Risk Control of Best Arm Identification in Multi-Armed Bandits via Successive Rejects

Xiaotian Yu, Irwin King and Michael R. Lyu
Shenzhen Key Laboratory of Rich Media Big Data Analytics and Applications,
Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
Email: {xtyu, king, lyu}@cse.cuhk.edu.hk

***Abstract*—Best arm identification in stochastic Multi-Armed Bandits (MAB) has become an essential variant in the research line of bandits for decision-making problems. In previous work, the best arm usually refers to an arm with the highest expected payoff in a given decision-arm set. However, in many practical scenarios, it would be more important and desirable to incorporate the risk of an arm into the best decision. In this paper, motivated by practical applications with risk via bandits, we investigate the problem of Risk Control of Best Arm Identification (RCBAI) in stochastic MAB. Based on the technique of Successive Rejects (SR), we show that the error resulting from the mean-variance estimation is sub-Gamma by setting mild assumptions on stochastic payoffs of arms. Besides, we develop an algorithm named as `RCMAB.SR`, and derive an upper bound for the probability of error for RCBAI in stochastic MAB. We demonstrate the superiority of the `RCMAB.SR` algorithm in synthetic datasets, and then apply the `RCMAB.SR` algorithm in financial data for yearly investments to show its superiority for practical applications.**

***Keywords*—Multi-armed bandits; Risk control; Successive rejects; Sub-Gamma noise**

I. Introduction

The popular decision-making model of Multi-Armed Bandits (MAB) elegantly characterizes a wide class of problems for sequential learning with partial feedbacks, which was originally proposed and investigated in [1]. Most of MAB algorithms are originally developed to maximize the expected cumulative payoffs during a number of rounds for sequential decisions but the algorithms have limited knowledge on the mechanism of generating a payoff for each round [2], [3]. In recent twenty years, there have been increasing investigations on various variants of the traditional MAB problem [4]–[7].

It is worth pointing out that best arm identification of stochastic MAB has become an important variant due to its inherent characteristic in finding the optimal decision with exploration. Intuitively, its goal is to find the best arm in a given decision-arm set at the end of exploration. In this case, there is no explicit trade-off between exploration and exploitation for each round of sequential learning.

The problem of Risk Control of Best Arm Identification (RCBAI) in stochastic MAB has been rarely investigated, which might be caused by different reasons. One reason is that risk control leads to analysis of high-order statistics, e.g., variance. The high-order statistics may bring the divergence of errors in sequential decisions, and need additional assumptions on distributions of payoffs. Another reason could be that the probability of error for selecting the best arm after exploration is affected by statistics with different orders. Since the problem of RCBAI plays an important role in real applications, it is urgent and meaningful to develop bandit algorithms for conducting best arm identification with consideration of risk control, especially in the case of introducing mild assumptions on the distributions of payoffs in MAB.

In this paper, motivated by the above discussions and two listed examples, we investigate the problem of RCBAI in stochastic MAB, where we adopt the metric of mean-variance. There are three main challenges in solving this problem. The first comes from the analysis of the error between the estimation of mean-variance and the true mean-variance for each arm in MAB, which contains the second order statistics. The second is how to guarantee the independence of samples in sequential decisions, which can affect the final decision of the best arm. The third is how to bound the probability of error for selecting the best arm after T rounds of sequential decisions. Here the best arm refers to the arm with the minimum mean-variance in a given decision-arm set. To solve the challenges, based on the popular technique of Successive Rejects (SR), we prove that the error resulting from the mean-variance estimation is sub-Gamma, where we set mild assumptions on payoffs of arms. Besides, we develop a bandit algorithm to solve the problem of RCBAI in stochastic MAB via SR, which is then simply named as `RCMAB.SR`. We derive an upper bound for the probability of error for RCBAI in stochastic MAB with the proposed algorithm. We demonstrate the superiority of the proposed algorithm in synthetic datasets, and then apply the `RCMAB.SR` algorithm in real financial data to show its superiority for practical applications.

II. Related Work and Preliminary

In this section, we first give a brief review on best arm identification of MAB, as well as risk control of MAB. Then, we present related notations and definitions used in this paper.

A. Related Work

Best arm identification in MAB has become an attracting branch in decision-making problems, where the goal is to identify the best arm after exploration among a given decision-arm set [4], [8]–[10]. It has been pointed out that best arm identification (also known as pure exploration) in MAB has

2374-8486/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDM.2017.153

many potential practical applications, such as communication network and online advertising.

