RISK-AVERSE ONLINE LEARNING UNDER MEAN-VARIANCE MEASURES

Sattar Vakili, Qing Zhao

Department of Electrical and Computer Enginerring
University of California, Davis, CA 95616
{svakili,qzhao}@ucdavis.edu

ABSTRACT

We study risk-averse multi-armed bandit problems under mean-variance measures. We consider two risk mitigation models. In the first model, the variations in the reward values obtained at different times are considered as risk and the objective is to minimize the mean-variance of the observed rewards. In the second model, the quantity of interest is the total reward at the end of the time horizon and the objective is to minimize the mean-variance of the total reward. Under both models, we establish asymptotic as well as finite-time lower bounds on regret and develop online learning algorithms that achieve the lower bounds.

***Index Terms*—** Multi-armed bandit, risk-aversion, mean-variance, regret.

1. INTRODUCTION

1.1. The Classic MAB Problems

Multi-armed bandit (MAB) is a class of online learning and sequential decision-making problems under unknown models. In the classic MAB formulation, there are K independent arms and a single player. At each time, the player chooses one arm to play and obtains a random reward drawn i.i.d. over time from an unknown distribution specific to the chosen arm. The design objective is a sequential arm selection policy that maximizes the total expected reward over a horizon of length T by striking a balance between earning immediate reward and learning the unknown reward models. The performance of an arm selection policy is measured by *regret* defined as the expected total reward loss with respect to the ideal scenario of known reward models (under which the arm with the highest expected reward is always played) [1]. A sublinear regret growth rate with T indicates that the maximum average reward under known models is asymptotically achievable as T approaches infinity, and the slower the regret growth rate, the faster the convergence rate of the average reward.

The MAB problem finds a wide range of applications including clinical trials, target tracking, dynamic spectrum access, internet advertising and web search, and social economical networks (see [2–4] and references therein).

This work is supported by the Army Research Office under Grant W911NF-12-1-0271.

1.2. Risk-averse MAB

In the classic MAB formulation, the performance measure is on the *expected* return of an online learning algorithm. In many applications, however, a player may be more interested in reducing the uncertainty (i.e., risk) in the outcome, rather than achieving the highest ensemble average. In this paper, we formulate and study *risk-averse* MAB problems.

The notions of risk and uncertainty have been widely studied, especially in economics. However, there is no consensus on a mathematical definition of risk and uncertainty, or on the relation between them [5]. Two commonly adopted risk measures are *risk-averse utility function* [6] and *mean-variance* [7]. The latter, introduced by Markowits in 1952, is particularly favored for portfolio selection in finance [8]. Specifically, the mean-variance $\xi(X)$ of a random variable X is given by

$$\xi(X) = \sigma^2(X) - \rho\mu(X), \tag{1}$$

where $\sigma^2(X)$ and $\mu(X)$ are, respectively, the variance and the mean of X, the coefficient $\rho > 0$ is the risk tolerance factor that balances the two objectives of high return and low risk.

In this paper, we study risk-averse MAB under mean-variance measures. We consider two risk mitigation models targeted for different applications. In the first model, the variations in the reward values obtained at different times are considered as risk, and the objective is to minimize the mean variance of the observations (i.e., the obtained rewards over time). One application of this model is clinical trial where, besides obtaining high average return, it is desirable to avoid high variations in the treatment outcomes of different patients [9]. In the second risk mitigation model, the player is only interested in the total reward at the end of the time horizon of length T (rather than each individual reward value obtained during the process). The objective is to minimize the mean-variance of the total reward accrued up to T. This model is of particular interest in economic and financial applications (for example, retirement investment). Under both models, we establish asymptotic as well as finite-time lower bounds on regret and develop online learning algorithms that achieve the lower bounds.

