Optimizing sequential decision-making under risk: Strategic allocation with switching penalties

Citation for the original published paper (version of record):
Malekipirbazari, M. (2025). Optimizing sequential decision-making under risk: Strategic allocation with switching penalties. European Journal of Operational Research, 321(1): 160-176.
http://dx.doi.org/10.1016/j.ejor.2024.09.023

(article starts on next page)

European Journal of Operational Research 321 (2025) 160–176

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/eor

Stochastics and statistics

Optimizing sequential decision-making under risk: Strategic allocation with switching penalties

Milad Malekipirbazari

Department of Computer Science & Engineering, Chalmers University of Technology, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Stochastic programming
Multiarmed bandit problem
Switching penalties
Risk-averse decision-making
Dynamic coherent risk measures

ABSTRACT

This paper considers the multiarmed bandit (MAB) problem augmented with a critical real-world consideration: the cost implications of switching decisions. Our work distinguishes itself by addressing the largely unexplored domain of risk-averse MAB problems compounded by switching penalties. Such scenarios are not just theoretical constructs but are reflective of numerous practical applications. Our contribution is threefold: firstly, we explore how switching costs and risk aversion influence decision-making in MAB problems. Secondly, we present novel theoretical results, including the development of the Risk-Averse Switching Index (RASI), which addresses the dual challenges of risk aversion and switching costs, demonstrating its near-optimal efficacy. This heuristic solution method is grounded in dynamic coherent risk measures, enabling a time-consistent evaluation of risk and reward. Lastly, through rigorous numerical experiments, we validate our algorithm's effectiveness and practical applicability, providing decision-makers with valuable insights and tools for navigating the multifaceted landscape of risk-averse environments with inherent switching costs.

1. Introduction

The multiarmed bandit (MAB) problem, originating in the seminal works of Robbins (1952), stands as a cornerstone in decision theory and operations research. It represents scenarios where an agent must choose from several options, each with its own reward structure. Initially devised to tackle sequential trials under uncertainty, the MAB problem has since evolved, mirroring the complexities of decision-making across diverse fields such as technology (Kumar & Saranga, 2010), healthcare (Villar et al., 2015), and finance (Bertsimas & Mersereau, 2007). According to Powell (2019), "the principles of bandit problems, long a niche community, should become a core dimension of mainstream stochastic optimization". This perspective underscores the growing importance and applicability of MAB frameworks in broader stochastic optimization contexts.

In addressing computational challenges associated with the MAB problem, particularly as the number of options – or "arms" – increases, Gittins and Jones (1974) introduced an index-based strategy. This method computes indices for each arm independently, using arm-specific data, and selects the arm with the highest "Gittins index" value at each step. The efficiency of this approach for large-scale MAB problems has been extensively discussed in the literature, including Weber (1992) and Gittins (1979), highlighting its computational advantages and limitations.

However, real-world scenarios often deviate from the idealized conditions of classical MAB models. A significant deviation is the presence of switching penalties, encompassing monetary, reputational, or operational costs incurred when changing from one option to another. These penalties add a layer of complexity to the MAB problem, affecting the performance of traditional bandit algorithms. Concurrently, the integration of risk considerations into decision-making, a concept rooted in the foundational work of Markowitz (1952) on portfolio theory, has prompted a shift from the traditional risk-neutral focus on maximizing expected rewards to a more detailed approach that also considers reward variability. This shift is particularly pertinent in domains like finance and operations research, where the stakes of decision-making under uncertainty are high.

Despite these advancements, the literature reveals a gap in models that comprehensively address both switching penalties and risk aversion within the MAB framework. This paper aims to bridge this gap and offer a more realistic and applicable model for decision-making in today's increasingly complex and uncertain world.

1.1. Motivating examples

To further elucidate the practical significance of combining risk-aversion and switching penalties in MAB problems, consider the following real-life examples:

E-mail address: miladma@chalmers.se.

https://doi.org/10.1016/j.ejor.2024.09.023
Received 25 March 2024; Accepted 11 September 2024
Available online 20 September 2024

Portfolio management in finance: In financial portfolio management, a risk-averse investor faces switching costs when reallocating assets within their portfolio. These costs include transaction fees and potential tax implications. Additionally, the investor must consider the risk associated with different asset classes. An optimal strategy would balance the risk of asset volatility with the costs of reallocating assets to maximize long-term returns.

Energy resource allocation: In energy resource allocation, the sector is moving toward using renewable energy sources. A utility company must decide how to distribute investments across various energy sources, such as fossil fuels, solar, and wind. Switching investments (e.g., from fossil fuels to renewable sources) involves significant costs, including infrastructure changes and workforce retraining. Moreover, the company must consider the risk factors associated with each energy source, such as regulatory changes or varying market demands. An effective strategy would minimize switching costs while managing the risks associated with each energy source's future viability and profitability.

E-commerce platform advertising: An e-commerce platform using online advertising must frequently decide which products to promote. Switching the focus from one product line to another incurs costs, including market research and new ad campaign development. Additionally, there is a risk associated with focusing on a new product line, such as uncertain consumer demand. The platform needs a strategy that judiciously balances the frequency of switching ad focus with the risk associated with new or untested products.

