# Risk-Averse Allocation Indices for Multiarmed Bandit Problem

Milad Malekipirbazari and Özlem Çavuş

***Abstract*—In classical multiarmed bandit problem, the aim is to find a policy maximizing the expected total reward, implicitly assuming that the decision-maker is risk-neutral. On the other hand, the decision-makers are risk-averse in some real-life applications. In this article, we design a new setting based on the concept of dynamic risk measures where the aim is to find a policy with the best risk-adjusted total discounted outcome. We provide a theoretical analysis of multiarmed bandit problem with respect to this novel setting and propose a priority-index heuristic which gives *risk-averse allocation indices* having a structure similar to Gittins index. Although an optimal policy is shown not always to have index-based form, empirical results express the excellence of this heuristic and show that with *risk-averse allocation indices* we can achieve optimal or near-optimal interpretable policies.**

***Index Terms*—Coherent risk measures, dynamic allocation index, dynamic risk-aversion, Gittins index, multiarmed bandit (MAB).**

## I. Introduction

In multiarmed bandit (MAB) problem, a gambler facing a line of arms (slot machines) should decide on which arm to play at each step. In classical MAB, each arm represents an independent Markov chain. Whenever a specific arm is played, its state changes to a new one with respect to state transition probabilities, providing an income depending on the current state. On the other hand, the arms that are not played maintain their current states. The typical goal is to maximize the expected value of the discounted total reward over an infinite horizon. This problem arises in many practical applications, such as clinical trials, portfolio design, Internet advertisement, and adaptive routing in networks.

In order to solve MAB with Markovian structure, we may consider the problem as a Markov decision process (MDP) and apply Markov decision theory to find the best policy. However, the size of bandit problem increases exponentially as the number of arms increases; therefore, solving large-scale MABs would be computationally challenging. In this regard, Gittins [1] provides a decomposition for the problem by proposing an index policy approach, where these indices are called *dynamic allocation indices*, afterwards referred to as *Gittins indices*. Regarding the specific setting of that problem, he shows that an index can be allocated to each state with respect to the data of the arm so that, in each step, by playing the arm having the highest index, the maximum expected total discounted reward is obtained.

Manuscript received January 27, 2020; revised January 29, 2020 and July 24, 2020; accepted January 6, 2021. Date of publication January 25, 2021; date of current version November 4, 2021. Recommended by Associate Editor Q. S. Jia. *(Corresponding author: Özlem Çavuş.)*

The authors are with the Department of Industrial Engineering, Bilkent University, Ankara 06800, Turkey (e-mail: milad@bilkent.edu.tr; ozlem.cavus@bilkent.edu.tr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2021.3053539.

Digital Object Identifier 10.1109/TAC.2021.3053539

Generalizations of Gittins work usually are based on new frameworks proposed as the consequence of relaxing some assumptions provided in the original study [1]. In this regard, Whittle [2] introduces *Whittle indices* to solve restless MAB problem in which states of the arms not being played also change. He expresses this index as the Lagrangian multiplier corresponding to the activation constraint, which specifies the number of arms activated at each step.

One other important assumption of Gittins work is that the decision-maker is risk-neutral, i.e., the aim is to maximize the expected total reward. This assumption can be a limitation when the decision-maker is risk-averse and wants to manage the variability of total reward. Applications of MAB with risk-aversion include portfolio selection [3], energy [4], routing [5], ambulance redeployment [6], and clinical trials [7]. Despite this wide range of application areas, literature on incorporating risk into MABs with Markovian setting is very poor. Denardo *et al.* [8] introduce the concept of risk in Markovian setting by employing concave utility functions. They introduce a state ranking method and illustrate that it is optimal to play the arm with the state of the highest rank. Improvements and modifications of this approach are presented in [9]. Another work with risk-aversion in Markovian setting is [5]. They study risk-aversion in route choosing, where they model the problem of choosing between a random or a safe selection as a one-armed bandit problem under different information regimes. In [5], similar to [8], the risk preference is incorporated into the model using utility functions.

As described above, existing studies on selection of the arm with the best risk-return tradeoff is limited to incorporating the risk by using some utility functions. However, utility functions have been the subject of considerable criticism over the years. The greatest burden of this approach is that the decision-maker needs to specify an appropriate utility function with respect to the degree of risk-aversion, which is not always an easy task to perform. It is also argued that it is impossible to measure utility across individuals objectively. Moreover, the resulting solutions may be difficult to interpret [10]. In order to handle these difficulties, in this article, we propose a different approach for risk incorporation by using coherent risk measures [11]. To model our problem, we employ the idea of the study [12], where discounted risk-averse discrete-time Markov models with infinite horizon are formulated using dynamic coherent measures of risk, and value and policy iteration algorithms are proposed to find an optimal policy.

Our article aims to develop the classical work of Gittins [1] and provide index-based solutions for risk-averse MAB with dynamic coherent risk measures. In this regard, we seek to achieve the following:

1) the characterization of an indexable and decomposable problem close to risk-averse MAB;

2) definition of indices to obtain an optimal solution to this new problem;

3) an algorithm to compute these indices;

4) an experimental approach to check how good these indices are for original problem.

To the best of authors' knowledge, this version of MAB is new, and no study exists on this setting.

## II. Problem Description and Preliminaries

In the classical risk-neutral MAB, the aim is to maximize the expected total discounted reward. More specifically, MAB can be modeled as an MDP, in which a decision-maker decides which of $K$ arms to play at each decision step $t \in \mathbb{N} = \{1, 2, 3, \dots\}$. Here, we are considering a risk-averse MAB with dynamic risk measures. For mathematical convenience, we take into account negative of rewards, which can be interpreted as costs. The details of our problem setting are described as follows.

1) With the assumption that there exists no common state in distinct arms, let $\mathcal{X} = \bigtimes_{i=1}^{K} \mathcal{X}^i$ be the *state space* of the resulting MDP, where each arm $i$ is a Markov chain with a finite state space $\mathcal{X}^i$, $i \in \mathcal{K} = \{1, 2, \dots, K\}$.
2) At step $t \in \mathbb{N}$, an *action* $u_t = \{u_t^1, u_t^2, \dots, u_t^K\}$ is applied, where $u_t^i \in \{-1, 0\}$ is the action applied to arm $i \in \mathcal{K}$. Here, $u_t^i = 0$ denotes that arm $i$ is played at step $t$; on the other hand, $u_t^i = -1$ represents that arm $i$ is not played. The *admissible actions* are the ones for which exactly one element of $u_t$ equals zero at each step $t \in \mathbb{N}$. That is, at each step, under any choice of admissible actions, the state of one and only one arm evolves independently from the states of other arms in a Markovian manner, and the states of the remaining arms are frozen.
3) The state of the arm we play at each step changes in a Markovian fashion with respect to the transition probabilities $P_{mn}^i(0) = \mathbb{P}(x_{t+1}^i = n \mid x_t^i = m,\ u_t^i = 0), i \in \mathcal{K}, m, n \in \mathcal{X}^i, t \in \mathbb{N}$. However, the state of an arm remains unchanged if it is not played, that is, for all $i \in \mathcal{K}$, $m, n \in \mathcal{X}^i$, $t \in \mathbb{N}$, we have
$$P_{mn}^i(-1) = \mathbb{P}(x_{t+1}^i = n \mid x_t^i = m,\ u_t^i = -1) = \begin{cases} 1, & \text{if } m = n \\ 0, & \text{otherwise.} \end{cases}$$
4) For each $i \in \mathcal{K}$, the *cost function* $g^i : \mathcal{X}^i \to \mathbb{R}_-$ is the finite and nonpositive cost of playing arm $i$ at state $x^i$. We define $c^i(x^i, u^i)$ as the cost incurred by arm $i$ at state $x^i \in \mathcal{X}^i$ under action $u^i \in \{-1, 0\}, i \in \mathcal{K}$:
$$c^i(x^i, u^i) := \begin{cases} g^i(x^i), & \text{if } u^i = 0 \\ 0, & \text{if } u^i = -1. \end{cases} \tag{1}$$
We also define $c(x, u)$ representing the cost over all arms, that is, $c(x, u) = \sum_{i \in \mathcal{K}} c^i(x^i, u^i)$, $x = (x^1, \dots, x^K) \in \mathcal{X}$, and $u = (u^1, \dots, u^K)$ where $u^i \in \{-1, 0\}, i \in \mathcal{K}$.
5) We denote a stationary (Markov) *policy* as $\pi : \mathcal{X} \to \{-1, 0\}^K$. Note that, with respect to our notation, $u_t = \pi(x_t)$, that is, both denote the decision to be taken at state $x_t \in \mathcal{X}$ for $t \in \mathbb{N}$. We also define $\pi^i : \mathcal{X}^i \to \{-1, 0\}$. which prescribes the action to be taken for arm $i \in \mathcal{K}$ at a particular state.