Generally, there are two settings for the line of best arm identification in MAB [6]. One is the setting with a fixed budget, which means that an algorithm would output an best arm after playing a fixed number of rounds. In this case, the theoretical guarantee of the bandit algorithm is to upper bound the probability of error for selecting the best arm [9]. The other setting is to fix a level of confidence to output the best arm, and the theoretical guarantee is to minimize the number of rounds for playing arms [11]. Recently, it has been pointed out that these two setting can be equivalent in the sense of sample complexity [4]. Besides, these two settings can be unified into a model in a recent study [6].

B. Notations

The total number of sequential rounds is T. For each round of $t \in [T]$ with $[T] \triangleq \{1, 2, \cdots, T\}$, an algorithm decides to play an arm among a given decision-arm set of $\mathbf{D}$. At the end of each round t, the algorithm observes a stochastic payoff. Let $K \in \mathbb{N}_+$ be the number of arms in $\mathbf{D}$ (i.e., the size of $\mathbf{D}$) and $\pi_t(y) \in \mathbb{R}$ the stochastic payoff of playing the arm y at round t with $y \in [K]$ and $[K] \triangleq \{1, 2, \cdots, K\}$.

C. Definitions

Definition 1. *A random variable ζ has a Gamma distribution if the probability density function of ζ is*

$$f(\zeta) = \begin{cases} \frac{\beta^\alpha \zeta^{\alpha-1}}{\Gamma(\alpha)} \exp(-\beta\zeta) & \text{if } \zeta \in \mathbb{R}_+, \\ 0 & \text{if } \zeta \notin \mathbb{R}_+, \end{cases} \tag{1}$$

where $\Gamma(y) = \int_0^\infty x^{y-1} \exp(-x) dx$ is a Gamma function, the shape parameter $\alpha > 0$, the scale parameter $\beta > 0$, and $\mathbb{R}_+$ is a set of positive real numbers.

By letting $\alpha = n/2$ and $\beta = 1/2$ in Gamma distributions, we have a Chi-square distribution with n degrees of freedom and $n \in \mathbb{N}_+$. For more details, one can refer to [12].

Definition 2. *(see [13]) A random variable ζ is sub-Gaussian if there exists a constant $R \geq 0$ such that*

$$\mathbf{E}[\exp(\lambda\zeta)] \leq \exp\Big(\frac{\lambda^2 R^2}{2}\Big), \tag{2}$$

where $\lambda \in \mathbb{R}$, $\mathbf{E}[\cdot]$ is the expectation of a random variable and $\exp(\cdot)$ denotes the exponential operation.

Definition 3. *(see [14]) A random variable ζ is sub-Gamma on the right tail if*

$$\mathbf{E}[\exp(\lambda(\zeta - \mathbf{E}[\zeta]))] \leq \exp\Big(\frac{\lambda^2 v}{2(1 - c\lambda)}\Big), \tag{3}$$

where $v > 0$ is a variance factor, $c \in \mathbb{R}$ is a scale parameter and $\lambda \in (0, 1/c)$.

Definition 4. *The measure of mean-variance for a random variable ξ is defined as*

$$\omega(\xi) \triangleq \sigma^2(\xi) - \kappa\mu(\xi), \tag{4}$$

where $\sigma^2(\xi)$ and $\mu(\xi)$ are, respectively, the variance and the mean of ξ, and the coefficient $\kappa \geq 0$ is the risk tolerance factor. It is worth mentioning that κ gives a trade-off between the mean of ξ and the risk of ξ, and κ is a given parameter based on practical needs in RCBAI of stochastic MAB. Besides, given T independent samples as $\{\xi_t\}_{t=1}^T$, we directly define the empirical mean and variance, respectively, as $\hat{\mu}_T \triangleq \sum_{t=1}^T \xi_t/T$ and $\hat{\sigma}_T^2 \triangleq \sum_{t=1}^T (\xi_t - \hat{\mu}_T)^2/(T-1)$. Then, the empirical mean-variance over T samples is $\hat{\omega}_T \triangleq \hat{\sigma}_T^2 - \kappa\hat{\mu}_T$.