1.3. Related Work

While there are a number of classic and recent studies on the traditional risk-neutral MAB (see [1, 10–12] and references therein), risk-averse MAB has received little attention. There are a

couple of recent studies that address risk-aversion in MAB. In an initial work on this topic, a risk-averse MAB problem using the measure of mean-variance of observations was formulated in [13]. Different from this paper where we consider stochastic risk-averse MAB, [13] adopts the so-called non-stochastic MAB framework first developed in [14] for risk-neutral MAB. The most relevant work to this paper is [9] which also considers the mean-variance of observations as a performance measure for risk-averse MAB within the stochastic framework. It was shown in [9] that a variation of the classic UCB policy developed in [11] for risk-neutral MAB achieves $O(\log^2 T)$ regret performance under the assumption of a positive difference between the mean-variance of the best and the second best arms, and a variation of the DSEE policy developed in [12] for risk-neutral MAB achieves $O(T^{2/3})$ regret performance without the positive difference assumption. While [9] only focuses on algorithm development, this paper also provides tight lower bounds on both the asymptotic and finite-time regret performance which serve as fundamental limits for gauging the optimality of learning algorithms. Furthermore, we provide a finer analysis of the UCB variation (referred to as MV-UCB), showing that it achieves an $O(\log T)$ regret rather than $O(\log^2 T)$ as given in [9]. Together with the regret lower bound developed in this paper, this finer analysis shows that MV-UCB is asymptotically order optimal under the measure of mean-variance of observations. In addition, we provide a parallel study under the model of mean-variance of total reward, a new risk mitigation model that has not been studied in the literature.

There are also a couple of studies on risk-averse MAB under a different risk measure based on the concept of value at risk [15, 16]. Due to the difference in risk measures, these studies and the work reported in this paper employ different techniques, and the corresponding results cannot be directly compared.

2. PRELIMINARIES

Consider a K-armed bandit and a single player. At each time t, the player chooses one arm to play. Playing arm i yields i.i.d. random reward $X_i(t)$ drawn from an unknown distribution f_i. Let $\mathcal{F} = (f_1, \cdots, f_K)$ denote the set of the unknown distributions. An arm selection policy π specifies a function at each time t that maps from the player's observations and decision history to the arm to play at time t.

2.1. Notations and Assumptions

Throughout the paper, $*$ is used to indicate the optimal arm that has the smallest mean-variance. For example, ξ_* and μ_* denote the mean-variance and mean value of the optimal arm, respectively. If there are more than one optimal arm, one of them is chosen arbitrarily as $*$. Let $\Gamma_{i,j} = \mu_i - \mu_j$ and $\Delta_i = \xi_i - \xi_*$ denote, respectively, the difference between the mean values of arm i and j, and the difference between the mean-variance of arm i and the optimal arm. The following notations are used for sample mean, sample variance, and sample mean-variance of arm i:

$$\overline{\mu}_i(t) = \frac{1}{\tau_i(t)} \sum_{s=1}^{\tau_i(t)} X_i(s), \tag{2}$$

$$\overline{\sigma^2}_i(t) = \frac{1}{\tau_i(t)} \sum_{s=1}^{\tau_i(t)} (X_i(s) - \overline{\mu}_i(t))^2, \tag{3}$$

$$\overline{\xi}_i(t) = \overline{\sigma^2}_i(t) - \rho \overline{\mu}_i(t), \tag{4}$$

where $\tau_i(t)$ denotes the number of times arm i has been played up to time t.

Throughout the paper it is assumed that for all random variables X_i assigned to the arms, $(X_i - \mu_i)^2$ has a light-tailed distribution. A random variable X is said to have a light-tailed distribution if

$$\mathbb{E}e^{uX} \leq e^{\zeta_0 u^2/2}, \tag{5}$$

for some constant $\zeta_0 > 0$ and $u \in (0, u_0)$ for some $u_0 > 0$.

2.2. Concentration of the Sample Mean-Variance

In the lemma below, we establish a concentration result on the sample mean-variance, which plays an important role in regret analysis. This result is similar to Chernoff-Hoeffding bound on the concentration of sample mean for light-tailed random variables. The proof is based on Chernoff-Hoeffding bound [17] and is omitted due to the space limit.