Here, we look for a stationary Markov policy in order to minimize the total risk-averse discounted cost incurred over an infinite horizon. With respect to the theory of [12], an infinite horizon stationary MDP with dynamic risk measures has a stationary optimal policy and it is possible to find this policy using *value iteration* or *policy iteration* algorithms. However, these algorithms may be computationally expensive or even practically infeasible for large-scale real-life problems. Furthermore, in some cases, the structure of optimal policy makes it difficult to interpret, which creates complication while applying them in real-life problems. Motivated to overcome these issues, we seek good index-based heuristic policies.

Before providing our model, we first introduce dynamic risk measures. Consider a probability space $(\Omega, \mathcal{F}, P)$, a filtration $\{\emptyset, \Omega\} = \mathcal{F}_1 \subset \dots \subset \mathcal{F}_T \subset \mathcal{F}$, and an adapted sequence of random variables $Z_t$, $t \in \{1, \dots, T\}$. Define the spaces $\mathcal{Z}_t$ of $\mathcal{F}_t$-measurable random variables on $\Omega$, $t \in \{1, \dots, T\}$ and $\mathcal{Z}_{1,T} = \mathcal{Z}_1 \times \dots \times \mathcal{Z}_T$. In our case, since each $\mathcal{F}_t$ is finite, the spaces $\mathcal{Z}_t$ can be identified with finite-dimensional vector spaces. With respect to the above setting, each policy would result in cost sequences of $Z_t = c(x_t, u_t) \in \mathcal{Z}_t$.

In this context, $\rho_t : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$, $t \in \{1, \dots, T\}$ satisfying the following axioms is defined as *one-step conditional risk measure* (see [13] and [14]):

A1) $\rho_t(\alpha Z + (1-\alpha)W) \le \alpha\rho_t(Z) + (1-\alpha)\rho_t(W) \quad \forall \alpha \in (0,1),\ Z, W \in \mathcal{Z}_{t+1}$;
A2) if $Z \preceq W$, then $\rho_t(Z) \le \rho_t(W)\ \forall Z, W \in \mathcal{Z}_{t+1}$;
A3) $\rho_t(Z + W) = Z + \rho_t(W)\ \forall Z \in \mathcal{Z}_t,\ W \in \mathcal{Z}_{t+1}$;
A4) $\rho_t(\alpha Z) = \alpha\rho_t(Z)\ \forall Z \in \mathcal{Z}_{t+1}, \alpha \ge 0$.

Conditions (A1)–(A4) are called *convexity*, *monotonicity*, *translation invariance*, and *positive homogeneity*, respectively. They are analogous to the axioms introduced by [11] for coherent measures of risk.

Two important coherent risk measures, which have been extensively studied in the literature, are *first-order mean-semideviation* [15] and *mean-average value-at-risk* (mean-AVaR) [16]. We now provide the conditional versions of these two risk measures. Conditional first-order mean-semideviation risk measure is defined as
$$\rho_t(Z_{t+1}) = \mathbb{E}[Z_{t+1}|\mathcal{F}_t] + \kappa\,\mathbb{E}[(Z_{t+1} - \mathbb{E}[Z_{t+1}|\mathcal{F}_t])_+|\mathcal{F}_t] \tag{2}$$
where $\kappa \in [0, 1]$ and $(a)_+ := \max\{a, 0\}$ for $a \in \mathbb{R}$. Given two parameters of $\alpha \in (0, 1)$ and $\lambda \in [0, 1]$, conditional mean-AVaR risk measure is defined as
$$\rho_t(Z_{t+1}) = \lambda\,\mathbb{E}[Z_{t+1}|\mathcal{F}_t] + (1-\lambda)\text{AVaR}_\alpha(Z_{t+1}|\mathcal{F}_t) \tag{3}$$
where conditional $\text{AVaR}_\alpha$ can be represented as
$$\text{AVaR}_\alpha(Z_{t+1}|\mathcal{F}_t) = \min_{\eta \in \mathcal{Z}_t} \left\{ \eta + \frac{1}{1-\alpha}\,\mathbb{E}\left[(Z_{t+1} - \eta)_+|\mathcal{F}_t\right] \right\}.$$

A *dynamic risk measure* is a sequence of one-step conditional risk measures (see [12] and the references therein). Thereby, dynamic risk measure $\varrho_{1,T}^\beta : \mathcal{Z}_{1,T} \mapsto \mathcal{Z}_1$ on a finite horizon of length $T$ with discount factor $\beta \in (0, 1)$ is defined as follows:
$$\varrho_{1,T}^\beta(Z_1, Z_2, \dots, Z_T) := Z_1 + \rho_1\big(\beta Z_2 + \rho_2\big(\beta^2 Z_3 + \dots + \rho_{T-1}\big(\beta^{T-1} Z_T\big)\dots\big)\big). \tag{4}$$
Accordingly, in the infinite horizon, the dynamic risk measure can be described as
$$\varrho^\beta(Z_1, Z_2, Z_3, \dots) = \lim_{T \to \infty} \varrho_{1,T}^\beta(Z_1, Z_2, \dots, Z_T). \tag{5}$$

Applying the dynamic risk measure (5), we can now evaluate the risk of the cost sequences $c(x_t, u_t) \in \mathcal{Z}_t$, $t \in \mathbb{N}$ resulted by policy $\pi$. Let $R(x_1)$ be the optimal risk-averse total discounted cost under the constraint that one arm will be played at each step
$$\begin{aligned} R(x_1) &= \min_{\pi \in \Pi} \varrho^\beta\left(c(x_1, u_1), c(x_2, u_2), \dots\right) \\ &= \min_{\pi \in \Pi} \sum_{i \in \mathcal{K}} c^i(x_1^i, u_1^i) + \rho_1\left(\beta \sum_{i \in \mathcal{K}} c^i(x_2^i, u_2^i)\right. \\ &\quad + \left.\rho_2\left(\beta^2 \sum_{i \in \mathcal{K}} c^i(x_3^i, u_3^i) + \dots\right)\right) \end{aligned} \tag{6}$$
where $\Pi$ is the class of stationary admissible policies for our problem, that is, at each step, we are only allowed to play one arm. Note that second equality is due to definition of $c(x_t, u_t)$. Since additivity of risk measure is one main assumption in [1] to guarantee the optimality

of an index policy, it follows that index policies may not be optimal for the risk-averse problem, as coherent risk measures are subadditive. Therefore, optimal policy can only be achieved by performing the far more difficult task of solving the risk-averse dynamic programming equations. At this point, following the classic works of [1] and [2], and based on the theory and presentation of [17], we aim to find a heuristic solution of the problem in the form of index policies. Therefore, in order to find proper actions in each state, we need to design a calibrating function for each arm. To come up with such indices, we need to perform a decomposition of our problem into $K$ separate problems, where each of which would relate to one individual arm. Moreover, with the help of these separate problems, we can acquire the appropriate calibrating functions.