At the end of round T, an algorithm $\mathcal{A}$ chooses an action $a_T \in [K]$. The empirical mean-variance of arm a_T is denoted by $\hat{\omega}(a_T)$. Let y^* denote the best arm with the minimum mean-variance shown as Eq. (4). Then, we can infer errors in terms of the difference between the mean-variance of the best arm and its empirical mean-variance. For $y \neq y^*$, we further introduce the following sub-optimality metric between arms y and y^* as

$$\Delta_y \triangleq \omega(y) - \omega(y^*), \tag{5}$$

where $y \in [K]$. Based on Eq. (5), we easily define the minimal sub-optimality as $\Delta_{y^*} \triangleq \min_{y \neq y^*, y \in [K]} \Delta_y$. We introduce the notation $(y) \in [K]$ to denote the y-th best arm, thus

$$\Delta_{y^*} = \Delta_{(1)} \leq \Delta_{(2)} \leq \Delta_{(3)} \leq \cdots \leq \Delta_{(K)}. \tag{6}$$

The sorted sequence of sub-optimality shown in Eq. (6) is helpful to analyze the probability of error for selecting the best arm. Besides, given K arms with a unique optimal arm, the number of sub-optimal arms is $K - 1$. Without loss of generality, the probability of error for selecting an best arm after t rounds can be presented as $\mathbb{P}[a_t \neq y^*]$. Inspired by [9], the probability of error is related to the sub-optimality in Eq. (5). We define the hardness as:

$$\mathbf{H}_1 \triangleq \sum_{y=1}^K \frac{1}{\Delta_y^2}, \ \mathbf{H}_2 \triangleq \max_{y \in [K]} y\Delta_{(y)}^{-2},$$

$$\mathbf{H}_3 \triangleq \sum_{y=1}^K \frac{1}{\Delta_y}, \mathbf{H}_4 \triangleq \max_{y \in [K]} y\Delta_{(y)}^{-1}.$$

Besides, by letting $\overline{\log}(K) = 1/2 + \sum_{i=2}^K 1/i$, we have the result of $\log(K+1) - 1/2 \leq \overline{\log}(K) \leq \log(2K)$. We can generalize the inequality on the hardness of [9] as

$$\mathbf{H}_2 \leq \mathbf{H}_1 \leq \overline{\log}(K)\mathbf{H}_2. \tag{7}$$

$$\mathbf{H}_4 \leq \mathbf{H}_3 \leq \overline{\log}(K)\mathbf{H}_4. \tag{8}$$

III. Assumptions, Problem and Challenges

In this section, we first present the assumptions. Then, we give the problem definition of RCBAI in stochastic MAB.

A. Assumptions

1) We assume that there are K arms for best arm identification in stochastic MAB, and payoffs of arm y (with $y \in [K]$) are independently drawn from a normal distribution as $\mathcal{N}(\mu(y), \sigma^2(y))$, where $\mu(y)$ is the mean of arm y and $\sigma^2(y)$ is the variance of arm y.

Algorithm 1 RCMAB.SR

1: **input** T, K, κ
2: construct a decision-arm set $\mathbf{D}_1 \leftarrow [K]$
3: calculate $\overline{\log}(K) \leftarrow \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$
4: **for** $y \in [K]$ **do**
5: $\quad \Phi(y) \leftarrow \emptyset$ ▷ *construct sets to store time index*
6: **end for**
7: $t \leftarrow 0$; $T_0 \leftarrow 0$
8: **for** $k \in [K-1]$ **do**
9: $\quad T_k \leftarrow \lceil \frac{T-K}{\overline{\log}(K)(K+1-k)} \rceil$ ▷ *take ceiling of the division*
10: $\quad n \leftarrow T_k - T_{k-1}$
11: $\quad$ **for** $y \in \mathbf{D}_k$ **do**
12: $\quad\quad$ **for** $i \in [n]$ **do**
13: $\quad\quad\quad t \leftarrow t + 1$
14: $\quad\quad\quad$ select arm y, and observe a payoff $\pi_t(y)$
15: $\quad\quad\quad \Phi(y) \leftarrow \Phi(y) \cup t$ ▷ *store the index for arm y*
16: $\quad\quad$ **end for**
17: $\quad$ **end for**
18: $\quad$ **for** $y \in \mathbf{D}_k$ **do**
19: $\quad\quad \hat{\omega}_t(y) \leftarrow \hat{\sigma}_t^2(y) - \kappa \hat{\mu}_t(y)$ ▷ *estimate mean-variance*
20: $\quad$ **end for**
21: $\quad a \leftarrow \arg\max_{y \in \mathbf{D}_k} \hat{\omega}_t(y)$ ▷ *break ties arbitrarily*
22: $\quad \mathbf{D}_{k+1} \leftarrow \mathbf{D}_k - \{a\}$ ▷ *successive rejects of arms*
23: **end for**
24: **return** $a_T \leftarrow \mathbf{D}_K$ ▷ *only one element in $\mathbf{D}_K$*

2) We assume that, given an arm $y \in [K]$, the variance of arm y is a constant denoted as $\sigma^2(y)$. Besides, we also assume that the variances among K arms are not all the same, and they are upper bounded by $\bar{\sigma}^2$ with $\bar{\sigma}^2 < +\infty$.
3) We assume that the best arm is unique among K arms, and thus the best arm is denoted by y^* for best arm identification in MAB among K arms.

