pi(T) \leq \frac{L}{T}\sum_{i \neq i^*} \mathbb{E}[\tau_i(T)]\left\|F^{(i^*)} - F^{(i)}\right\|, \tag{15}$$

where L is a problem dependent parameter. The proof of this argument as well as an explicit expression for L appear in the full version of the paper.

Learning algorithm. We present $U - UCB$, a natural adaptation of $(\alpha, \psi) - UCB$ (see Bubeck and Cesa-Bianchi (2012)) to a strongly stable EDPM. Let,

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

where a, b, q are the parameters of Definition 3. The $U - UCB$ policy is given by,

$$\pi_t^{U-UCB} \in \arg\max_{i \in \mathbb{K}}\left[U\left(\hat{F}_{\tau_i(t-1)}^{(i)}\right) + \phi^{-1}\left(\frac{\alpha \log t}{\tau_i(t-1)}\right)\right], \quad t \geq K+1, \tag{16}$$

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

Theorem 2 ($\mathbf{U-UCB}$ Proxy Regret) *Suppose that $\Delta_i > 0$ for all $i \neq i^*$, and U is a quasiconvex and strongly stable EDPM. Then for $\alpha > 2$ and L taken from (15) we have that*

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

Example ($\mathbf{CVaR}_\alpha$). Unlike stability, strong stability of $CVaR_\alpha$, requires control of both upper and lower tails of the distribution. This leads us to consider the norm

$$\|F\| = \max \left\{ \|F\|_\infty , \left| \int_{-\infty}^{0} x dF \right| , \left| \int_{0}^{\infty} x dF \right| \right\} .$$

Similarly to stability, verifying requirement 1 becomes mostly technical, and results with $q = 2$, and a value of b which depends on an upper bound of the $CVaR_\alpha$ and VaR_α values of the arm distributions. Requirement 2 then follows by Dvoretzky-Kiefer-Wolfowitz (Massart (1990)), and a sub-Gaussian assumption on the arm distributions ($F^{(i)}$, $i \in \mathbb{K}$). We conclude that, for sub-Gaussian arms, $CVaR_\alpha$ incurs $\mathcal{O}(\frac{\log T}{T})$ proxy regret.

Discussion: from proxy regret to regret The proxy regret is a relatively easy metric to analyze but leaves open the question of its relationship to the regret. We refer to the difference between regret and proxy regret as "the gap," and by analyzing it we obtain regret bounds. The framework developed thus far plays a major role in quantifying the gap. More specifically, we find that a strongly stable EDPM has a gap of $\mathcal{O}(\frac{1}{\sqrt{T}})$ (up to logarithmic factors). As seen in Theorem 2 and its application to $CVaR_\alpha$, the proxy regret is of order $\mathcal{O}(\frac{\log T}{T})$, and as such we obtain $\mathcal{O}(\frac{1}{\sqrt{T}})$ as the regret upper bound. Thus it is not clear whether this analysis is tight. For example, consider the classic bandit average reward which is essentially equivalent to a linear EDPM, i.e., when U is linear. In this case the definitions of regret and proxy regret coincide and the gap is clearly zero.

In light of the above, we pursue the needed structural assumptions for obtaining a smaller gap. The salient structural element here is *smoothness*, and it essentially requires that EDPM U have a good (local) linear approximation. When satisfied this additional requirement provides a gap of $\mathcal{O}(\frac{\log T}{T})$, and as an immediate consequence, a similar regret bound. Finally, we note that *smoothness* typically imposes little to no constraints on admissible arm distributions. This implies that strongly stable EDPMs are typically smooth and as such enjoy logarithmic regret.

5. Illustrative Examples

The purpose of this section, first and foremost, is to illustrate how 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. We refer the interested reader to the full version of the paper for the complete details.

Differentiable EDPMs. Assuming that the "correct" norm is chosen, typical EDPMs are differentiable, a fact that essentially implies smoothness. Table 1 introduces some well-known criteria that are compositions of linear functionals, and as such differentiable. Table 2 presents the associated choice of norm and the constraints on arm distributions ($F^{(i)}$) under which our framework yields logarithmic regret. As the emerging pattern in Table 2 suggests, "well behaved" EDPMs only require sub-Gaussian type assumptions to satisfy our framework. Furthermore, it is worth noting