## III. Relaxation and Decomposition

With the aim of attaining an index heuristic for problem (6), we obtain an equivalent optimization problem by extending $\Pi$ to the class $\Pi'$ of stationary policies, which allows a free choice of action, i.e., no arm or more than one arm can also be played at a step

$$R(x_1) = \min_{\pi\in\Pi'} \sum_{i\in\mathcal{K}} c^i(x_1^i,u_1^i) + \rho_1\Bigg(\beta\sum_{i\in\mathcal{K}} c^i(x_2^i,u_2^i) + \rho_2\left(\beta^2\sum_{i\in\mathcal{K}} c^i(x_3^i,u_3^i)+\cdots\right)\Bigg) \tag{7}$$

$$\text{s.t.} \sum_{i\in\mathcal{K}} u_t^i = 1-K \;\forall t\in\mathbb{N}, \tag{8}$$

$$u_t^i \in \{-1,0\}\;\forall i\in\mathcal{K},\; t\in\mathbb{N}. \tag{9}$$

Here, constraint (8) implies that playing exactly one arm is allowed at each step. Further, we can relax this constraint by replacing it with

$$\varrho^\beta\left(\sum_{i\in\mathcal{K}} u_1^i, \sum_{i\in\mathcal{K}} u_2^i, \dots\right) = \frac{1-K}{1-\beta}$$

which can also be written as follows using definitions (4) and (5):

$$\sum_{i\in\mathcal{K}} u_1^i + \rho_1\Bigg(\beta\sum_{i\in\mathcal{K}} u_2^i + \rho_2\left(\beta^2\sum_{i\in\mathcal{K}} u_3^i+\cdots\right)\Bigg) = \frac{1-K}{1-\beta}. \tag{10}$$

Now, replacing constraint (8) with (10) and relaxing it in Lagrangian manner, we obtain the following Lagrangian dual function:

$$\mathcal{L}_D(\nu,x_1) = \min_{\pi\in\Pi'} \;\; \mathcal{L}(\nu,x_1)$$
$$\text{s.t.} \quad u_t^i\in\{-1,0\}\;\forall i\in\mathcal{K},\; t\in\mathbb{N} \tag{11}$$

where $\nu\in\mathbb{R}$ is the Lagrangian multiplier and

$$\mathcal{L}(\nu,x_1) = \sum_{i\in\mathcal{K}} c^i(x_1^i,u_1^i) + \rho_1\Bigg(\beta\sum_{i\in\mathcal{K}} c^i(x_2^i,u_2^i) + \rho_2\left(\beta^2\sum_{i\in\mathcal{K}} c^i(x_3^i,u_3^i)+\cdots\right)\Bigg) + \nu\Bigg(\sum_{i\in\mathcal{K}} u_1^i + \rho_1\Bigg(\beta\sum_{i\in\mathcal{K}} u_2^i + \rho_2\left(\beta^2\sum_{i\in\mathcal{K}} u_3^i+\cdots\right)\Bigg)\Bigg) - \nu\left(\frac{1-K}{1-\beta}\right). \tag{12}$$

Note that, by duality, we have

$$\mathcal{L}_D(\nu,x_1) \le R(x_1)$$

for any $\nu\in\mathbb{R}$. Therefore, if we restrict $\nu\in\mathbb{R}_+$, the above inequality is still satisfied. Before going any further, we need to show subadditivity of dynamic risk measures, which will be used to obtain an indexable problem.

*Lemma III.1:* Consider the dynamic risk measure $\varrho_{1,T}^\beta$ in (4) and random costs $Z_t^i\in\mathcal{Z}_t$, $t\in\{1,\dots,T\}$, $i\in\mathcal{K}$. Then we have

$$\varrho_{1,T}^\beta\left(\sum_{i\in\mathcal{K}} Z_1^i, \sum_{i\in\mathcal{K}} Z_2^i, \dots, \sum_{i\in\mathcal{K}} Z_T^i\right) \le \sum_{i\in\mathcal{K}} \left(Z_1^i + \rho_1\left(\beta Z_2^i + \cdots + \rho_{T-1}\left(\beta^{T-1}Z_T^i\right)\dots\right)\right).$$

*Proof:* From axioms (A1) and (A4), we get

$$\rho_{T-1}\left(\beta^{T-1}\sum_{i\in\mathcal{K}} Z_T^i\right) \le \sum_{i\in\mathcal{K}} \rho_{T-1}\left(\beta^{T-1}Z_T^i\right). \tag{13}$$

Then, we obtain

$$\rho_{T-2}\left(\beta^{T-2}\sum_{i\in\mathcal{K}} Z_{T-1}^i + \rho_{T-1}\left(\beta^{T-1}\sum_{i\in\mathcal{K}} Z_T^i\right)\right) \le \rho_{T-2}\left(\sum_{i\in\mathcal{K}}\left(\beta^{T-2}Z_{T-1}^i + \rho_{T-1}\left(\beta^{T-1}Z_T^i\right)\right)\right) \le \sum_{i\in\mathcal{K}} \rho_{T-2}\left(\beta^{T-2}Z_{T-1}^i + \rho_{T-1}\left(\beta^{T-1}Z_T^i\right)\right)$$

where the first inequality follows from monotonicity axiom (A2) and inequality (13), and the second inequality is due to (A1) and (A4). Iterating similarly down to $t=1$, the assertion of the lemma follows. ■

Let us now define a function $\mathcal{L}'(\nu,x_1)$ as follows:

$$\mathcal{L}'(\nu,x_1) = \sum_{i\in\mathcal{K}}\Big(c^i(x_1^i,u_1^i) + \rho_1\big(\beta c^i(x_2^i,u_2^i) + \rho_2(\beta^2 c^i(x_3^i,u_3^i)+\cdots)\big) + \nu\left(u_1^i + \rho_1\big(\beta u_2^i + \rho_2(\beta^2 u_3^i+\cdots)\big)\right)\Big) - \nu\left(\frac{1-K}{1-\beta}\right) \tag{14}$$

where $\nu\in\mathbb{R}_+$ and $x_1\in\mathcal{X}$. Note that, for $\nu\in\mathbb{R}_+$, using Lemma III.1 and (5), we get

$$\mathcal{L}(\nu,x_1) \le \mathcal{L}'(\nu,x_1). \tag{15}$$

We now define a dual function $\mathcal{L}'_D(\nu,x_1)$ based on $\mathcal{L}'(\nu,x_1)$ and use it in order to obtain an approximation of $R(x_1)$

$$\mathcal{L}'_D(\nu,x_1) = \min_{\pi\in\Pi'} \mathcal{L}'(\nu,x_1)$$
$$\text{s.t. } u_t^i\in\{-1,0\}\;\forall i\in\mathcal{K},\; t\in\mathbb{N} \tag{16}$$

where $\nu\in\mathbb{R}_+$. Note that, from (15), for $\nu\in\mathbb{R}_+$, we obtain

$$\mathcal{L}_D(\nu,x_1) \le \mathcal{L}'_D(\nu,x_1). \tag{17}$$

Although $\mathcal{L}_D(\nu,x_1)$ is a lower bound for $R(x_1)$, $\mathcal{L}'_D(\nu,x_1)$ might not be smaller than $R(x_1)$. On the other hand, problem (16) provides us with an index-based policy, which is an interpretable feasible policy for the risk-averse MAB problem (6). For the rest of the study, we will derive this index-based policy and computationally show that it is close to an optimal policy of (6).