B. Problem Definition

Without loss of generality, in this paper, we focus on the problem of RCBAI for stochastic MAB in the setting of a fixed budget, which is an important scenario [4], [6]. For an algorithm $\mathcal{A}$, the goal of RCBAI of MAB is to identify the best arm of y^* with the smallest mean-variance shown in (4). Specifically, given a fixed budget of T, we design bandit algorithms to minimize the probability of error, which is explicitly shown as

$$\min \ \mathbb{P}[a_T \neq y^*]. \tag{9}$$

It is very difficult to directly solve the problem of Eq. (9). A potential solution is to upper bound the probability of error in $\mathbb{P}[a_T \neq y^*]$, which is popular in [4], [9]. Compared with the problem of best arm identification in the traditional MAB, the problem of RCBAI will encounter second order statistics.

IV. Algorithm

In order to solve the problem of RCBAI in stochastic MAB (shown as Eq. (9)), we develop a bandit algorithm, which is simply named as RCMAB.SR (i.e., Algorithm 1). The RCMAB.SR algorithm is based on the technique of SR with estimations of mean-variance.

In light of the definition of mean-variance, at time t, we can design the mean-variance estimation as follows:

$$\hat{\omega}_t(y) = \hat{\sigma}_t^2(y) - \kappa \hat{\mu}_t(y), \tag{10}$$

where $\hat{\mu}_t(y)$ is the estimation of the true expected payoff of $\mu(y)$ at time t. We can calculate $\hat{\mu}_t(y)$ as

$$\hat{\mu}_t(y) = \sum_{i \in \Phi(y)}^{s_t(y)} \frac{\pi_i(y)}{s_t(y)}, \tag{11}$$

where $\Phi(y)$ is a set to store time instants of selecting arm y, and $s_t(y)$ is the size of the set $\Phi(y)$ at time t. Besides, we have $\hat{\sigma}_t^2(y)$ as

$$\hat{\sigma}_t^2(y) = \frac{1}{s_t(y) - 1} \sum_{i \in \Phi(y)}^{s_t(y)} (\pi_i(y) - \hat{\mu}_t(y))^2. \tag{12}$$

A. Successive Rejects

In the research line of bandits, the technique of SR has become popular in best arm identification with stochastic MAB [4], [6], [9]. The main idea of SR is to divide $K-1$ phrases of exploration given a fixed budget T among K arms. For each phrase, algorithms with SR eliminate a worst arm in the decision-arm set. When the $(K-1)$-th phrase is finished, the decision-arm set should always has only one arm, which is selected as the best arm. Because the number of pulls for each arm in the decision-arm set is the same in a phrase, then the samples for pulling arms are independent.

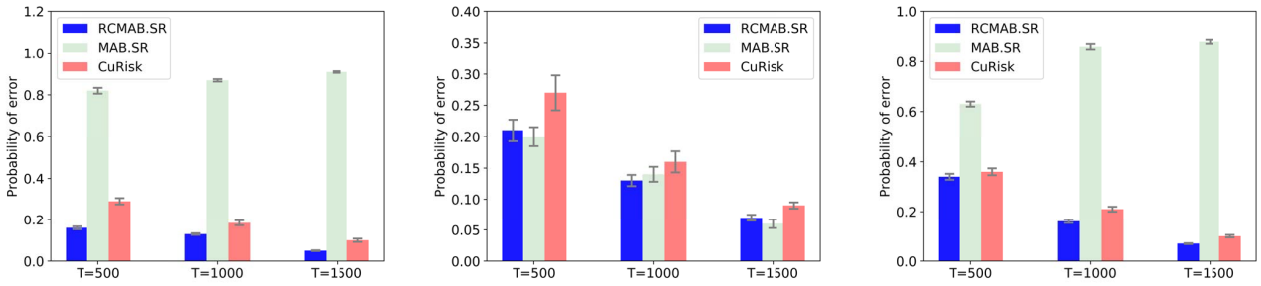
B. Description of Algorithm

We show the proposed RCMAB.SR algorithm to solve the problem of RCBAI in Algorithm 1. Specifically, given a fixed budget T, K arms and a risk tolerance factor κ, the RCMAB.SR algorithm should output the best arm after T rounds. We divide T rounds into $K-1$ phrases, and the phrases are not uniform in the sense of the number of time for pulling arms. We follow the phrase design in [9], which is technical and helpful for theoretical analysis. We can infer the number of pulling arms in RCMAB.SR is $T_{\text{pull}} = \sum_{i=1}^{K-1} T_i + T_{K-1}$, which is bounded as $T - K \leq T_{\text{pull}} \leq T$.

