2019 IEEE 58th Conference on Decision and Control (CDC)
Palais des Congrès et des Expositions Nice Acropolis
Nice, France, December 11-13, 2019

Decision Variance in Risk-Averse Online Learning

Sattar Vakili[1], Alexis Boukouvalas[1], and Qing Zhao[2]

Abstract— **Online learning has traditionally focused on the expected rewards. In this paper, a risk-averse online learning problem under the performance measure of the mean-variance of the rewards is studied. Both the bandit and full information settings are considered. The performance of several existing policies is analyzed, and new fundamental limitations on risk-averse learning is established. In particular, it is shown that although a logarithmic distribution-dependent regret in time T is achievable (similar to the risk-neutral problem), the worst-case (i.e. minimax) regret is lower bounded by $\Omega(T)$ (in contrast to the $\Omega(\sqrt{T})$ lower bound in the risk-neutral problem). This sharp difference from the risk-neutral counterpart is caused by the the variance in the player's decisions, which, while absent in the regret under the expected reward criterion, contributes to excess mean-variance due to the non-linearity of this risk measure. The role of the decision variance in regret performance reflects a risk-averse player's desire for robust decisions and outcomes.**

I. Introduction

A. Risk-Neutral Online Learning

Consider an online decision making problem with a finite set $[K] = \{1, 2, \ldots, K\}$ of actions and a learner who chooses the actions sequentially. Each chosen action $k \in [K]$ at time t results in a random reward $X_{k,t}$ drawn independently over time from an unknown distribution.

Classic formulations of the problem target at the *expected* cumulative reward over a horizon of length T. A commonly adopted performance measure is regret defined as the cumulative reward loss in expectation as compared to the optimal policy with the knowledge of the reward distribution under each action. A sublinear regret order in T implies that not knowing the reward distributions results in diminishing reward loss per play, and the specific regret order gives a finer measure on the efficiency of the learning policies.

We are yet to specify the observations available to the learner for decision-making at each time. Two common feedback models have been considered in the literature: the full-information setting and the bandit setting (see, for example, [1]). In the former, after taking an action $X_{k,t}$ at time t, the random rewards of all K actions are revealed to the learner. This feedback model applies to applications such as stock investment and portfolio management. In the latter, only the reward of the chosen action k is revealed. This model arises naturally from applications such as online ads placement where the payoff of a particular action is only observed after the action is tried out. This coupling between information gathering and reward earning under the bandit setting leads to the exploration-exploitation tradeoff that significantly complicates the problem.

When comparing learning policies in their regret performance, there are two approaches to handling the bias toward specific reward distributions (consider, for example, a policy that always chooses action 1; it works perfectly when this action does lead to the highest expected reward). In the first approach, only policies offering uniformly good performance across all reward distributions (in a certain class) are admissible. These admissible policies are then compared under each possible set of reward distributions. Such a distribution-dependent regret typically depends on certain statistics of the underlying reward distributions such as the Kullback-Leibler (KL) divergence and the gap in the mean values. In the second approach, all policies are admissible. The performance of a policy, however, is taken as the worst among all reward distributions. The regret (referred to as the worst-case or minimax regret) of a policy is thus independent of specific distributions, and policies are compared at different reward distributions, i.e., their specific worst scenarios. It is known that in the full-information setting, the distribution-dependent regret and the minimax regret are lower bounded by $\Omega(\log K)$ [2] and $\Omega(\sqrt{T})$ [3], respectively, with order-optimal policies given in [4], [3]. In the bandit setting, the distribution-dependent regret and the minimax regret are lower bounded by $\Omega(K \log T)$ [5] and $\Omega(\sqrt{KT})$ [6], [7], respectively, with order-optimal policies given in, for example, [7], [8], [9].

B. Risk-Averse Online Learning and Main Results

In this paper, we consider risk-averse online learning. We adopt Markowitz's mean-variance measure, a common risk measure especially for modern portfolio selection [10]. The mean-variance of a random variable X is defined as

$$\text{MV}(X) = \sigma^2(X) - \lambda\mu(X), \tag{1}$$

a linear combination of its mean $\mu(X)$ and variance $\sigma^2(X)$ [11]. The parameter λ is the risk-tolerance factor. It can be interpreted as the inverse Lagrangian multiplier in the constrained optimization of maximizing the expected return $\mu(X)$ subject to a given variance level.

Let $\{\pi_t\}_{t=1}^T$ denote the sequence of actions chosen by a policy π and $X_{\pi_t,t}$ the reward obtained at time t under action π_t. The objective is to minimize the cumulative risk given

[1]Prowler.io, Cambridge, UK, {sattar,alexis}@prowler.io.
[2]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA, qz16@cornell.edu. The work of Qing Zhao was supported by the National Science Foundation under Grant CCF-1815559 and the European Union Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 754412.

by the total mean-variance:

$$\text{MV}_\pi(T) = \sum_{t=1}^{T} \text{MV}(X_{\pi_t,t}).$$

The above cumulative mean-variance measure is an extension of the risk measure of a random variable X to a risk measure of a random process $\{X_{\pi_t,t}\}_{t=1}^T$. In particular, the risk constraint on the variance is imposed locally for each time t. This is particularly relevant to applications such as clinical trial where the risk in each action (i.e. for each patient) needs to be controlled.