Note that problem (16) can be decomposed into $K$ subproblems, each of which relates to different arms. For each arm $i\in\mathcal{K}$, we have

$$\mathcal{L}'_{Di}(\nu, x_1^i) = \min_{\pi^i \in \Pi'^i} c^i(x_1^i, u_1^i) + \rho_1 \big(\beta c^i(x_2^i, u_2^i) + \rho_2(\beta^2 c^i(x_3^i, u_3^i) + \cdots)\big) + \nu\big(u_1^i + \rho_1\big(\beta u_2^i + \rho_2(\beta^2 u_3^i + \cdots)\big)\big) \quad \text{s.t. } u_t^i \in \{-1, 0\}\ \forall t \in \mathbb{N} \tag{18}$$

where $\Pi'^i$ is a class of stationary policies for arm $i$ which has no restriction regarding the action to be taken at each step. Also, the term $\nu(\frac{1-K}{1-\beta})$ in (14) does not appear in subproblem (18), without changing the optimal policies.

In the next section, we discuss the indexability of subproblem (18) and explain the structure of the indices by developing a calibrating function for each arm.

## IV. Indexability and Indices

In this step, in search of the structure of optimal policy for subproblem (18), we need to follow [2] and [17]. Regarding Whittle's discussion in [2], the indexability argument is related with the class of bandit problems centering on individual arms. Therefore, we focus our attention on subproblem (18), which we call as arm $i$, and provide below a lemma describing a key property of its optimal policy.

*Lemma IV.1:* Consider the subproblem (18). For $i \in \mathcal{K}$, if action "not play" is optimal for state $x_t^i \in \mathcal{X}^i$, then it is also optimal for state $x_{t+1}^i \in \mathcal{X}^i\ \forall t \in \mathbb{N}$.

*Proof:* If action "not play" is optimal for $x_t^i$, with respect to the rested property of the problem, in the next step, the arm will be in the same state. And due to stationary property of the feasible policies, the same action will be optimal for $x_{t+1}^i$. ■

Now, in order to find an optimal policy of (18), we will consider the evolution of arm $i \in \mathcal{K}$ under some policy $\pi^i \in \Pi'^i$ starting from the initial state $x_1^i$, and from Lemma IV.1, we will restrict our attention to two specific stationary policies of never playing as well as playing the arm from the beginning until some random time $\tau^i$, which, for brevity, will be referred to as $\tau^i$ policy. To do so, we need to find a proper calibration function $\nu^i : \mathcal{X}^i \to \mathbb{R}_+$ for arm $i \in \mathcal{K}$. The index $\nu^i(x^i)$ of arm $i$ when in state $x^i \in \mathcal{X}^i$ is defined as the value of $\nu$ which makes two actions of "play" and "not play" equally attractive.

However, for $\nu^i(x^i)$ to be meaningful, it requires to induce a consistent ordering of the states, such that if arm $i$ being at state $x^i$ is not played under $\nu^i(x^i)$, it will also not be played under $\nu > \nu^i(x^i)$. For this, we define $\Theta_\nu^i$ as the set of states of arm $i$, for which the action is "not play" for a given value of $\nu$. According to [2], arm $i \in \mathcal{K}$ is indexable if there exists a family of policies in which each policy, corresponding to a specific $\nu \in \mathbb{R}_+$, is optimal for subproblem (18) and the cardinality of $\Theta_\nu^i$ is nondecreasing as $\nu$ increases. Under these conditions, the corresponding index would be

$$\nu^i(x^i) = \inf\{\nu \in \mathbb{R}_+ : x^i \in \Theta_\nu^i\}. \tag{19}$$

As an interpretation of (19), $\nu^i(x^i)$ is a fair charge we pay to go from "not play" to "play" when arm $i$ is in state $x^i$. For any $\nu \in \mathbb{R}_+$, if $\nu \le \nu^i(x^i)$, then playing arm $i$ is optimal; otherwise the optimal action will be not playing it. In this case, for a given $\nu \in \mathbb{R}_+$, the stopping time $\tau^i$ can be computed as

$$\tau^i = \inf\{t \ge 2 : x_t^i \in \Theta_\nu^i\}. \tag{20}$$

From Lemma IV.1 and using (20), we have $x_t^i \in \Theta_\nu^i, \forall t \ge \tau^i$. Therefore, finding an optimal policy boils down to searching for some stopping time, which would be obtained by employing the indices described in Definition IV.1. Here, with a slight abuse of notation and referring to (6) in [12], we consider $\varrho^\beta_{1,\tau-1}(Z_1, Z_2, \ldots, Z_{\tau-1}) = \varrho^\beta(Z_1, Z_2, \ldots, Z_{\tau-1}, 0, 0, \ldots)$.

*Definition IV.1:* For $i \in \mathcal{K}$, the *risk-averse allocation index* (RAI) for each initial state $x_1^i \in \mathcal{X}^i$ is given by

$$\nu^i(x_1^i) := \sup_{\tau^i \ge 2} \frac{\varrho^\beta_{1,\tau^i-1}(c^i(x_1^i, 0), c^i(x_2^i, 0), \ldots, c^i(x^i_{\tau^i-1}, 0))}{\varrho^\beta_{1,\tau^i-1}(-1, -1, \ldots, -1)}. \tag{21}$$

Note that based on (19) and (20), the stopping time in which we attain the supremum in (21) is $\tau^{i*} = \inf\{t \ge 2 : x_t^i \in \Theta^i_{\nu^i(x_1^i)}\}$.

*Remark:* It is not difficult to show finiteness of RAI using axioms (A2) and (A4) of coherent risk measures. RAI is, therefore, guaranteed to be finite and has the similar structure and interpretation as the Gittins index and may be considered as a generalized form of Gittins index that is capable of taking dynamic risk-aversion into account.

In the next theorem, we show that each arm $i \in \mathcal{K}$ is indexable. For this purpose, let us define $\mathcal{T}^i$ as the class of stationary positive-valued stopping times for arm $i$, that is, the set of all feasible $\tau^i$ policies for all possible $\Theta_\nu^i \subseteq \mathcal{X}^i,\ \nu \in \mathbb{R}_+$.

*Theorem IV.1:* Each risk-averse arm $i \in \mathcal{K}$ is indexable with the RAI introduced in Definition IV.1.