Lemma 1 *Let $\overline{\xi}(X)$ be the sample mean-variance of a random variable X calculated from s observations. Assume that $(X - \mu)^2$ has a light-tailed distribution. We have, for some constant $a > 0$*

$$\begin{cases} \mathbb{P}[\overline{\xi}(X) - \xi(X) < -\delta] & \leq 2\exp(-\frac{as\delta^2}{(2+\rho)^2}), \\ \mathbb{P}[\overline{\xi}(X) - \xi(X) > \delta] & \leq 2\exp(-\frac{as\delta^2}{(1+\rho)^2}). \end{cases} \tag{6}$$

3. MEAN-VARIANCE OF OBSERVATIONS

Employing a policy π by the player results in a set of reward observations $\{X_{\pi(t)},\ t = 1, ..., T\}$. The objective in the first risk mitigation model is to minimize the total mean-variance of the observations defined as

$$\overline{\xi}_\pi(t) = \sum_{t=1}^{T} [X_{\pi(t)} + (X_{\pi(t)} - \frac{1}{T} \sum_{t=1}^{T} X_{\pi(t)})^2]. \tag{7}$$

Regret under this model is given by

$$R_\pi^{(1)}(T) = \mathbb{E}[\overline{\xi}_\pi(t)] - T\xi_*. \tag{8}$$

Regret is the difference between the expected total mean-variance of the observations under policy π and that under known models (where the player always plays the optimal arm). Lemma 2 below gives an expression of $R_\pi^{(1)}(T)$ in terms of τ_i and Δ_i. This lemma is used in the following subsections for driving lower bounds on regret and analyzing the performance of the learning algorithms.

Lemma 2 *The regret of a policy π under the measure of mean-variance of observations can be written as*

$$R_\pi^{(1)}(T) = \frac{T-1}{T}\sum_{i=1}^{K}\mathbb{E}[\tau_i(T)]\Delta_i + \frac{1}{T^2}\mathbb{E}[\sum_{i=1}^{K}\tau_i(T)(\sum_{j=1}^{K}\tau_j(T)\Gamma_{i,j})^2]. \quad (9)$$

Proof is omitted due to space limit.

3.1. Lower Bounds on Regret

In the MAB problem two types of regret, referred to as the asymptotic regret and the finite-time regret are distinguished, based on the order of taking the worst-case over all arm distributions and letting T grow to infinity. The asymptotic regret is obtained by letting T grow to infinity first and then consider the worst-case among all possible arm distributions. The finite time regret, however, assumes that the learning algorithm faces a sequence of independent finite-horizon problems with increasing horizon length T. For each finite-horizon problem with length T, the performance of the algorithm is measured against the worst-case arm distribution specific to the horizon length T. The following theorem gives lower bounds on the asymptotic and finite-time regret under the measure of mean-variance of observations. Proof is omitted due to space limit.

Theorem 1 *Consider a risk-averse MAB problem formulation under the measure of mean-variance of observations with $\Delta > 0$. The asymptotic regret of any policy π satisfies, for some constants $c > 0$ and $T_0 \in \mathbb{N}$,*

$$R_\pi^{(1)}(T) \geq \frac{c\log T}{\Delta^2} \quad \textit{for all } T > T_0. \quad (10)$$

The finite-time regret of any policy π satisfies, for some constants $c > 0$ and $T_0 \in \mathbb{N}$,

$$R_\pi^{(1)}(T) \geq cT^{2/3} \quad \textit{for all } T > T_0. \quad (11)$$

3.2. Risk-Averse Learning Algorithms

The first algorithm we consider here is a variation of the classic UCB policy developed in [11] for risk-neutral MAB. This risk-averse learning algorithm, referred to as MV-UCB, assigns an index $I(t)$ to each arm and plays the arm with the smallest index at time t. The index depends on the sample mean-variance calculated from past observations and the number of times that the arm has been played up to time t. Specifically, the index of arm i at time t is given by

$$I_i(t) = \overline{\xi}_i(t) - b\sqrt{\frac{\log t}{\tau_i(t)}}, \quad (12)$$

where b is a policy parameter whose value depends on the risk measure. The regret performance of MV-UCB is given in the theorem below.