Similar to the risk-neutral online learning, regret is defined as the excess in cumulative mean-variance in comparison to the optimal policy π^* under known reward distributions:

$$R_\pi(T) = \text{MV}_\pi(T) - \text{MV}_{\pi*}(T).$$

The regret definition in risk-averse online learning is similar to the one in risk-neutral online learning except that the measure of expected value is replaced with the measure of mean-variance. In the risk-neutral setting, due to the linearity of the expectation operator (and by Wald first identity), regret can be expressed as a weighted sum of the expected number of times suboptimal actions are chosen where the weights are the suboptimality gap of the corresponding action. In the risk-averse setting, however, due to the non-linearity of the performance measure, regret is no longer merely determined by the mean-variance of the rewards of the selected actions, but importantly also, as shown in Sec. III, by the variance in the decisions; hence, the title of the paper. Under the mean-variance measure, in addition to choosing the suboptimal actions, the uncertainty in the actions with different outcomes is penalized, which is motivated by learner's interest in robust decisions and outcomes.

In Sec. III, we establish fundamental limits on the performance of policies under the mean-variance measure. Specifically, we show that the impact of decision variance on the distribution-dependent regret is absorbed by the leading constants of the regret. In other words, the same $\Omega(K \log T)$ and $\Omega(\log K)$ lower bounds on distribution-dependent regret holds under the mean-variance risk measure for bandit and full information cases, respectively. In contrast and rather surprisingly, the variance in the decisions makes an $\Omega(T)$ worst-case regret inevitable under both bandit and full-information feedback models, which is striking in comparison to the sublinear regret order of $\Omega(\sqrt{T})$ in the corresponding risk-neutral problems.

We also analyze the performance of several policies under the risk-averse measure. In the bandit setting, we consider Mean-Variance Lower Confidence Bound (MV-LCB), a modification of the classic UCB introduced in [8] for risk-neutral bandits, and Confidence Bounds based Action Elimination (CB-AE), a more structured policy based on an action elimination method introduced in [12] for risk-neutral bandits. CB-AE considerably reduces the regret by reducing the variance in the decisions. We show that, while an $\mathcal{O}(K \log T)$ distribution-dependent regret is achievable, both MV-LCB and CB-AE have a linear worst-case regret in time. In parallel, in the full information case, we study a modification of Follow the Leader policy [4], referred to as MV-FL as well as CB-AE. We show that, while an $\mathcal{O}(\log K)$ distribution-dependent regret is achievable, both MV-FL and CB-AE have a linear worst-case regret in time. The analysis of the policies shows the tightness of the lower bound results.

C. Related Work

In contrast to the long history of extensive studies on risk-neutral online learning dating back to Thompson's work in 1933 [13], risk-averse online learning is receiving research attention only fairly recently. A couple of existing studies have extended the mean-variance measure to the bandit problem. In defining the mean-variance of a random reward sequence under a given policy, two other approaches exist in the literature, which we refer to as the empirical risk constraint and the global risk constraint. Together with the local risk constraint considered in this work, these models target different applications, depending on which type of uncertainty is deemed as risk. In the empirical risk constraint model first introduced in [14], temporal fluctuations over the *empirical* mean of the *realized* reward sequence are deemed undesired (e.g. volatility in financial security). The risk measure is given by the empirical mean and empirical variance of the realized reward sequence. The global risk constraint model concerns with only the variance of the total reward seen at the end of the time horizon (e.g. retirement investment). The risk measure is thus given by the mean-variance of the sum of the rewards.

The first and yet incomplete study of the empirical risk constraint model was given in [14], which established an $\mathcal{O}(\sqrt{T})$ upper bound on distribution-specific and an $\mathcal{O}(T^{2/3})$ upper bound on distribution-independent regrets. The upper bound of $\mathcal{O}(\sqrt{T})$ on the distribution-specific regret offered by MV-UCB is loose, and no result on achievable lower bounds was given in [14]. The result for the empirical risk constraint model was completed in [15] with lower bounds of $\Omega(\log T)$ for distribution-specific regret and $\Omega(T^{2/3})$ for minimax regret, as well as a tight analysis of MV-UCB showing its optimal $\Theta(\log T)$ distribution-specific regret. Incomplete studies of the global risk constraint model have been reported in [16]. But regret lower bounds remain open, without which, the optimality of policies cannot be assessed.

This work gives the first and complete set of results on local risk constraint model: problem-specific and minimax, full-information and bandit feedbacks, lower bounds and order-optimal policies. Local risk constraint is fundamentally different from empirical and global risk constraints. The differences in objective functions lead to different regret expressions, different feasible minimax regret orders ($T^{\frac{2}{3}}$ vs. linear), and different techniques used in analysis.

In [17], the quality of an action was measured by a general function of the mean and the variance of the random variable. Authors in [18] considered an online variance minimization model. The model in [18] is different than ours in that it allows for linear actions that distribute a budget over

actions at each time (i.e. choose a weighted sum of the actions), which differs from the atomic actions in our model. Note such linear actions can reduce variance (e.g. a linear combination of two i.i.d. random variables has a lower variance than both). Also, [18] assumed direct observation of the variance instead of the value of random rewards. These studies are closer to the risk-neutral bandit problems than to the problem studied in this paper in that the variance in the decisions does not effect the regret as it dominantly does in our results.