Clinical trials for new treatments: In clinical trials, particularly in drug development, researchers must allocate resources across multiple potential treatments. Each treatment represents an "arm" in the multi-armed bandit problem. Switching from one treatment to another incurs significant penalties, including the logistical costs of setting up new trials and ethical considerations. Concurrently, there is a high degree of risk involved, as each treatment carries its own efficacy and side effect profiles. A risk-averse strategy is crucial due to the implications for patient health and safety. The challenge lies in minimizing switching between treatments while managing the risks associated with each, aiming to identify the most promising treatment efficiently and safely.

These examples demonstrate the real-world relevance of studying MAB problems that incorporate both risk-aversion and switching penalties. Our research aims to address the gap in current methodologies by proposing strategies that consider both these factors, offering more realistic and applicable solutions for decision-makers in various fields.

1.2. Related work

Our research lies in the Markovian MAB literature, which we explore from two distinct perspectives: MAB problems with switching penalties and risk-averse MAB problems.

1.2.1. MAB problem with switching penalties

The MAB problem with switching costs is a complex extension of the classical MAB framework, which traditionally did not account for this cost. The literature on this problem evolved significantly, moving from foundational theoretical work to more complex, application-driven research. This evolution reflects a growing recognition of the complexities inherent in real-world decision-making scenarios and the need for sophisticated, adaptable strategies. A comprehensive survey by Jun (2004) explores theoretical foundations, algorithmic advancements, and practical applications in MAB problems with switching costs. This survey provides an extensive resource for understanding the impact of switching penalties and the strategies developed to address them.

The initial acknowledgment of switching costs in the MAB framework can be traced back to works that identified the limitations of traditional approaches in practical scenarios. For instance, Whittle (1988) introduced the concept of "restless bandits", which laid the groundwork for considering scenarios where the state of unchosen options (arms) could change, indirectly hinting at the potential costs of not switching.

The seminal paper by Banks and Sundaram (1994) was among the first to explicitly address the challenge of switching costs in MAB problems. They demonstrated that the introduction of switching costs renders traditional index policies, such as the Gittins index, suboptimal. This work was crucial in steering the focus of MAB research towards developing strategies that could incorporate these additional costs. Following this work, several researchers explored the complexities introduced by switching costs. For instance, Asawa and Teneketzis (1994) examined optimal switching times in the presence of these costs, proposing a modified Gittins index to account for them. Their work highlighted the intricacies involved in calculating the optimal time for transitioning between arms, especially when each switch incurs a penalty.

Subsequent research has further developed computational methods for these indices. Niño-Mora (2008) introduced a faster index algorithm, significantly improving the computational efficiency for bandits with switching costs. Additionally, Niño-Mora (2010) presented a comprehensive study on computing an index policy for such bandits, offering valuable insights into the practical implementation of these strategies.

The impact of switching costs has been examined across various domains. For example, in queuing systems, Van Oyen et al. (1992) analyzed optimal scheduling policies for parallel queues without arrivals, a special case of MAB with switching costs. Washburn (2008) applied MAB to sensor management, incorporating switching delays in mechanically pointed sensors. Similarly, Caro and Gallien (2007) investigated dynamic assortment optimization for seasonal consumer goods using a finite horizon multiarmed bandit model. They addressed how retail firms should modify product assortments over time to maximize profits, while considering implementation delays, switching costs, and demand substitution effects.

Recent advancements in the domains of regret minimization and pure exploration have also incorporated switching costs into the MAB framework. With regard to regret minimization, algorithms have been designed to balance the trade-off between exploration and exploitation while minimizing the cumulative regret and the costs associated with switching between arms. For instance, Rouyer et al. (2021) and Amir et al. (2022) propose novel strategies that effectively manage switching costs in both stochastic and adversarial settings. In the domain of pure exploration, recent work by Mwai et al. (2024) introduces a batched bandit algorithm for fixed-confidence pure exploration with constraints on the frequency of arm switching. Their algorithm demonstrates that it is possible to achieve quick stopping times while respecting the strict switching limits, providing an efficient approach for scenarios where switching costs are a significant concern.

1.2.2. Risk-averse MAB problem

The integration of risk considerations into MAB problems, despite its relevance in numerous applications, remains a relatively underexplored area in the literature. Denardo et al. (2007) pioneered the incorporation of risk in a Markovian context, utilizing concave utility functions. They introduced a novel state ranking system and established its optimality in selecting the highest-ranking arm. Their comprehensive study spans three distinct models: one with a risk-averse exponential utility, another with a risk-seeking exponential utility, and a third employing a linear utility. Further enhancements to this methodology are discussed in Denardo et al. (2013)'s subsequent work. Chancelier et al. (2007) tackled risk aversion in a Markovian setting by examining the choice between a random and a safe route in various information regimes, framing it as a one-armed bandit problem. Similar to Denardo et al. (2007), their approach hinges on utility functions to integrate risk preferences into the model. However, these methods, centered on utility functions for risk incorporation, are limited by the necessity for decision-makers to explicitly define suitable utility functions, which can

be challenging and may lead to solutions that are difficult to interpret, as noted by Shapiro et al. (2009).