*Proof:* Consider the subproblem (18) for arm $i \in \mathcal{K}$ and for a given constant $\nu \in \mathbb{R}_+$. Given the previous discussion, we have two policies to consider: either playing the arm until some stopping time or not playing it at all. Playing this arm from the initial state $x_1^i \in \mathcal{X}^i$ until stopping time $\tau^i \in \mathcal{T}^i$ results in the following objective function value:

$$\varrho^\beta(c^i(x_1^i, 0), \ldots, c^i(x^i_{\tau^i-1}, 0), 0, 0, \ldots) + \nu\varrho^\beta(\underbrace{0, 0, \ldots, 0}_{\tau^i-1}, -1, -1, \ldots) \tag{22}$$

for which referring to (6) in [12], we have

$$\varrho^\beta(c^i(x_1^i, 0), \ldots, c^i(x^i_{\tau^i-1}, 0), 0, 0, \ldots) = \varrho^\beta_{1,\tau^i-1}(c^i(x_1^i, 0), \ldots, c^i(x^i_{\tau^i-1}, 0)).$$

Similarly, if the arm is not played in state $x_1^i$, then from Lemma IV.1, the objective function value is

$$\varrho^\beta(0, 0, \ldots) + \nu\varrho^\beta(-1, -1, \ldots) \tag{23}$$

where $\varrho^\beta(0, 0, \ldots) = 0$ as a result of axiom (A4). Therefore, the "play" action is strictly optimal for $x_1^i$ when there exists a stopping policy $\tau^i \in \mathcal{T}^i$ for which the quantity in (22) falls behind the quantity in (23), that is

$$\varrho^\beta_{1,\tau^i-1}(c^i(x_1^i, 0), \ldots, c^i(x^i_{\tau^i-1}, 0)) + \nu\varrho^\beta(\underbrace{0, \ldots, 0}_{\tau^i-1}, -1, -1, \ldots) < \nu\varrho^\beta(-1, -1, \ldots). \tag{24}$$

Organizing this inequality, we obtain

$$\varrho^\beta_{1,\tau^i-1}(c^i(x_1^i, 0), c^i(x_2^i, 0), \ldots, c^i(x^i_{\tau^i-1}, 0)) < \nu\big(\varrho^\beta(-1, -1, \ldots) - \varrho^\beta(\underbrace{0, \ldots, 0}_{\tau^i-1}, -1, -1, \ldots)\big)$$

$$= \nu\bigg(-\sum_{t=0}^{\tau^i-2}\beta^t\bigg) = \nu\varrho^\beta_{1,\tau^i-1}(-1, -1, \ldots, -1)$$

which results in

$$\varrho^\beta_{1,\tau^i-1}(c^i(x_1^i, 0), \ldots, c^i(x^i_{\tau^i-1}, 0)) < \nu\varrho^\beta_{1,\tau^i-1}(-1, \ldots, -1).$$

Therefore, action "play" is optimal for arm $i \in \mathcal{K}$ at initial state $x_1^i$, whenever there exists $\tau^i \in \mathcal{T}^i$ for which

$$\frac{\varrho^{\beta}_{1,\tau^i-1}(c^i(x_1^i,0),c^i(x_2^i,0),\ldots,c^i(x^i_{\tau^i-1},0))}{\varrho^{\beta}_{1,\tau^i-1}(-1,-1,\ldots,-1)} > \nu$$

and, from (21), we have $\nu^i(x_1^i) > \nu$.

Similarly, action "not play" is strictly optimal for $x_1^i$ if $\forall \tau^i \geq 2$

$$\frac{\varrho^{\beta}_{1,\tau^i-1}(c^i(x_1^i,0),c^i(x_2^i,0),\ldots,c^i(x^i_{\tau^i-1},0))}{\varrho^{\beta}_{1,\tau^i-1}(-1,-1,\ldots,-1)} \leq \nu^i(x_1^i) < \nu.$$

In conclusion, action "play" is optimal in state $x_1^i$ if and only if $\nu^i(x_1^i) \geq \nu$. Also, action "not play" is optimal in this state if and only if $\nu^i(x_1^i) \leq \nu$. From this argument, we can infer the existence of a family of optimal policies whose associated inactive sets $\Theta^i_\nu$, $i \in \mathcal{K}$ are nondecreasing in $\nu$. This argument establishes the indexability of each arm $i \in \mathcal{K}$ with respect to Definition IV.1. ■

In addition, we need to show that, in our definition of RAI, the supremum is always achieved by stopping time(s) that have a simple characterization.

*Proposition IV.1:* For each arm $i \in \mathcal{K}$, starting from $x_1^i \in \mathcal{X}^i$, a positive stopping time $\tau^i \in \mathcal{T}^i$ achieves the supremum in Definition IV.1.

*Proof:* For a fixed nonnegative value of $\nu$, with respect to the previous proof, we know that it is optimal to play at step $t$ if and only if $\nu^i(x_t^i) \geq \nu$. It is also optimal not to play if and only if $\nu^i(x_t^i) \leq \nu$. Particularly, it is optimal to both play and not to play at step one if $\nu^i(x_1^i) = \nu$. If we take action "play" at step one, then it would be optimal to continue taking this action before step $\tau^i$ if $\nu^i(x_s^i) \geq \nu$ for $1 \leq s \leq \tau^i - 1$. In this case, action "not play" will be optimal afterwards if $\nu^i(x^i_{\tau^i}) \leq \nu$. Thus, with respect to the rested nature of our problem and by assuming that decisions are taken in a stationary manner, only the sequences of actions where we either do not play or we play throughout $[1, \tau^i)$, where $\tau^i \in \mathcal{T}^i$, are optimal. Equating the objective values of these two policies for subproblem (18), provides us the indifference between these alternative actions for state $x_1^i$. The result of this, which is calculated by using equality in (24), concludes that

$$\nu^i(x_1^i) = \frac{\varrho^{\beta}_{1,\tau^i-1}(c^i(x_1^i,0),c^i(x_2^i,0),\ldots,c^i(x^i_{\tau^i-1},0))}{\varrho^{\beta}_{1,\tau^i-1}(-1,-1,\ldots,-1)}$$

which establishes the statement of proposition. ■

Note that RAI policy is a priority index policy, that is, a policy based on playing the arm with the highest index value at each decision step. It is an optimal solution to problem (16) but does not necessarily provide optimal policies for our risk-averse MAB. The reason lies in the fact that problem (16) is the result of several relaxations of the original problem (6), including the relaxation of activation constraint in a Lagrangian manner as well as taking the summations over all arms out in the objective function with respect to subadditivity and monotonicity properties of dynamic risk measures. Therefore, RAI policy is a feasible policy for our original problem (6) and provides an upper bound to the risk-averse value function $R(x_1)$ given in (6).

An important exception to the discussion above is *one-armed bandit problem*, as it is indexable with RAI, which we state without proof in Theorem IV.2. One-armed bandit problem is also known as the two-armed bandit problem, where one arm contains a single state and the other one is a Markovian arm with finite number of states. This problem is of interest from both applied and theoretical viewpoints. The reader is referred to [5], [18], and [19] for some interesting applications of one-armed bandit problem.

**Algorithm 1:** Computing RAI for Arm $i$.

**for** $k = 1$ to $|\mathcal{X}^i|$ **do**
  **if** $k = 1$ **then**
    $\Phi_k^i = \mathcal{X}^i$
  **else**
    $\Phi_k^i = \Phi_{k-1}^i \backslash \alpha_{k-1}^i$
  **end if**
  $\alpha_k^i = \arg\max_{x^i \in \Phi_k^i} N^i_{x^i}(\Phi_k^i)$
  $\nu^i(k) = N^i_{\alpha_k^i}(\Phi_k^i)$
**end for**
**return** $\nu^i(k),\ k \in \{1,\ldots,|\mathcal{X}^i|\}$

*Theorem IV.2:* One-armed bandit problem is indexable with the RAIs introduced in Definition IV.1.