In [16], [19], bandit problem under the measure of value at risk was studied. In [19], learning policies using the measure of conditional value at risk were developed. However, the performance guarantees were still within the risk-neutral framework (in terms of the loss in the expected total reward) under the assumption that the best action in terms of the mean value is also the best action in terms of the conditional value at risk. Logarithm of moment generating function was considered as a risk measure for bandit problems in [20] and high probability bounds on regret were obtained. We point out that the logarithm of the moment generating function reduces to mean-variance for a random variable with Gaussian distribution. Even under this special case, [20] uses the mean-variance conditioned on the action at each t, thus measures only randomness in the reward itself for a fixed action, but not the randomness in actions which has complex dependencies on past observations. Thus, [20] is close to the risk-neutral case and has similar regret bounds, while this work shows drastically different bounds.

We point out that both bandit and full information problems have been studied under a different, the so-called adversarial setting where the reward process is non-stochastic and designed adversarially. Under a full information setting, [21] considered a linear combination of mean and empirical standard deviation (in contrast to mean-variance) and established a negative result showing the infeasibility of sublinear regret. The adversarial setting is fundamentally different than the stochastic setting in the assumptions and solution methods.

II. Problem Formulation and Preliminaries

Consider a stochastic online learning problem with a discrete set $[K] = \{1, 2, \ldots, K\}$ of actions. At each time t, a learner chooses an action $k \in [K]$ and receives the corresponding reward $X_{k,t}$, drawn from an unknown distribution f_k. The rewards are independent over k, and i.i.d. over t. Let $\mathcal{F} = \{f_k\}_{k=1}^K$ denote the set of distributions. We use $\mathbb{E}_\mathcal{F}$ and $\Pr_\mathcal{F}$ to denote the expectation and probability with respect to $\mathcal{F}$ and drop the subscript $\mathcal{F}$ when it is clear from the context. Let μ_k, σ_k^2 and MV_k denote the mean, variance and mean-variance of the random reward X_k of action k.

An action selection policy π specifies a sequence of mappings $\{\pi_t\}_{t\geq 1}$ from the history of observations to the action to choose at each time t. In the bandit information setting the learner only observes the reward of the selected action at each time, thus, we have $\pi_t : [K]^{t-1} \times \mathbb{R}^{t-1} \to [K]$. In the full information setting, the learner observes the rewards of all actions at each time, thus we have $\pi_t : [K]^{t-1} \times \mathbb{R}^{K\times(t-1)} \to [K]$.

The objective is an action selection policy π that minimizes regret defined with respect to the optimal policy π^* under known reward distributions:

$$R_\pi(T) = \sum_{t=1}^{T} \textsc{mv}(X_{\pi_t,t}) - \sum_{t=1}^{T} \textsc{mv}(X_{\pi_t^*,t}), \tag{2}$$

where π_t denotes the action taken by policy π at time t, and $\textsc{mv}(\cdot)$ denotes the mean-variance of a random variable as defined in (1). We point out that different from the risk-neutral case where the optimal policy π^* under known reward distributions is easily known to be a single-action policy, the corresponding statement cannot be easily made under the mean-variance measure.

a) Concentration Inequalities: Most existing work on risk-averse (e.g. [14], [15]) and risk-neutral ([8], [9]) online learning assume bounded support distribution. We assume the random variable $(X_{k,1} - \mu_k)^2 - \sigma_k^2$, for all k, is sub-Gaussian with parameter b^2, i.e., its moment generating function is bounded by that of a Gaussian distribution with variance b^2:

$$\mathbb{E}\left[\exp\left(u\left((X_{k,1} - \mu_k)^2 - \sigma_k^2\right)\right)\right] \leq \exp(\frac{u^2b^2}{2}).$$

As a result of the Chernoff-Hoeffding bound ([22]), we have the concentration inequalities on the sample mean and the sample mean-variance given in Lemma 1. This class includes all distributions (of action rewards) with bounded support. The extension to light-tailed distributions is fairly standard as similar concentration inequalities exist for light-tailed distributions (e.g. see [9], [23]).

Let $\mathbb{I}[.]$ denote the indicator function that is, for an event $\mathcal{E}$, $\mathbb{I}[\mathcal{E}] = 1$ if and only if $\mathcal{E}$ is true, and $\mathbb{I}[\mathcal{E}] = 0$, otherwise. Let $\tau_{k,t} = \sum_{s=1}^{t} \mathbb{I}[\pi_s = k]$ denote the number of times that action k has been chosen until time t. The sample mean, the sample variance[1] and the sample mean-variance of each action k up to time t are, respectively, denoted by $\bar{\mu}_{k,t}$, $\bar{\sigma}_{k,t}^2$ and $\bar{\mathrm{MV}}_{k,t} = \bar{\sigma}_{k,t}^2 - \lambda\bar{\mu}_{k,t}$. Specifically, under bandit information $\bar{\mu}_{k,t} = \frac{1}{\tau_{k,t}} \sum_{s=1}^{t} \mathbb{I}[\pi_s = k]X_{k,s}$ and $\bar{\sigma}_{k,t}^2 = \frac{1}{\tau_{k,t}} \sum_{s=1}^{t} \mathbb{I}[\pi_s = k](X_{k,s} - \bar{\mu}_{k,t})^2$; and, under full information $\bar{\mu}_{k,t} = \frac{1}{t} \sum_{s=1}^{t} X_{k,s}$ and $\bar{\sigma}_{k,t}^2 = \frac{1}{t} \sum_{s=1}^{t} (X_{k,s} - \bar{\mu}_{k,t})^2$. To keep the notation uncluttered we drop the specification of the policy from $\tau_{k,t}$, $\bar{\mu}_{k,t}$, $\bar{\sigma}_{k,t}^2$ and $\bar{\mathrm{MV}}_{k,t}$.