Theorem 2 *Set the value of b in the MV-UCB policy such that $b \geq \frac{\sqrt{2}(2+\rho)}{\sqrt{a}}$. The regret offered by the MV-UCB policy under the measure of mean-variance of observations is upper bounded by*

$$R_{MV-UCB}^{(1)}(T) \leq \sum_{i\neq *}(\frac{4b^2\log T}{\Delta_i^2} + 2)(\Delta_i + 2\Gamma_{\max}^2). \quad (13)$$

This theorem gives a finer analysis of the performance of MV-UCB policy comparing to [9] (where an upper bound of $O(\frac{\log^2 T}{\Delta^4})$ was shown). This result together with the lower bound result in Theorem 1 show the asymptotic optimality of the MV-UCB policy under the measure of mean-variance of observations when $\Delta > 0$.

As discussed in [9] the regret performance of MV-UCB algorithm is linear with time when $\Delta = 0$. Thus, the second algorithm as a variation of the DSEE policy developed in [12] for risk-neutral MAB was proposed, achieving $O(T^{2/3})$ regret performance. In the DSEE policy, time is divided into two interleaving sequences: an exploration sequence denoted by $\mathcal{E}(t)$ and an exploitation sequence. In the former, the player plays all arms in a round-robin fashion. In the latter, the player plays the arm with the smallest sample mean. This policy is modified to a risk-averse learning algorithm, referred to as MV-DSEE, by replacing the sample mean with sample mean-variance. To achieve the $O(T^{2/3})$ regret performance it is sufficient to set $|\mathcal{E}(t)| = \lceil t^{2/3}\rceil$. From Theorem 1 it is seen that the MV-DSEE algorithm archives the order optimal finite-time regret performance under the measure of mean-variance of observations. We also mention that the MV-DSEE policy with the parameter setup

$$|\mathcal{E}(t)| = \lceil f(t)\log t\rceil, \quad (14)$$

with $f(t)$ a positive increasing unbounded function, achieves an asymptotic regret performance of $O(f(t)\log t)$ with no assumption on Δ. This means the asymptotic performance can get close to logarithm up to such a function $f(t)$. For example $f(t)$ can be set to $\log t$ or $\sqrt{\log t}$.

4. MEAN-VARIANCE OF TOTAL REWARD

The risk-neutral MAB optimizes the expected total reward obtained by a policy π up to time T. The objective in the second risk mitigation model we study here is to optimize the mean-variance of total reward up to time T. The regret under this model is given by

$$R_\pi^{(2)}(T) = \xi(\sum_{t=1}^{T}X_{\pi(t)}) - T\xi_*, \quad (15)$$

where the first term is the mean-variance of the total reward obtained by policy π and the second term is the mean-variance of the total reward achievable when the arm reward models are known. Parallel to Lemma 2, the following lemma gives an expression of the regret under the measure of mean-variance of total reward in terms of τ_i and Δ_i.

Lemma 3 *The regret of a policy π under the measure of mean-variance of total reward can be written as*

$$R_\pi^{(2)}(T) = \sum_{i=1}^{K}\mathbb{E}\tau_i(T)\Delta_i + \mathbb{E}[(\sum_{i\neq *}(\tau_i - \mathbb{E}\tau_i)\Gamma_{i,*})^2]. \quad (16)$$

4.1. Lower Bounds on Regret

The minimum asymptotic regret growth rate under the measure of mean-variance of total reward can be shown to be logarithmic. This follows readily from the regret lower bound under the classic risk-neutral MAB formulation (see [11, 18]). Details are omitted due to space limit. Next, we consider the finite-time regret under the measure of mean-variance of total reward and show that it grows at least linearly with time. In other words, for any policy π and any time horizon T, there exists an assignment of distributions to the arms that renders sublinear regret orders impossible. The reason behind this result can be intuitively understood by examining the trade-off between learning and risk-aversion. The classic MAB problem is known to address the exploration-exploitation trade-off in learning. Here, we find a higher order trade-off between learning and risk-aversion. The consequence of learning manifests in the dependency of current actions on past observations. This dependency on random observations, however, leads to randomness in the outcome of algorithm, causing uncertainty. In the light of Lemma 3, it can be said upper bounding the number of times a suboptimal arm is played results in an inevitable variance in τ_i. Theorem 3 formally states this result. Proof is omitted due to space limit.