## V. Computation of the Risk-Averse Indices

So far, we presented the structure of RAI for risk-averse MAB problem. Now, we need to provide an algorithm for the computation of these indices. As stated, RAI has a similar structure to Gittins index, thereby its computation is not that much more difficult than the Gittins index calculation. The approach we suggest here is a modified version of the algorithm presented by [20] for computing Gittins indices. For both cases, we solve single-arm optimal stopping problems. Let $x^i$ be a generic state of arm $i$ and consider the arm evolving under "play" action starting from $x^i$. In contrast to Section IV, where the set of states with "not play" action is provided with respect to a given $\nu$ as $\Theta^i_\nu$, here, we define $\Phi_k^i$ as the *stopping set* related to the $k$th highest index state. Therefore, $\Phi_k^i$ has the same interpretation as $\Theta^i_{\nu^i(x^i)}$ provided that $x^i$ has the $k$th highest index value. In other words, $\Phi_k^i$ is the set of all states of arm $i \in \mathcal{K}$ excluding the best $k-1$ index states. In this regard, we define $\tau^i_{x^i}(\Phi_k^i)$ as the first time that the state of arm $i$ lies in subset $\Phi_k^i$ starting from $x^i$. Then, $\tau^i_{x^i}(\Phi_k^i)$ provides the supremum in (21) given that $x^i$ has the $k$th highest index value. Also, for each set $\Phi_k^i \subseteq \mathcal{X}^i$ and $x^i \in \mathcal{X}^i$, we define $N^i_{x^i}(\Phi_k^i)$ as

$$N^i_{x^i}(\Phi_k^i) = \frac{\varrho^{\beta}_{1,\tau^i_{x^i}(\Phi_k^i)-1}(c^i(x^i,0),c^i(x_2^i,0),\ldots,c^i(x^i_{\tau^i_{x^i}(\Phi_k^i)-1},0))}{\varrho^{\beta}_{1,\tau_{x^i}(\Phi_k^i)-1}(-1,-1,\ldots,-1)} \tag{25}$$

which yields the optimal value in (21) as long as $x^i$ is the $k$th highest index state.

Algorithm 1 computes the RAI for arm $i \in \mathcal{K}$, with the inputs as the costs, discount factor, and the parameters of the risk measure. The output of this algorithm, integer state identifiers, $\nu^i(k),\ k \in \{1,\ldots,|\mathcal{X}^i|\}$, provide the RAI of all states of arm $i$, numbered in increasing order of their index values, that is, state 1 having the highest and state $|\mathcal{X}^i|$ having the lowest RAI. In this algorithm, first, we identify the highest index state, where all the states in arm $i$ are members of the stopping set $\Phi_1^i$, imposing the occurrence of the supremum in (21) in $\tau^i = 2$ almost surely. Afterwards, the second-highest index is computed by repeating the same computation on a new stopping set obtained by removing the state with the highest index from previous stopping set. The next steps employ this idea and progress in similar fashion, until the last iteration, where $\Phi^i_{|\mathcal{X}^i|}$ is left with only one state which has the smallest index value.

TABLE I
STATISTICS OF MAXIMUM SUBOPTIMALITY PERCENTAGE OF POLICIES RN AND RA FOR FIRST-ORDER MEAN-SEMIDEVIATION

| $\sigma$ | $\beta$ | | $\kappa=0.25$ | | $\kappa=0.50$ | | $\kappa=0.75$ | | $\kappa=1$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | RN | RA | RN | RA | RN | RA | RN | RA |
| 0.01 | 0.90 | mean | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | max | 0.002 | 0 | 0.004 | 0.002 | 0.008 | 0.001 | 0.015 | 0.006 |
| | 0.95 | mean | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | max | 0.003 | 0.001 | 0.005 | 0.002 | 0.007 | 0.003 | 0.009 | 0.002 |
| 0.5 | 0.90 | mean | 0.020 | 0.003 | 0.071 | 0.011 | 0.142 | 0.019 | 0.230 | 0.032 |
| | | max | 1.137 | 0.356 | 2.355 | 0.953 | 3.586 | 1.136 | 4.838 | 1.800 |
| | 0.95 | mean | 0.017 | 0.003 | 0.060 | 0.010 | 0.126 | 0.020 | 0.207 | 0.030 |
| | | max | 1.035 | 0.487 | 2.501 | 1.143 | 4.030 | 1.769 | 5.628 | 2.362 |
| 1 | 0.90 | mean | 0.046 | 0.010 | 0.160 | 0.031 | 0.329 | 0.059 | 0.557 | 0.087 |
| | | max | 2.050 | 1.103 | 4.555 | 2.100 | 7.536 | 3.209 | 11.810 | 3.813 |
| | 0.95 | mean | 0.035 | 0.008 | 0.133 | 0.023 | 0.282 | 0.048 | 0.487 | 0.067 |
| | | max | 1.571 | 1.167 | 3.942 | 1.870 | 6.637 | 3.134 | 10.003 | 4.535 |

TABLE II
AVERAGE SIMILARITY PERCENTAGES OF POLICIES RN AND RA FOR FIRST-ORDER MEAN-SEMIDEVIATION

| $\sigma$ | $\beta$ | $\kappa=0.25$ | | $\kappa=0.50$ | | $\kappa=0.75$ | | $\kappa=1$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | RN | RA | RN | RA | RN | RA | RN | RA |
| 0.01 | 0.90 | 99.963 | 100 | 99.906 | 99.981 | 99.839 | 99.939 | 99.747 | 99.950 |
| | 0.95 | 99.950 | 100 | 99.919 | 99.975 | 99.844 | 99.952 | 99.722 | 99.963 |
| 0.5 | 0.90 | 98.280 | 99.603 | 96.789 | 99.217 | 95.353 | 98.844 | 93.841 | 98.459 |
| | 0.95 | 98.181 | 99.580 | 96.467 | 99.078 | 94.920 | 98.628 | 93.442 | 98.291 |
| 1 | 0.90 | 98.030 | 99.681 | 96.272 | 99.394 | 94.575 | 98.967 | 93.027 | 98.745 |
| | 0.95 | 97.702 | 99.594 | 95.811 | 99.205 | 93.983 | 98.933 | 92.203 | 98.669 |

## VI. COMPUTATIONAL EXPERIMENTS

In this section, we conduct a series of computational experiments in order to assess the performance of the proposed risk-averse index policy. For each test instance, we compute the following policies:

1) an optimal policy of our problem, which is obtained by using the *risk-averse policy iteration algorithm with convex optimization method* (see [21]);
2) the risk-averse index policy (RA) obtained via RAIs;
3) the risk-neutral policy (RN) obtained via Gittins indices.

We consider a problem of three-armed bandit, where each arm contains four states and randomly generate 1000 test instances. In each test instance, the transition probabilities of each arm under the play action are sampled from $U(0,1)$, where $U$ represents uniform distribution, and the resulting transition matrix is normalized across the rows, i.e., entries of each row is divided by its sum. Following [22], the costs in each state are drawn from a truncated normal distribution with mean value generated from $U(-6,-5)$ and standard deviation $\sigma$ taking values from the set $\{0.01, 0.5, 1\}$. The experiments are conducted with discount factor $\beta \in \{0.9, 0.95\}$.

As risk measures, we employ first-order mean-semideviation, given in (2), with $\kappa \in \{0.25, 0.50, 0.75, 1\}$, as well as mean-AVaR, given in (3), with parameters of $\alpha \in \{0.80, 0.90, 0.95\}$ and $\lambda \in \{0, 0.5\}$. For each test instance, we compare RA and RN policies based on two different performance measures.