It is worth pointing out that the estimation of mean-variance in Line 19 brings the challenges in the RCMAB.SR algorithm. The technique of SR guarantees the independence of samples, which is helpful for analyzing the estimation errors of mean-variance. At the end of exploration, we always has a unique element in the decision-arm set, which is directly output as the best arm. For the time consumption of RCMAB.SR, we can calculate its time complexity as $\mathcal{O}(TK)$.

V. Theoretical Analyses

Theorem 1. *For stochastic MAB with K arms, each arm follows a normal distribution as $\mathcal{N}(\mu(y), \sigma^2(y))$ and $y \in [K]$. In best arm identification of MAB, we assume payoffs $\{\pi_i(y)\}$ are independently drawn from arm y with $i \in \Phi(y)$ and $\Phi(y)$*



(a) S1 (b) S2 (c) S3

Fig. 1. Probability of error for best arm identification in MAB with $\kappa = 1.0$ and different values of T.

being an index set for choosing arm y. If the sample mean-variance of arm y is designed as $\hat{\omega}_t(y) = \hat{\sigma}_t^2(y) - \kappa\hat{\mu}_t(y)$, where we have

$$\hat{\sigma}_t^2(y) = \frac{1}{s_t(y) - 1} \sum_{i \in \Phi(y)}^{s_t(y)} (\pi_i(y) - \hat{\mu}_t(y))^2,$$

$\hat{\mu}_t(y) = \sum_{i \in \Phi_t(y)}^{s_t(y)} \pi_i(y)/s_t(y)$ *with $s_t(y)$ being the size of $\Phi(y)$ at time t and $\kappa \geq 0$, then the variable of $\rho_t(y) = \hat{\omega}_t(y) - \omega(y)$ is sub-Gamma on the right tail, which means that*

$$\mathbf{E}[\exp(\lambda(\rho_t(y) - \mathbf{E}[\rho_t(y)]))] \leq \exp\Big(\frac{\lambda^2 v}{2(1 - c\lambda)}\Big),$$

where $c = 2\bar{\sigma}^2$, $v = (2\bar{\sigma}^2 + \kappa^2/2)\bar{\sigma}^2$, and $\lambda \in (0, 1/c)$.

Proof. We give a sketch of proof as follows. We first derive moment generating functions of mean and variance separately. Then, due to independence, we combine them together to obtain a sub-Gamma distribution. □

Theorem 2. *For stochastic MAB with K arms, each arm follows a normal distribution as $\mathcal{N}(\mu(y), \sigma^2(y))$, where $\sigma^2(y)$ is upper bounded by $\bar{\sigma}^2$ (with $\bar{\sigma}^2 \geq 1$) and $y \in [K]$. Suppose Assumptions 1-3 are satisfied for best arm identification in stochastic MAB. If Algorithm 1 is run with a fixed budget of T, we have an upper bound of the probability of error as*

$$\mathbb{P}[a_T \neq y^*] \leq \frac{K(K-1)}{2} \exp\Big(-\frac{T - K}{\overline{\log}(K)\mathbf{H}}\Big), \tag{13}$$

where $\mathbf{H} = \max\{\mathbf{H}_4(6\bar{\sigma}^4 + 3\kappa^2\bar{\sigma}^2), \mathbf{H}_2(18\bar{\sigma}^4 + 9\kappa^2\bar{\sigma}^2)\}$ and $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$.

Proof. The proof can be generalized from [9]. □

VI. Experiments

In this section, we conduct a series of experiments via synthetic and real datasets to evaluate the performance of the proposed `RCMAB.SR` algorithm. We compare the `RCMAB.SR` algorithm with two state-of-the-art algorithms in best arm identification of bandits, i.e., `MAB.SR` [9] and `CuRisk` [15].