Theorem 3 *Consider a risk-averse MAB problem formulation under the measure of mean-variance of total reward. The finite-time regret of any policy π satisfies, for some constants $c > 0$ and $T_0 \in \mathbb{N}$,*

$$R_{\pi}^{(2)}(T) \geq cT \quad \text{for all } T > T_0. \tag{17}$$

4.2. Risk-averse Learning Algorithms

After some modifications to policy parameters, both the MV-UCB and the MV-DESS polices apply to the risk mitigation model under the measure of mean-variance of total reward. The following theorem states their performance in terms of asymptotic regret order.

Theorem 4 *Set the value of b in the MV-UCB policy such that $b \geq \frac{\sqrt{5}(2+\rho)}{\sqrt{a}}$. The regret offered by the MV-UCB policy under the measure of mean-variance of total reward is upper bounded by*

$$R_{MV-UCB}^{(2)}(T) = O(\frac{\log^2 T}{\Delta^4}). \tag{18}$$

Set the cardinality of exploration sequence $|\mathcal{E}(t)| = \lceil f(t) \log t \rceil$, with $f(t)$ a positive increasing unbounded function. The regret offered by the MV-DSEE policy under the measure of mean-variance of total reward satisfies

$$R_{MV-DSEE}^{(2)}(T) = O(f(t) \log t). \tag{19}$$

For $\Delta = 0$, similar to the case under the measure of mean-variance of observations, MV-UCB policy shows a poor performance. The MV-DSEE policy achieves the $O(f(t) \log t)$ asymptotic regret. This superior performance of MV-DSEE comparing to MV-UCB policy, in this case, is due to the deterministic structure of the MV-DSEE policy. The cardinality of exploration sequence predetermines the number of times each suboptimal arm is played during the exploration sequence. The uncertainty in the exploitation sequence causes only a constant amount of regret.

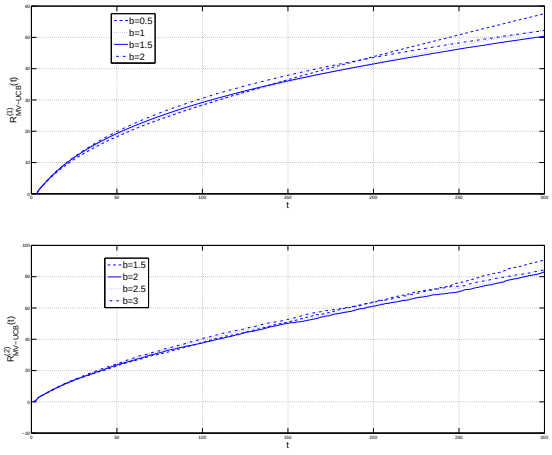

Fig. 1. The effect of parameter b on the performance of MV-UCB policy under different measures.

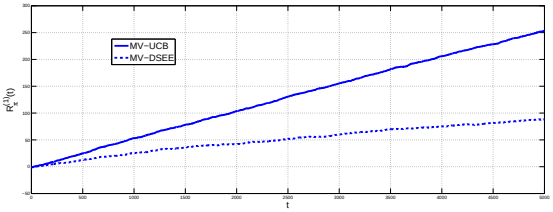

Fig. 2. Comparison between regret performance of MV-UCB and MV-DSEE algorithms when $\Delta = 0$.