Empirical reward	EDPM Definition	Description
Mean	$U^{ave}(F) = \int_{-\infty}^{\infty} x dF$	The traditional MAB average reward.
Second moment	$U^{E^2}(F) = \int_{-\infty}^{\infty} x^2 dF$	An average of the squared reward.
Below target semi-variance	$U^{-TSV_r}(F) = -\int_{-\infty}^{\infty} (x-r)^2 \mathbb{1}\{x \leq r\} dF$	Measures the negative variation from a threshold $r \in \mathbb{R}$.
Entropic Risk	$U^{ent}(F) = -\frac{1}{\theta} \log\left(\int_{-\infty}^{\infty} \exp(-\theta x) dF\right)$	A risk assessment using an exponential utility function with risk aversion parameter $\theta > 0$.
Negative variance	$U^{-\sigma^2}(F) = -\left[U^{E^2}(F) - [U^{ave}(F)]^2\right]$	Empirical variance of the reward.
Mean-variance (Markowitz)	$U^{MV}(F) = U^{ave}(F) + \rho U^{-\sigma^2}(F)$	A weighted sum (using $\rho \geq 0$) of the empirical mean and variance.
Sharpe ratio	$U^{Sh_r}(F) = \frac{U^{ave}(F)-r}{\sqrt{\varepsilon_\sigma - U^{-\sigma^2}(F)}}$	A ratio between the empirical mean and variance, where r is a minimum average reward, and $\varepsilon_\sigma > 0$ is a regularization factor.
Sortino ratio	$U^{So_r}(F) = \frac{U^{ave}(F)-r}{\sqrt{\varepsilon_\sigma - U^{-TSV_r}(F)}}$	Sharpe ratio with variance replaced by the below target semi-variance measure.

Table 1: Differentiable EDPMs

Empirical reward	The function $g(F)$ in $\Vert F \Vert = \max\left\{\Vert F \Vert_{L_\infty}, \lvert g(F) \rvert\right\}$	Constraints on the random rewards $X^{(i)} \sim F^{(i)}$ for all $i \in \mathbb{K}$	linear/ convex/ quasi -convex
Mean	$U^{ave}(F)$	$X^{(i)}$ are sub-Gaussian	linear
Second moment	$U^{E^2}(F)$	${X^{(i)}}^2$ are sub-Gaussian	linear
Below target semivariance	$U^{TSV_r}(F)$	$\max\left\{0, -X^{(i)}\right\}^2$ are sub-Gaussian	linear
Entropic Risk	$\theta \exp\left(U^{ent}(F)\right)$	$\exp\left(-\theta X^{(i)}\right)$ are sub-Gaussian	convex
Variance	$\max\left\{\lvert U^{ave}(F) \rvert, \left\lvert U^{E^2}(F) \right\rvert\right\}$	${X^{(i)}}^2$ are sub-Gaussian	convex
Mean-variance (Markowitz)	$\max\left\{\lvert U^{ave}(F) \rvert, \left\lvert U^{E^2}(F) \right\rvert\right\}$	${X^{(i)}}^2$ are sub-Gaussian	convex
Sharpe ratio	$\max\left\{\lvert U^{ave}(F) \rvert, \left\lvert U^{E^2}(F) \right\rvert\right\}$	${X^{(i)}}^2$ are sub-Gaussian	quasi-convex
Sortino ratio	$\max\left\{\lvert U^{ave}(F) \rvert, \left\lvert U^{-TSV_r}(F) \right\rvert\right\}$	$X^{(i)}$ and $\max\left\{0, -X^{(i)}\right\}^2$ are sub-Gaussian	quasi-convex

Table 2: EDPM properties

that we did not find any known examples of risk criteria that are not either linear, convex, or quasiconvex.

Non-differentiable EDPMs. We conclude with two examples of non-differentiable criteria. The first, $CVaR_\alpha$, is found to be smooth and strongly stable under appropriate conditions. The second, VaR_α, is strongly stable but appears to be non-smooth. In both cases, our analysis yields particular conditions that are beyond the sub-Gaussian requirements observed thus far. In particular, these relate to the behavior of the arm distributions around the α percentile. Focusing our attention on $CVaR_\alpha$, we observe two types of behavior. First, unlike previous examples, requiring only sub-Gaussian arm distributions may lead to $CVaR_\alpha$ being strongly stable yet non-smooth. In such a case our framework predicts that the gap between regret and proxy regret is of $\mathcal{O}(\frac{1}{\sqrt{T}})$. This is further supported by a simulation result. This observation indicates the necessity of the smoothness condition to achieve a logarithmic gap. Second, further requiring that all $F \in \mathcal{D}^\Delta$ have positive density around their α percentile, we obtain smoothness and thus logarithmic regret.

6. Open Problems and Future Directions

One main question that we leave open is the dependence of the regret on problem parameters such as the number of arms K, and the sub-optimality gaps Δ_i. As our regret analysis passes through the proxy regret, the optimal order of the regret remains open as well. This will likely be resolved by means of a matching lower bound. 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.

Acknowledgments

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

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, pages 39–1, 2012.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331. IEEE, 1995.

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

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

Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *ACML*, pages 245–260, 2013.

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

Odalric-Ambrym Maillard. Robust risk-averse stochastic multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 218–233. Springer, 2013.

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *COLT*, pages 497–514, 2011.

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

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

R Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

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

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

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Alexander Zimin, Rasmus Ibsen-Jensen, and Krishnendu Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833*, 2014.