TABLE I
Statistics of used datasets.

dataset	#arms	$\{\mu(y)\}$	$\{\sigma^2(y)\}$
S1	3	{1.0,1.5,1.3}	{0.2,0.5,0.1}
S2	5	[1.0,1.4] with a uniform gap	{0.1,0.2,0.4, 0.5,0.3}
S3	8	[1.0,1.7] with a uniform gap	$\{\sigma^2(5)$=0.3, $\sigma^2(8)$ =0.5, others 0.1}

A. Settings

We conduct experiments on a personal computer with Intel CPU@3.70GHz and 16GB memory. In order to evaluate the performance of algorithms in synthetic datasets, we calculate the probability of error based on frequency of wrong decision after exploration. Specifically, we run multiple epochs of experiments, with each epoch containing ten independent experiments for best arm identification. For each independent experiment, algorithms output an estimated best arm at T. We label 1 for an experiment if the output arm is the optimal arm in hindsight. Otherwise we label 0. For the ten experiments of an epoch, we evaluate the probability of error in terms of frequency of zero in labels. Clearly, we have an estimated probability of error in an epoch. By running multiple epochs, we obtain an average of probability of error and its standard error. In experiments, we set the number of epochs as 10.

B. Synthetic Datasets and Results

For verifications, we adopt three synthetic datasets (named as S1-S3) in the experiments, of which statistics are shown in Table I. It is worth mentioning that, with $\kappa = 1.0$, the best arm with true mean-variance in S2 is equivalent to that with mean. For comparisons, we set the parameters in `MAB.SR` the same as those in `RCMAB.SR`.

Via experimental results in Fig. 1, we find superior performance of the proposed `RCMAB.SR` algorithm in terms of the probability of error for selecting the best arm with a fixed budget of T. We also find that different values of κ affect the performance of `RCMAB.SR`, as shown in Fig. 2.

From Fig. 1, we find that `RCMAB.SR` beats `CuRisk` for best arm identification, because the probability of error for

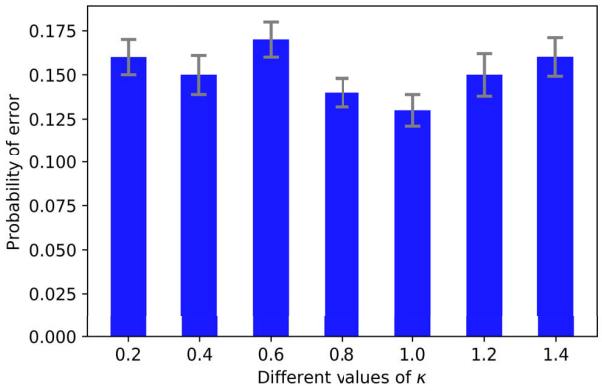


Fig. 2. Probability of error via RCMAB.SR in S2 with $T = 1000$.

RCMAB.SR is the lower. Besides, we find that RCMAB.SR is comparable to MAB.SR in S2. Note that the best arm with mean-variance in S2 with $\kappa = 1.0$ is equivalent to that with mean. In this case, the RCMAB.SR incurs higher errors than MAB.SR. It means that the first order statistics is sufficient for best arm identification with $\kappa = 1.0$ in S2. Finally, by increasing T, we observe that the performance of RCMAB.SR will increase. This is consistent with theoretical analysis of the probability of error.

In Fig. 2, we set different values of κ in experiments. From the figure, we know that different values of κ affect the performance RCMAB.SR, and the effect is nonlinear.

C. Financial Data and Results

The real data for experiments are historical returns on stocks, bonds and bills of United States from 1928 to 2016[1]. The dataset contains 89 samples of annual returns on SP500, 3-month Treasury Bill and 10-year Treasury Bond.

We design the experiment as follows, of which the scheme is shown in Fig. 3. For yearly investments of the above dataset, we should output the best arm (i.e., the red dot with dash line in Fig. 3) for investments in each year. For example, at the beginning of 2015, we first determine which choice is the best among SP500, 3-month Treasury Bill and 10-year Treasury Bond, and then invest all the available money on that choice. After a year (i.e., at the beginning of 2016), we observe the realized return of the choice in 2015, and sequentially determine the best choice for investments in 2016. We define the cumulative returns as

$$C_{ret}(N) = \prod_{i=1}^{N}(1 + r_i), \tag{14}$$

where r_i is the realized return for the i-th investment period, and N is the total periods in investments. Clearly, an algorithm performs better if $C_{ret}(N)$ is higher.