Lemma 1 ([15]): Let $\bar{\mathrm{MV}}_t$ be the sample mean-variance of a random variable X obtained from t i.i.d. observations. Let $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathbb{E}[(X-\mu)^2]$, and assume that $(X-\mu)^2 - \sigma^2$ has a sub-Gaussian distribution, i.e.,

$$\mathbb{E}[e^{u((X-\mu)^2-\sigma^2)}] \leq e^{\zeta_1 u^2/2}$$

for some constant $\zeta_1 > 0$. As a result $X - \mu$ has a sub-Gaussian distribution, i.e.,

$$\mathbb{E}[e^{u(X-\mu)}] \leq e^{\zeta_0 u^2/2}.$$

[1]The use of the biased estimator for the variance is for the simplicity of the expression. The results presented in this work remain the same with the use of the unbiased estimator with $\tau_{k,t}$ (t) replaced by $\tau_{k,t} - 1$ ($t-1$) in the expression of $\bar{\sigma}_{k,t}^2$ under bandit (full information) setting.

Let $\zeta = \max\{\zeta_0, \zeta_1\}$. We have, for all constants $\alpha \in (0, \frac{1}{2\zeta}]$ and $\delta \in (0, 2+\lambda]$,

$$\begin{cases} \mathbb{P}[\bar{\text{MV}}_t - \text{MV}(X) > \delta] \leq 2\exp(-\frac{\alpha t \delta^2}{(2+\lambda)^2}), \\ \mathbb{P}[\bar{\text{MV}}_t - \text{MV}(X) < -\delta] \leq 2\exp(-\frac{\alpha t \delta^2}{(2+\lambda)^2}). \end{cases}$$

III. Lower Bounds

A. The Decision Variance and the Decomposition of the Regret

In this subsection, we derive a compact analytical expression of the regret of any given policy π. This expression shows a decomposition of regret into two terms. The first term is given by the expected number of times suboptimal actions are chosen. The second term, which is absent in the regret under the expected reward criterion, captures the role of the variance in the actions (due to the mapping from past random observations) in excess mean-variance. This result also shows that the optimal policy π^* under known models is an optimal single action policy, a fact that is not obvious as in the risk-neutral case.

Lemma 2 provides an expression of regret which is used throughout the paper to analyze the performance of the policies. Let $k^* = \texttt{argmin}_k \text{MV}_k$ (with ties broken arbitrarily), $\Gamma_k = \text{MV}_k - \text{MV}_{k^*}$ and $\Delta_k = \mu_k - \mu_{k^*}$.

Lemma 2: The regret of a policy π under the measure of total mean-variance of rewards can be expressed as

$$R_\pi(T) = \sum_{k=1}^{K} \mathbb{E}[\tau_{k,T}]\Gamma_k + \sum_{t=1}^{T} \mathbb{E}\left[\left(\sum_{k\in[K]\setminus k^*} (\mathbb{I}[\pi_t = k] - \Pr[\pi_t = k])\Delta_k\right)^2\right]. \quad (3)$$

Proof. Omitted due to space limit[2].

The regret expression given in Lemma 2 shows that $R_\pi(T) \geq 0$ for any policy π, and $R_{\pi^*}(T) = 0$ for $\pi_t^* = k^*$ (for all t), which proves that the optimal single-action policy is the optimal policy under the risk-averse measure.

B. Distribution-Dependent Regret

The first term in the regret expression given in Lemma 2 captures choosing suboptimal actions similar to the risk-neutral setting. Since the second term is always positive, the similar distribution-dependent lower bounds as in the risk-neutral problem hold. Specifically, under bandit information setting, an $\Omega(K \log T)$ lower bound for distribution-dependent regret can be established following the similar lines as in the proof of the lower bound results for risk-neutral bandit information setting provided in [5], [6]. Under full information setting, an $\Omega(\log K)$ lower bound for distribution-dependent regret can be established following the similar lines as in the proof of the lower bound results for risk-neutral full information setting provided in [2].

These results are order optimal since, assuming constant distribution parameters ($\Gamma_k > 0$, Δ_k), the distribution-dependent regret incurred due to decision variance is in the same order as the regret incurred due to choosing suboptimal actions. The upper bound results presented in Section IV confirm this observation.

Although the two terms in regret show similar distribution-dependent performance, they are different in the dependence to the distribution parameters; specifically Δ_k and Γ_k. This different scaling, in comparison to the risk-neutral setting, results in different worst-case regret performance as shown next.

C. Worst-case Regret

We prove a linear lower bound for risk-averse regret under worst case distribution assignment which is striking in contrast to the sublinear risk-neutral regret. The lower bound is proven under the full information setting. The same lower bound immediately follows under the bandit information setting since the more limited information in the bandit setting cannot improve the performance. In other words, since the bandit information policies are a subset of the full information policies, any lower bound result on the latter also holds for the former.

Our lower bound proof is based on a coupling argument in a problem with 2 actions. Let $\mathcal{F} = (f_1, f_2)$ and $\mathcal{F}' = (f_1, f_2')$ denote two different distribution models. Let $f_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, a normal distribution with mean $\mu_1 = \frac{3}{2}$ and variance $\sigma_1^2 = \frac{3}{16} - 4\Gamma^2$, for some $\Gamma \in (0, \frac{1}{8})$. Also, let $f_2 \sim \mathcal{B}(p)$, a Bernoulli distribution with $p = 1/4 + 2\Gamma$, and $f_2' \sim \mathcal{B}(q)$ a Bernoulli distribution with $q = 1/4 - 2\Gamma$. For any action selection policy π, we prove that, under at least one of the two systems, the number of times the suboptimal action is chosen is high in expectation.