5. SIMULATIONS

In this section, we provide numerical experiments on the performance of the algorithms. Fig 1 shows the performance of MV-UCB policy with different values of parameter b. Under the measure of mean-variance of observations the best performance is obtained with $b = 1.5$ while under the measure of mean-variance of total reward the best performance is obtained with $b = 2.5$. This result confirms that the parameter b needs to be tuned differently because of the different objectives under two risk-averse models. The higher value of b under the model of mean-variance of total reward shows a higher rate of exploration under this model. The experiment is run on a 4-armed bandit with normal distributions where $\mu_1 = 0.1$, $\mu_2 = 0.15$, $\mu_3 = 0.2$, $\mu_4 = 0.25$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1.5$, $\sigma_3^2 = 2$, $\sigma_4^2 = 2.5$. Fig. 2 compares the regret performance of the MV-UCB and MV-DSEE algorithms for a case with $\Delta = 0$. While the regret performance of the MV-UCB approaches to linear with time, the MV-DSEE algorithm shows an $O(T^{2/3})$ regret performance. The experiment is run on a 2-armed bandit with normal distributions.

6. REFERENCES

[1] T. Lai, H. Robbins, "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4-22, 1985.

[2] H. Robbins, "Some Aspects of the Sequential Design of Experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527-535, 1952.

[3] T. Santner, A. Tamhane, "Design of Experiments: Ranking and Selection", CRC Press, 1984.

[4] A. Mahajan and D. Teneketzis, "Multi-armed Bandit Problems," *Foundations and Applications of Sensor Management, A. O. Hero III, D. A. Castanon, D. Cochran and K. Kastella, (Editors)*, Springer-Verlag, 2007.

[5] S. Rachev, S. Ortobelli, S. Stoyanov, F. J. Fabozzi, A. Biglova, "Desirable Properties of an Ideal Risk Measure in Portfolio Theorey," *International Journal of Theoretical and Applied Finance (IJTAF)*, vol. 11, no. 01, pages 19-54, 2008.

[6] A. Mas-Collel, M. D. Whinston, J. R. Green, "Microeconomic Theory,"Oxford, England, Oxford University Press, 1995,

[7] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1 pp. 7791, 1952.

[8] M. C. Steinbach, "Markowitz Revisited: Meanvariance Models in Financial Portfolio Analysis," *SIAM Review* vol. 43, no. 1, pp. 3185, 2001.

[9] A. Sani, A Lazaric, R Munos, "Risk Aversion in Multi-armed bandits," *Neural Information Processing Systems (NIPS)*, 2013

[10] R. Agrawal, "Sample Mean Based Index Policies with $O(\log n)$ Regret for the Multi-armed Bandit Problem," *Advances in Applied Probability*, vol. 27, pp. 1054-1078, 1995.

[11] P. Auer, N. Cesa-Bianchi, P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, vol. 47, pp. 235-256, 2002.

[12] S. Vakili, K. Liu, 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, 2013.

[13] E. Even-Dar, M. Kearns, J. Wortman, "Risk-sensitive Online Learning," *17th international conference on Algorithmic Learning Theory (ALT06)*, pp. 199213, 2006.

[14] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, "The Non-stochastic Multi-armed Bandit Problem," *SIAM Journal on Computing*, Vol. 32, pp. 48-77, 2003.

[15] N. Galichet, M. Sebag, O. Teytaud, "Exploration vs Exploitation vs Safety: Risk-averse Multi-Armed Bandits," *Asian Conference on Machine Learning*, 2013.

[16] J. Y. Yu, E. Nikolova, "Sample Complexity of Risk-averse Bandit-arm Selection," *23rd international joint conference on Artificial Intelligence*, pp. 2576-2582, 2013.

[17] R. Agrawal, "The Continuum-Armed Bandit Problem," *SIAM J. Control and Optimization*, vol. 33, no. 6, pp. 1926-1951, November, 1995.

[18] S. Bubeck, V. Perchet, P. Rigollet, "Bounded Regret in Stochastic Multi-armed Bandits," arXiv:1302.1611 [math.ST], 2013.