We apply best arm identification of bandits in each investment period. For best arm identification, the sliding window

[1]http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/

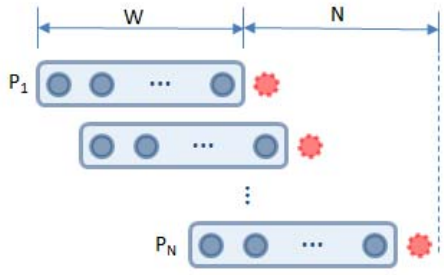


Fig. 3. Yearly investments over N rounds with a sliding window W.

contains historical W samples for each arm, and the fixed budget in RCMAB.SR is set as $T = W$. We show the empirical performance of $C_{ret}(N)$ based on three bandit algorithms (i.e., RCMAB.SR, MAB.SR and CuRisk) in Figs. 4 and 5. We set different κ as $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$, and different observation window W as $\{10, 40\}$.

In Fig. 4, we find that RCMAB.SR always outperforms CuRisk in terms of cumulative returns. This superiority comes from the key idea of the technique of SR in best arm identification. Besides, it is interesting to find that RCMAB.SR outperforms MAB.SR in some region of κ, which means that, based on Eq. (4), one should not overweight or neglect the mean of payoff in investments. When the mean of payoff is overweighted, the dominant term is the mean of payoff. This can be verified by the performance in Fig. 4(f).

We demonstrate similar results in Fig. 5, where $\kappa = 0.5, 1.5, 3.0$. In Fig. 5(a), it is surprising to find that the yearly investments over 49 years via RCMAB.SR have a non-negative return for each year. This reveals that the sufficient exploration (i.e., large T in RCMAB.SR) brings better identification of the best arm given a decision-arm set.

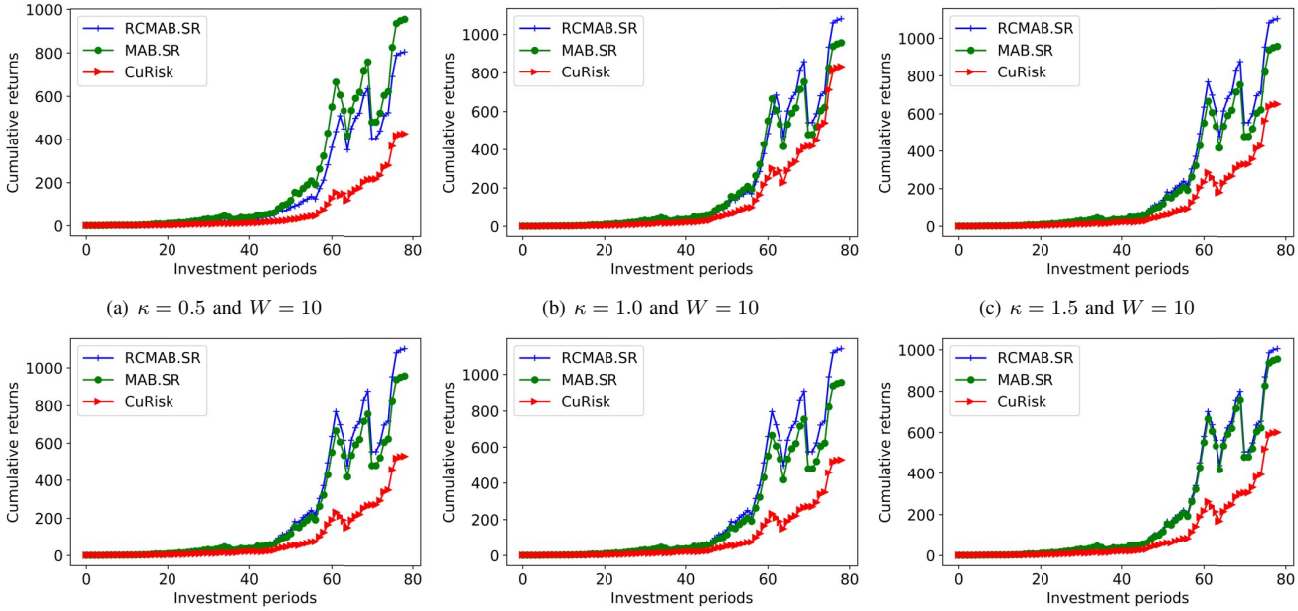
Overall, by comparing with MAB.SR and CuRisk, we show the superiority of RCMAB.SR in yearly investments with the measure of cumulative returns.

VII. Conclusion

In this paper, motivated by risk control in best arm identification for real applications, we studied the problem of RCBAI in stochastic MAB. We proved the error resulting from the estimation of mean-variance for best arm identification is sub-Gamma by setting mild assumptions on stochastic payoffs of arms. We developed an efficient bandit algorithm named RCMAB.SR to solve the problem, and derived an upper bound for the probability of error. By comparing with two baselines, we showed superior performance of the RCMAB.SR algorithm with synthetic and real datasets. We showed that RCMAB.SR helped to bring stable cumulative returns in yearly investments.