1) The suboptimality percentage of policy $\pi \in \{\text{RA}, \text{RN}\}$ for each initial state $x \in \mathcal{X}$. This percentage is computed as $100 \times (R(x) - R^{\pi}(x))/R(x)$, where $R^{\pi}(x)$ denotes the value of objective function in (6) under policy $\pi$.
2) Similarity percentage of policy $\pi \in \{\text{RA}, \text{RN}\}$ compared to optimal policy $\pi^*$. It is computed as $100 \times \sum_{x \in \mathcal{X}} \mathbb{1}_{(\pi,\pi^*)}(x)/|\mathcal{X}|$, where $\mathbb{1}_{(\pi,\pi^*)}(x)$ is equal to one if $\pi^*(x) = \pi(x)$ and zero otherwise.

For each test instance and risk measure, we compute maximum of suboptimality percentages over all states. The maximum suboptimality percentages for first-order mean-semideviation risk measure are summarized in Table I . In the third column, we report mean and maximum of these percentages over 1000 test instances. It can be seen that RA policy is much better than RN policy regarding this performance measure. For instance, for $\sigma = 1$, $\beta = 0.90$, and $\kappa = 1$, the maximum value for maximum suboptimality percentage for RN is 11.81% while it is 3.813% for RA. As indicated in Table I, by increasing the variability of cost values, that is $\sigma$, we observe higher maximum suboptimality percentages for both polices RN and RA. However, this increase is more noticeable for RN. Moreover, increasing the level of risk-aversion (increasing the value of $\kappa$) results in higher maximum suboptimality percentages for both policies; yet, higher increase belongs to RN. Concerning the discount factor $\beta$, no consistent behavior is observed. For instance, when $\kappa = 1$, the worst suboptimality percentage of policy RA corresponding to $\sigma = 0.01$ decreases as $\beta$ increases; on the other hand, it increases for $\sigma = 1$.

The average similarity percentage for this risk measure is reported in Table II. With respect to these results, the performance of our risk-averse index policy RA is exceptionally well. For instance, for $\sigma = 1$, $\beta = 0.95$, $\kappa = 1$, there is more than 6% difference in the average similarity percentages of RN and RA.

Moreover, we focus on the percentage of instances RN and RA policies provide optimal policy. In this respect, RA is optimal in almost 90% of instances, and in no occasion, it is more than 5% suboptimal, whereas RN is optimal only in 70% of the instances and the suboptimality percentage can be more than 10%.

The results related to maximum suboptimality percentage for mean-AVaR risk measure are summarized in Table III. Similar to first-order mean-semideviation, increase in cost value variability $\sigma$ results in increase in maximum suboptimality percentage of RN and RA, where the increase in the former is more notable. And, no consistent behavior is observed with respect to the discount factor $\beta$. However, increasing the level of risk-aversion has a diverse effect in RN and RA. By increasing the level of risk-aversion, i.e., increasing the value of $\alpha$ or reducing $\lambda$, the performance of RN decreases in general. On the other hand, as $\alpha$ increases, the performance of RA, in general, increases. Additionally, for $\alpha \in \{0.80, 0.90\}$, as $\lambda$ decreases, the performance of RA decreases; however, it increases for $\alpha = 0.95$. Our RA policy is superior to RN policy. It can be seen that, when $\sigma = 1, \beta = 0.90, \lambda = 0$, and $\alpha = 0.90$, although the worst suboptimality percentage of RN is 39.692%, it is 2.761% for RA.

The average similarity percentages for this risk measure are summarized in Table IV. In all the performance measures, RA is superior to RN. While the average similarity percentage is more than 99% for RA, it can drop to around 85% for RN.

In general, the performance of our index policy RA is more superior to RN when mean-AVaR is used as the risk measure instead of first-order mean-semideviation. This superiority is more clear when we focus on the maximum values in Table III for $\sigma = 1$, where, in general, the suboptimality percentage of RN is more than 10 times than that of RA.

Additionally, our index policy is optimal in 97% of instances and in no occasion it is more than 7.6% suboptimal, whereas RN policy is optimal only in 50% of the instances and can have a suboptimality percentage around 40%.

Moreover, in a separate experiment, in order to visualize the time efficiency of our index policy compared to solving the corresponding MDP optimally, we record the CPU time of the algorithms for 12

TABLE III
STATISTICS OF MAXIMUM SUBOPTIMALITY PERCENTAGE OF POLICIES RN AND RA FOR MEAN-AVAR

| $\sigma$ | $\beta$ | | $\lambda=0$ | | | | | | $\lambda=0.5$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\alpha=0.80$ | | $\alpha=0.90$ | | $\alpha=0.95$ | | $\alpha=0.80$ | | $\alpha=0.90$ | | $\alpha=0.95$ | |
| | | | RN | RA | RN | RA | RN | RA | RN | RA | RN | RA | RN | RA |
| 0.01 | 0.90 | mean | 0.001 | 0 | 0.001 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | max | 0.061 | 0 | 0.090 | 0 | 0.089 | 0 | 0.016 | 0 | 0.020 | 0 | 0.020 | 0 |
| | 0.95 | mean | 0 | 0 | 0.001 | 0 | 0.001 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | max | 0.047 | 0 | 0.075 | 0 | 0.074 | 0 | 0.012 | 0 | 0.012 | 0 | 0.012 | 0 |
| 0.5 | 0.90 | mean | 1.449 | 0.028 | 1.577 | 0.006 | 1.613 | 0 | 0.454 | 0.018 | 0.499 | 0.002 | 0.512 | 0.002 |
| | | max | 14.103 | 4.677 | 20.102 | 1.323 | 20.053 | 0.044 | 6.555 | 1.979 | 9.155 | 0.232 | 9.072 | 0.237 |
| | 0.95 | mean | 1.341 | 0.021 | 1.465 | 0.003 | 1.506 | 0 | 0.400 | 0.015 | 0.448 | 0.006 | 0.463 | 0.001 |
| | | max | 13.545 | 4.498 | 19.862 | 0.749 | 19.835 | 0.012 | 6.075 | 1.641 | 8.316 | 2.018 | 8.234 | 0.157 |
| 1 | 0.90 | mean | 3.547 | 0.027 | 3.921 | 0.012 | 4.056 | 0.001 | 1.029 | 0.024 | 1.172 | 0.005 | 1.211 | 0.005 |
| | | max | 28.588 | 7.538 | 39.692 | 2.761 | 39.603 | 0.285 | 11.524 | 1.817 | 15.691 | 0.828 | 15.517 | 0.999 |
| | 0.95 | mean | 3.467 | 0.022 | 3.820 | 0.006 | 3.940 | 0.001 | 0.915 | 0.020 | 1.059 | 0.007 | 1.096 | 0.005 |
| | | max | 28.568 | 7.050 | 38.861 | 1.44 | 38.809 | 0.133 | 11.935 | 1.251 | 14.492 | 1.022 | 16.694 | 1.173 |