Lemma 3: For any policy π with full information and any parameter $\Gamma > 0$, in the 2-action problem described above with the number of rounds $T \geq 100$,

$$\{\mathbb{E}_{\mathcal{F}}[\tau_{2,T}] \vee \mathbb{E}_{\mathcal{F}'}[\tau_{1,T}]\} \geq \left\{\frac{0.01}{\Gamma^2} \wedge \frac{T}{2e}\right\}^3. \quad (4)$$

Proof. Omitted due to space limit.

Using Lemma 3, we establish a lower bound on the worst case regret performance of any policy π.

Theorem 1: For any action selection policy π with full information, there exists a distribution assignment $\mathcal{F}$ to a 2-action problem where

$$R_\pi(T) \geq \frac{T}{4e}. \quad (5)$$

Proof:

The first and the second terms in the regret expression given in Lemma 2 correspond to the expected value and the variance of choosing suboptimal actions, respectively. We prove that there exists a mapping from any policy π to a new policy whose expected number of choosing suboptimal

[2]The proofs of Lemma 2, Lemma 3, Theorem 2 and Theorem 3 are omitted from this manuscript due to space limit. The detailed proofs are available in a full version of the paper at https://arxiv.org/abs/1807.09089.

[3]The notation $\{a \vee b\}$ ($\{a \wedge b\}$) denotes the maximum (minimum) of two real numbers a and b.

actions gives a lower bound on the total expected variance of π. This interesting observation together with Lemma 3 proves the theorem. A detailed proof is given below.

Let $[T] = \{1, 2, \ldots, T\}$ denote the set of time instances. For each $S \subseteq [T]$ and any policy π in a 2-action problem, we construct a new policy π^S, based on π, that is obtained by altering the decision of policy π on set S. In particular,

$$\begin{cases} \pi_t^S = \pi_t, & \text{if } t \notin S \\ \pi_t^S = 3 - \pi_t, & \text{if } t \in S. \end{cases} \tag{6}$$

In a 2-action problem, let $\Delta = \Delta_k$ where $k \in \{1, 2\}$ and $k \neq k^*$. In the second term in regret expression given in (3), we have

$$\begin{aligned} \mathbb{E}_{\mathcal{F}} \left[\left(\sum_{\substack{k=1 \\ k \neq k^*}}^{K} (\mathbb{I}[\pi_t = k] - \Pr_{\mathcal{F}}[\pi_t = k]) \Delta_k \right)^2 \right] \\ = \mathbb{E}_{\mathcal{F}} \left[\left((\mathbb{I}[\pi_t \neq k^*] - \Pr_{\mathcal{F}}[\pi_t \neq k^*]) \Delta \right)^2 \right] \\ = \Pr_{\mathcal{F}}[\pi_t \neq k^*](1 - \Pr_{\mathcal{F}}[\pi_t \neq k^*]) \Delta^2. \end{aligned}$$

The first term in the regret expression given in (3), is always positive. Thus

$$R_\pi(T) \geq \sum_{t=1}^{T} \Pr_{\mathcal{F}}[\pi_t \neq k^*](1 - \Pr_{\mathcal{F}}[\pi_t \neq k^*]) \Delta^2. \tag{7}$$

For $t \in S$, $\Pr[\pi_t^S \neq k^*] = \Pr[\pi_t \neq k^*]$ because $\pi_t^S = \pi_t$; and for $t \notin S$, $\Pr[\pi_t^S \neq k^*] = 1 - \Pr[\pi_t \neq k^*]$ because $\pi_t^S = 3 - \pi_t$. We thus have, for all $S \subseteq [T]$

$$\begin{aligned} \Pr_{\mathcal{F}}[\pi_t^S \neq k^*](1 - \Pr_{\mathcal{F}}[\pi_t^S \neq k^*]) \Delta^2 = \\ \Pr_{\mathcal{F}}[\pi_t \neq k^*](1 - \Pr_{\mathcal{F}}[\pi_t \neq k^*]) \Delta^2. \end{aligned} \tag{8}$$

By construction of $\{\pi^S\}_{S \subseteq [T]}$, there exists a $S_0 \subseteq [T]$ that $\Pr_{\mathcal{F}}[\pi_t^{S_0} \neq k^*] \leq \frac{1}{2}$ for all $t \in [T]$. For S_0, we have

$$\begin{aligned} \sum_{t=1}^{T} \Pr_{\mathcal{F}}[\pi_t^{S_0} \neq k^*](1 - \Pr_{\mathcal{F}}[\pi_t^{S_0} = 2]) \Delta^2 \\ \geq \frac{1}{2} \sum_{t=1}^{T} \Pr_{\mathcal{F}}[\pi_t^{S_0} \neq k^*] \Delta^2. \end{aligned} \tag{9}$$

From Lemma 3, there exists a distribution $\mathcal{F}$ for a 2-action problem where

$$\sum_{t=1}^{T} \Pr_{\mathcal{F}}[\pi_t^{S_0} \neq k^*] \geq \{ \frac{0.01}{\Gamma^2} \wedge \frac{T}{2e} \}. \tag{10}$$