Acknowledgments

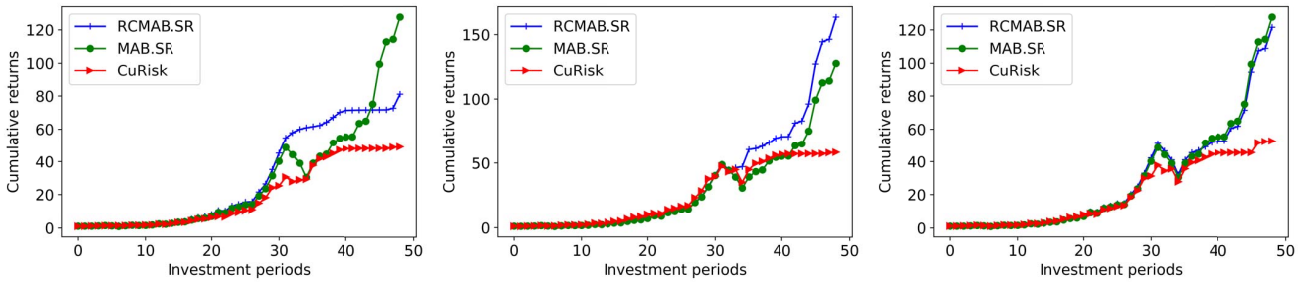
The work described in this paper was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK14234416 and No. CUHK14208815 of the General Research Fund), and 2015



(a) $\kappa = 0.5$ and $W = 10$ (b) $\kappa = 1.0$ and $W = 10$ (c) $\kappa = 1.5$ and $W = 10$

(d) $\kappa = 2.0$ and $W = 10$ (e) $\kappa = 2.5$ and $W = 10$ (f) $\kappa = 3.0$ and $W = 10$

Fig. 4. Cumulative returns in yearly investments on SP500, 3-month Treasury Bill and 10-year Treasury Bond with sliding window $W = 10$. The investment is one-year forward from 1937 to 2016.



(a) $\kappa = 0.5$ and $W = 40$ (b) $\kappa = 1.5$ and $W = 40$ (c) $\kappa = 3.0$ and $W = 40$

Fig. 5. Cumulative returns in yearly investments on SP500, 3-month Treasury Bill and 10-year Treasury Bond with sliding window $W = 40$. The investment is one-year forward from 1967 to 2016.

Microsoft Research Asia Collaborative Research Program (Project No. FY16-RES-THEME-005).

REFERENCES

[1] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[2] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[3] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, vol. 3, pp. 397–422, 2002.

[4] S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen, "Combinatorial pure exploration of multi-armed bandits," in *NIPS*, 2014, pp. 379–387.

[5] X. Yu, H. Yang, I. King, and M. R. Lyu, "Online non-negative dictionary learning via moment information for sparse poisson coding," in *IJCNN*. IEEE, 2016, pp. 5094–5101.

[6] V. Gabillon, M. Ghavamzadeh, and A. Lazaric, "Best arm identification: A unified approach to fixed budget and fixed confidence," in *NIPS*, 2012, pp. 3212–3220.

[7] X. Yu, M. R. Lyu, and I. King, "Cbrap: Contextual bandits with random projection." in *AAAI*, 2017, pp. 2859–2866.

[8] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *ALT*, 2009, pp. 23–37.

[9] J.-Y. Audibert and S. Bubeck, "Best arm identification in multi-armed bandits," in *COLT*, 2010, pp. 13–p.

[10] E. Kaufmann, O. Cappé, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1–42, 2016.

[11] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *CISS*, 2014, pp. 1–6.

[12] H. O. Lancaster and E. Seneta, *Chi-Square Distribution*. Wiley Online Library, 1969.

[13] V. V. Buldygin and Y. V. Kozachenko, "Sub-gaussian random variables," *Ukrainian Mathematical Journal*, vol. 32, no. 6, pp. 483–489, 1980.

[14] S. Boucheron, M. Thomas *et al.*, "Concentration inequalities for order statistics," *Electronic Communications in Probability*, vol. 17, no. 51, pp. 1–12, 2012.

[15] J. Y. Yu and E. Nikolova, "Sample complexity of risk-averse bandit-arm selection." in *IJCAI*, 2013.