TABLE IV
AVERAGE SIMILARITY PERCENTAGES OF POLICIES RN AND RA FOR MEAN-AVAR

| $\sigma$ | $\beta$ | $\lambda=0$ | | | | | | $\lambda=0.5$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\alpha=0.80$ | | $\alpha=0.90$ | | $\alpha=0.95$ | | $\alpha=0.80$ | | $\alpha=0.90$ | | $\alpha=0.95$ | |
| | | RN | RA | RN | RA | RN | RA | RN | RA | RN | RA | RN | RA |
| 0.01 | 0.90 | 99.352 | 99.968 | 99.336 | 99.984 | 99.320 | 100 | 99.753 | 99.984 | 99.722 | 99.984 | 99.658 | 100 |
| | 0.95 | 99.288 | 99.968 | 99.272 | 99.984 | 99.257 | 100 | 99.674 | 99.968 | 99.658 | 99.984 | 99.61 | 100 |
| 0.5 | 0.90 | 87.329 | 99.543 | 87.071 | 99.881 | 86.979 | 99.952 | 92.510 | 99.579 | 92.251 | 99.777 | 92.132 | 99.932 |
| | 0.95 | 86.486 | 99.360 | 86.168 | 99.773 | 85.985 | 99.948 | 91.846 | 99.535 | 91.543 | 99.781 | 91.412 | 99.889 |
| 1 | 0.90 | 86.363 | 99.610 | 86.041 | 99.825 | 85.922 | 99.988 | 91.635 | 99.837 | 91.460 | 99.964 | 91.309 | 99.960 |
| | 0.95 | 85.234 | 99.431 | 84.912 | 99.750 | 84.741 | 99.956 | 90.879 | 99.797 | 90.565 | 99.889 | 90.430 | 99.976 |

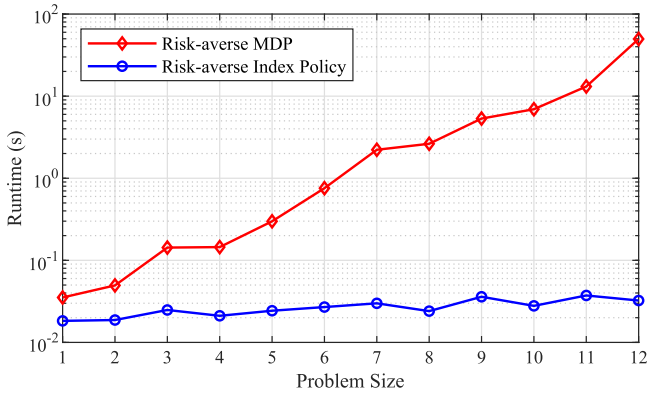


Fig. 1. Runtime of obtaining the optimal policy and index policy for first-order mean-semideviation.

problem sizes for first-order mean-semideviation. These problem sizes start from problem size 1 where we have two arms with three states each and ends with problem size 12 where we have three arms with five states each. The results are presented in Fig. 1 where the CPU time is in logarithmic scale to provide better comparison. It is seen that the computation of our risk-averse index heuristic takes significantly less time than that of finding optimal policy. By increasing the problem size, computation time of our risk-averse indices seems to grow linearly while the computation time of solving the risk-averse MDP optimally shows an exponential growth. The computation time of the same instances for mean-AVaR is longer but provides a similar pattern.

We further remark that an optimal solution of risk-averse MAB is not always priority-index based, and in first-order mean-semideviation and mean-AVaR risk measures, in 0.175% and 0.042% of cases the optimal policy could not be represented as an index policy, respectively.

## VII. CONCLUSION

We introduce a new approach in incorporating risk into MAB problem by using dynamic coherent measures of risk and propose a priority-index heuristic, analogous to the structure of Gittins index. The experiments we conduct to test the performance of our index policy indicate the excellence of the policy. Compared to the policy derived from Gittins index, the proposed index policy provides much lower suboptimality percentages and it is more similar to the optimal policy. Our experiments comparing the computational complexity of our index heuristic to solving the corresponding MDP reveal that our method is time efficient and the computation time grows linearly with the problem size, whereas the risk-averse MDP computation time shows an exponential growth.

As future research, one direction would be to perform a regret analysis of the RAI heuristic for this MAB problem, which would further solidify the current computational results indicating that the RAI provides near-optimal policies. The other direction would be to extend this framework to the restless setting or to the settings with imperfect information.

## REFERENCES

[1] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Stat. Soc. Ser. B*, vol. 41, no. 2, pp. 148–177, 1979.

[2] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probability*, vol. 25, no. A, pp. 287–298, 1988.
[3] X. Huo and F. Fu, "Risk-aware multi-armed bandit problem with application to portfolio selection," *Roy. Soc. Open Sci.*, vol. 4, no. 11, 2017, Art. no. 171377.
[4] N. Galichet, M. Sebag, and O. Teytaud, "Exploration vs exploitation vs safety: Risk-aware multi-armed bandits," in *Proc. ACML*, 2013, pp. 245–260.
[5] J.-P. Chancelier, M. De Lara, and A. De Palma, "Risk aversion, road choice, and the one-armed bandit problem," *Transp. Sci.*, vol. 41, no. 1, pp. 1–14, 2007.
[6] Ü. Şahin, V. Yücesoy, A. Koç, and C. Tekin, "Risk-averse ambulance redeployment via multi-armed bandits," in *Proc. IEEE 26th Signal Process. Commun. Appl. Conf.*, 2018, pp. 1–4.
[7] A. Sani, A. Lazaric, and R. Munos, "Risk-aversion in multi-armed bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 3275–3283.
[8] E. V. Denardo, H. Park, and U. G. Rothblum, "Risk-sensitive and risk-neutral multiarmed bandits," *Math. Oper. Res.*, vol. 32, no. 2, pp. 374–394, 2007.
[9] E. V. Denardo, E. A. Feinberg, and U. G. Rothblum, "The multi-armed bandit, with constraints," *Ann. Oper. Res.*, vol. 208, no. 1, pp. 37–62, 2013.
[10] A. Shapiro, D. Dentcheva, and A. Ruszczyński, Lectures on Stochastic Programming: Modeling and Theory. Philadelphia, PA, USA: SIAM, 2009.
[11] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Math. Finance*, vol. 9, no. 3, pp. 203–228, 1999.
[12] A. Ruszczyński, "Risk-averse dynamic programming for Markov decision processes," *Math. Program.*, vol. 125, no. 2, pp. 235–261, 2010.
[13] F. Riedel, "Dynamic coherent risk measures," *Stochastic Processes Appl.*, vol. 112, no. 2, pp. 185–200, 2004.
[14] A. Ruszczyński and A. Shapiro, "Conditional risk mappings," *Math. Oper. Res.*, vol. 31, no. 3, pp. 544–561, 2006.
[15] W. Ogryczak and A. Ruszczyński, "From stochastic dominance to mean-risk models: Semideviations as risk measures," *Eur. J. Oper. Res.*, vol. 116, no. 1, pp. 33–50, 1999.
[16] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *J. Banking Finance*, vol. 26, no. 7, pp. 1443–1471, 2002.
[17] K. D. Glazebrook and R. Minty, "A generalized Gittins index for a class of multiarmed bandits with general resource requirements," *Math. Oper. Res.*, vol. 34, no. 1, pp. 26–44, 2009.
[18] M. Woodroofe, "A one-armed bandit problem with a concomitant variable," *J. Amer. Stat. Assoc.*, vol. 74, no. 368, pp. 799–806, 1979.
[19] J.-P. Chancelier, M. De Lara, and A. de Palma, "Risk aversion in expected intertemporal discounted utilities bandit problems," *Theory Decis.*, vol. 67, no. 4, pp. 433–440, 2009.
[20] P. Varaiya, J. Walrand, and C. Buyukkoc, "Extensions of the multiarmed bandit problem: The discounted case," *IEEE Trans. Autom. Control*, vol. AC-30, no. 5, pp. 426–439, May 1985.
[21] Ö. Çavuş and A. Ruszczyński, "Computational methods for risk-averse undiscounted transient Markov models," *Oper. Res.*, vol. 62, no. 2, pp. 401–417, 2014.
[22] V. Kuleshov and D. Precup, "Algorithms for multi-armed bandit problems," 2014, *arXiv:1402.6028*.