Thus, combining (7), (8), (9) and (10), there exists a distribution model $\mathcal{F}$ for the 2-action problem where

$$\begin{aligned} R_\pi(T) &\geq \sum_{t=1}^{T} \Pr_{\mathcal{F}}[\pi_t \neq k^*](1 - \Pr_{\mathcal{F}}[\pi_t \neq k^*]) \Delta^2 \\ &= \sum_{t=1}^{T} \Pr_{\mathcal{F}}[\pi_t^{S_0} \neq k^*](1 - \Pr_{\mathcal{F}}[\pi_t^{S_0} \neq k^*]) \Delta^2 \\ &\geq \frac{1}{2} \sum_{t=1}^{T} \Pr_{\mathcal{F}}[\pi_t^{S_0} \neq k^*] \Delta^2 \\ &\geq \{ \frac{0.005}{\Gamma^2} \wedge \frac{T}{4e} \} \Delta^2. \end{aligned}$$

Choosing the worst case $\Gamma = \sqrt{\frac{0.02e}{T}}$, and for $\Delta = 1$, we have

$$R_\pi(T) \geq \frac{T}{4e},$$

which completes the proof. ∎

We point out that considering only 2 actions does not limit the extension of the lower bound result to the problems with more than 2 actions. Specifically the same lower bound with the same proof holds for a problem with $K > 2$ actions where the actions $k = 3, 4, \ldots, K$ are suboptimal in both $\mathcal{F}$ and $\mathcal{F}'$. Our lower bound proof however lacks the dependency on the number of actions. Nevertheless, notice that a linear lower bound on regret shows the impossibility of converging to the performance of the optimal policy regardless of dependency on K.

The linear lower bound on the regret holds irrespective to the value of λ. The reason is that λ appears only in the first term in the regret corresponding to choosing suboptimal actions. The second term in the regret which corresponds to the decision variance (and has a dominant effect on the worst case regret lower bound) is independent of λ.

IV. Risk-Averse Policies

In this section, we introduce and analyze the performance of several risk-averse policies under both bandit and full information settings.

A. The Bandit Setting

Under bandit information setting we analyze the performance of Mean-Variance Lower Confidence Bound (MV-LCB) policy and Confidence Bounds based Action Elimination (CB-AE) policy.

MV-LCB is a modification of the classic UCB policy first introduced in [8] for risk-neutral bandits and then adopted for risk-averse bandits in [14], [15]. At each time t, MV-LCB chooses the action with the smallest lower confidence bound on mean-variance:

$$\pi_t^{\text{MV-LCB}} = \texttt{argmin}_k \bar{\text{MV}}_{k,t} - \sqrt{\frac{c \log t}{\tau_{k,t}}}, \tag{11}$$

where c is a constant that depends on the distribution class parameter α (as specified in Lemma 1).

Algorithm 1 MV-LCB Policy.

1: Initialization: $T \in \mathbb{N}$, $[K]$, $\bar{\mathtt{MV}}_{k,1} = 0$, $\tau_{k,1} = 0$, for all $k \in [K]$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Play $\pi_t^{\text{MV-LCB}} = \mathtt{argmin}_k \bar{\mathtt{MV}}_{k,t} - \sqrt{\frac{c \log t}{\tau_{k,t}}}$
4: Update $\bar{\mathtt{MV}}_{k,t}$ and $\tau_{k,t}$.
5: **end for**

Theorem 2: When there is a positive gap in the mean-variances of the best and the second best actions, for $c \geq \frac{3(2+\lambda)^2}{\alpha}$, the regret of MV-LCB policy satisfies[4]

$$R^{\pi^{\text{MV-LCB}}}(T) \leq \sum_{k \in [K] \setminus k^*} \left(\frac{4c \log T}{\Gamma_k^2} + 5 \wedge T \right) \left(\Gamma_k + \frac{(K-1)\Delta_k^2}{4} \right). \quad (12)$$

Proof: Omitted due to space limit.

Theorem 2 shows a logarithmic upper bound on the distribution-dependent regret of MV-LCB for easy problems where there is a positive gap $\Gamma = \min_k \{\Gamma_k : \Gamma_k > 0\}$ in the mean variances of the best and the second best actions. Notice that when $\Gamma \to 0$ the upper bound grows to be linear in T.

The CB-AE policy is a modification of Improved UCB introduced in [12] which proceeds in steps $n = 0, 1, 2, \ldots$. At each step n, a set of actions $\mathcal{K}_n$, initialized at $\mathcal{K}_0 = [K]$, are chosen, each $u_n = \lceil \frac{C \log T}{\widehat{\Gamma}_n^2} \rceil$ times where $\widehat{\Gamma}_n = \widehat{\Gamma}_0 2^{-n}$ is initialized at $\widehat{\Gamma}_0 > 0$ and $C > 0$ is a constant that depends only on the distribution class parameter α. At each step, a number of actions are potentially removed from $\mathcal{K}_n$ based on upper and lower confidence bounds on their mean-variance, respectively, in the from of $\bar{\mathtt{MV}}_k^{(n)} + \frac{\hat{\Gamma}_n}{4}$ and $\bar{\mathtt{MV}}_j^{(n)} - \frac{\hat{\Gamma}_n}{4}$, where $\bar{\mathtt{MV}}_k^{(n)}$ is the sample mean-variance obtained from the u_n observations at step n. If the lower confidence bound of action k is bigger than the minimum of the upper confidence bounds of all other remaining actions, action k is removed $\mathcal{K}_{n+1} = \mathcal{K}_n \setminus \{k\}$; see lines 6-10 in Algorithm 2.

Let $n_k = \min\{n : \widehat{\Gamma}_n \leq \Gamma_k\}$ and $n_{\max}$ be the number of steps taken in CB-AE. Let $\Delta_{\max} = \max_{k \in [K] \setminus *} |\Delta_k|$.

Theorem 3: The risk-averse regret performance of CB-AE policy, for $C \geq \frac{64}{\alpha}$, satisfies

$$\begin{aligned} & R^{\pi^{\text{CB-AE}}}(T) \\ & \leq \sum_{k \in [K] \setminus k^*} \left(\frac{\frac{4C}{3} \log T}{\Gamma_k^2} + \log_2 \left(\frac{1}{\Gamma_k} \right) + \frac{K \log_2 T + 2}{T^3} \wedge T \right) \Gamma_k \\ & + \frac{1}{2} \log_2 T \Delta_{max}^2 \sum_{k \in [K] \setminus k^*} \left(\left(\frac{C \log T}{\Gamma_k^2} + 1 \right) \mathbb{I}[n_k \leq n_{\max}] \right. \\ & + \left. \left(\frac{\frac{C}{4} \log T}{\Gamma_k^2} + 1 \right) \mathbb{I}[n_k - 1 \leq n_{\max}] \right) \\ & + \left(\frac{K \log_2 T + 2}{T^4} + \frac{K \log_2 T}{T} \right) \left(\frac{(K-1)^2 T \Delta_{\max}^2}{4} \right). \quad (13) \end{aligned}$$

Proof: Omitted due to space limit.

[4] α is the distribution class parameter specified in concentration inequalities in Lemma 1.

Theorem 2 shows a logarithmic upper bound on the distribution-dependent regret of CB-AE. The worst case regret of CB-AE corresponds to the cases where there exists a k with $\Gamma_k = \Theta(\frac{1}{\sqrt{T}})$. Unlike MV-LCB, CB-AE recovers the sublinear regret for the smaller orders of Γ_k. Specifically, with equally good actions in terms of their mean variance, CB-AE has a 0 regret which is not the case with MV-LCB , as it is shown in the simulations section.

Algorithm 2 CB-AE Policy.

1: Initialization: $\widehat{\Gamma}_0 = 1$, $n = 0$, $T \in \mathbb{N}$, $\mathcal{K}_0 = [K]$.
2: **while** time is left **do**
3: $\mathcal{K}_{n+1} = \mathcal{K}_n$
4: $u_n = \lceil \frac{C \log T}{\widehat{\Gamma}_n^2} \rceil$.
5: Choose each action $k \in \mathcal{K}_n$ for u_n times.
6: **for** $k \in \mathcal{K}_n$ **do**
7: **if** $\bar{\mathtt{MV}}_k^{(n)} - \frac{\widehat{\Gamma}_n}{4} > \min_{j \in \mathcal{K}_n} \bar{\mathtt{MV}}_j^{(n)} + \frac{\widehat{\Gamma}_n}{4}$ **then**
8: Remove action k: $\mathcal{K}_{n+1} \leftarrow \mathcal{K}_{n+1} \setminus \{k\}$.
9: **end if**
10: **end for**
11: n=n+1
12: $\widehat{\Gamma}_{n+1} = \frac{\widehat{\Gamma}_n}{2}$
13: **end while**

B. The Full Information Setting

Full information from actions renders the need for bandit exploration obsolete. The simple Follow the Leader (FL) policy is a common policy in the risk-neutral problem. A straightforward modification of FL for risk-averse problem gives us the policy

$$\pi_t^{MV-FL} = \mathtt{argmin} \bar{\mathtt{MV}}_{k,t}. \quad (14)$$

Theorem 4: The risk-averse regret performance of MV-FL satisfies

$$R^{\pi^{\text{MV-FL}}}(T) \leq \left(\frac{4}{\alpha \Gamma^2} (\log K + 1) + 1 \wedge T \right) \left(\Gamma + \frac{(K-1)\Delta_{\max}^2}{4} \right). \quad (15)$$

Parallel to the bandit information setting, a more structured policy based on action elimination is expected to offer a better risk-averse regret. Specifically, the same CB-AE policy can be used in the full information setting with two changes: first, the sample mean-variance is calculated based on full information available at each step, second, leveraging the full information the value of u_n is reduced to $u_n = \lceil \frac{C \log T}{|\mathcal{K}_n| \widehat{\Gamma}_n^2} \rceil$.

V. Simulations

In this section, we provide simulation results on the performance of MV-LCB, CB-AE, and MV-FL. We compare the performance of MV-LCB and CB-AE in Figure 1. As it is expected, CB-AE shows a better regret performance in the simulations in comparison to MV-LCB. The reason is that CB-AE, by fixing the action elimination structure, reduces the variance in the decisions. While both policies show a linear worst case regret performance, MV-LCB has a linear

regret performance for all the settings where there exists a $k \neq k^*$ with $\Gamma_k = \mathcal{O}(\frac{1}{\sqrt{T}})$ and $\Delta_k >> 0$. On the other hand, CB-AE, as it can be seen from the upper bound in Theorem 3, has a linear regret for the particular case of $\Gamma_k = \Theta(\frac{1}{\sqrt{T}})$ and $\Delta_k >> 0$. Specifically, the CB-AE policy recovers the sublinear regret for the smaller values of Γ_k (when $\Gamma_k \to 0$).

Figure 2 shows the comparison of MV-FL and CB-AE under full feedback setting. While for easy models with relatively large Γ, MV-FL works well and has a sublinear regret, with $\Gamma \to 0$ the regret grows to linear with time. CB-AE , on the other hand, recovers the sublinear regret when $\Gamma \to 0$.

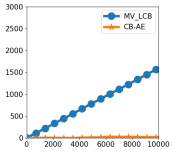
(a) $\Gamma = 0.50$ (b) $\Gamma = 0.20$ (c) $\Gamma = 0.10$

(d) $\Gamma = 0.05$ (e) $\Gamma = 0.01$ (f) $\Gamma = 0.00$

Fig. 1. Comparison of the performance of MV-LCB and CB-AE in terms of their regret over time for different values of Γ.

In this simulation, $K = 4$ actions are Binomially distributed with mean $\mu_* = 1$ and variance $\sigma_*^2 = 1$ for the optimal action. For other actions we choose $\mu_k = 2$ and vary the variance over the set $\{2.5, 2.2, 2.1, 2.05, 2.01, 2.0\}$ simulating different Γ values. The time horizon is varied from $T = 1$ to $T = 10000$ and the regret curves are average performance over 1000 Monte Carlo runs. The parameters for MV-LCB and CB-AE are $c = 1$, $\Gamma_0 = 1$, and $C = 16$.

(a) $\Gamma = 0.50$ (b) $\Gamma = 0.20$ (c) $\Gamma = 0.10$



(d) $\Gamma = 0.05$ (e) $\Gamma = 0.01$ (f) $\Gamma = 0.00$

Fig. 2. Comparison of the performance of MV-FL and CB-AE in terms of their regret over time for different values of Γ.

VI. Conclusion

In this paper, we studied online learning problems under a mean-variance measure. We showed that a dominant term in risk-averse regret comes from the variance in the decisions. We established fundamental limits on learning policies; while a logarithmic distribution-dependent regret is achievable by UCB and FL type policies, similar to the risk-neutral settings, an $\Omega(T)$ worst case regret is inevitable in contrast to the $\Omega(\sqrt{T})$ counterpart lower bound in the risk-neutral setting.

References

[1] V. Dani, T. P. Hayes, and S. M. Kakade, "The price of bandit information for online optimization," in *Proceedings of NIPS*, 2007.

[2] J. Mourtada and S. Gaffas, "Anytime hedge achieves optimal regret in the stochastic regime," *available at arXiv:1809.01382 [stat.ML]*, 2018.

[3] N. Cesa-Bianchi and G. Lugosi, "Prediction, learning, and games," *Cambridge University Press*, 2006.

[4] M. K. Warmuth and W. M. Koolen, "Open problem: Shifting experts on easy data," *JMLR: Workshop and Conference Proceedings*, vol. 35, pp. 1295–1298, 2014.

[5] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[6] S. Bubeck, V. Perchet, and P. Rigollet, "Bounded regret in stochastic multi-armed bandits," *available at http://arxiv.org/abs/1302.1611*, 2013.

[7] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multi-armed bandit problem." *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[8] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *achine Learning*, vol. 47, pp. 235–256, 2002.

[9] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 759–767, Oct 2013.

[10] M. C. Steinbach, "Markowitz revisited: Mean-variance models in financial portfolio analysis," *SIAM Review*, vol. 43, no. 1, pp. 31–85, 2001.

[11] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[12] P. Auer and R. Ortner, "Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem," *Periodica Mathematica Hungarica*, vol. 61, pp. 55–65, 2010.

[13] W. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, pp. 285–294, 1933.

[14] A. Sani, A. Lazaric, and R. Munos, "Risk aversion in multi-armed bandits," in *Neural Information Processing Systems (NIPS)*, 2012.

[15] S. Vakili and Q. Zhao, "Risk-averse multi-armed bandit problems under mean-variance measure," *IEEE Journal of Selected Topics in Signal Processing (JSTSP): Special Issue on Financial Signal Processing and Machine Learning for Electronic Trading*, vol. 10, no. 6, pp. 1093–1111, 2016.

[16] ——, "Mean-variance and value at risk in multi-armed bandit problems," in *Proceedings of 53rd Annual Allerton Conference on Communication, Control, and Computing*, 2015.

[17] A. Zimin, R. Ibsen-Jensen, and K. Chatterjee, "Generalized risk-aversion in stochastic multi-armed bandits," *available at http://arxiv.org/abs/1405.0833*, 2014.

[18] M. K. Warmuth and D. Kuzmin, "Online variance minimization," in *Algorithmic Learning Theory*. COLT, 2006.

[19] N. Galichet, M. Sebag, and O. Teytaud, "Exploration vs exploitation vs safety: Risk-averse multi-armed bandits," in *Proceedings of Asian Conference on Machine Learning*, 2013.

[20] O. Maillard, "Robsut risk-averse stochastic multi-armed bandits," *Algorithmic Learning Theory*, vol. 8139, pp. 218–233, 2013.

[21] E. Even-Dar, M. Kearns, and J. Wortman, "Risk-sensitive online learning," in *Proceedings of 17th international conference on Algorithmic Learning Theory (ALT-06)*, 2006, pp. 199–213.

[22] R. G. Antonioni, Y. Kozachenko, and A. Volodin, "Convergence of series of dependent ϕ-subgaussian random variables," *Mathematical Analysis and Applications*, vol. 338, no. 2, pp. 1188–1203, 2008.

[23] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, "Bandits with heavy tail," *IEEE Transactions on Information Theory*, vol. 59, pp. 7711–7717, 2013.