In contrast, more recent methodologies employing coherent risk measures overcome the complexities associated with utility functions (Artzner et al., 1999; Grechuk & Zabarankin, 2016). In this vein, Cohen and Treetanthiploet (2019) innovatively adapted the Gittins index into a nonlinear operator using coherent risk measures. They revisited the Gittins index theorem, known as the prevailing charge formulation by Weber (1992), in a generic discrete-time framework, moving away from the Markov assumption to embrace the approach of El Karoui and Karatzas (1994). Their method, which finely balances exploration and exploitation in decision-making, formulates an optimal stopping problem under a nonlinear expectation. They also showed that Gittins indices can provide optimal solutions in scenarios where arms are strongly independent, albeit with a relaxed definition of optimality.

Building on these developments, Malekipirbazari and Çavuş (2021, 2024) designed a new MAB framework based on dynamic coherent risk measures, inspired by the risk-averse discrete-time Markov models for discounted infinite horizon problems by Ruszczyński (2010). Malekipirbazari and Çavuş (2021) used Lagrangian duality theory to decompose the problem and introduced a priority-index heuristic. Malekipirbazari and Çavuş (2024) established a theoretical foundation based on Whittle's retirement problem (Whittle, 1980) and proposed a different index-based policy.

As discussed, coherent risk measures provide a more robust framework for risk-averse optimization compared to utility functions, particularly due to their desirable properties like coherence and convexity. However, we acknowledge that specifying the parameters of these risk measures, including the weighting parameter, can be challenging for practitioners. This difficulty is due to the lack of straightforward interpretation and the need for careful calibration based on the specific decision-making context.

1.3. Our contributions

This paper pioneers the integration of switching penalties and risk aversion within the MAB framework, marking a significant leap towards more realistic decision-making models. Our contributions are threefold, each addressing a critical gap in the existing literature and offering both theoretical insights and practical solutions:

1. **Qualitative analysis of decision-making strategies**: We explore how switching costs and risk aversion influence decision-making in MAB problems. Our work here discusses the limitations of existing strategies under these conditions by highlighting the complexities introduced by these factors.
2. **Development of novel strategies – The RASI policy**: We introduce the RASI policy, a heuristic method based on dynamic coherent risk measures. This strategy is specifically designed to navigate the complexities introduced by risk aversion and switching costs, offering a systematic approach to arm selection that balances the dual considerations of risk and penalty for switching.
3. **Insights from numerical experiments**: Through rigorous numerical experiments, we validate the effectiveness of the RASI policy, showcasing its superiority over existing strategies. Our findings reveal that the RASI policy outperforms traditional risk-neutral policies by achieving an average optimality percentage close to 99% in environments characterized by high switching costs and risk aversion. These insights underscore the practical applicability of our approach, providing decision-makers with a robust tool for enhancing resource allocation in risk-averse settings.

1.4. Outline of paper

The paper is structured as follows: Section 1 introduces the MAB problem, highlighting the integration of risk aversion and switching penalties. Section 2 details the problem formulation, presenting the risk-averse MAB problem with switching costs as a Markov decision process. Section 3 analyzes the impact of switching costs in a risk-averse setting, including theoretical insights into model equivalence and the non-existence of optimal index policies. Section 4 introduces a novel index-based heuristic and discusses its application in various MAB environments, along with computational aspects and behavior in restricted settings. Section 5 presents computational experiments to assess the efficiency of the proposed index-based policy, focusing on suboptimality and optimality percentages. Section 6 concludes the paper with a summary of key findings and future research directions.

2. Problem formulation

In this section, we study the intricate problem of formulating a risk-averse multiarmed bandit framework that incorporates switching penalties. We begin by describing the MAB problem as a Markov decision process (MDP), highlighting the key aspects of our approach. This section is structured to first lay the groundwork with a detailed problem description, followed by an exploration of the fundamental concepts of dynamic coherent risk measures. We then proceed with a comprehensive formulation of the risk-averse MAB problem with switching penalties. This formulation not only captures the complexity of real-world decision-making scenarios but also sets the foundation for developing practical, heuristic strategies that balance computational feasibility with the intricate demands of risk-averse environments.

2.1. Problem description

The expected total discounted reward is targeted to be maximized in the classical risk-neutral bandits. More specifically, we may describe MAB as an MDP, in which a decision maker chooses which arm to play from a pool of K potential options at each decision step $t \in \mathbb{N}$, with $\mathbb{N} = \{1, 2, 3, \dots\}$. In this work, a risk-averse MAB problem with dynamic coherent risk measures is taken into consideration. We consider minimizing negative rewards, which may be interpreted as costs, for the sake of mathematical convenience.

The following is a description of our problem setting:

(i) Let each arm i forms a Markov chain with a finite state space $\mathcal{X}^i$, $i \in \mathcal{K}$, and $\mathcal{K} = \{1, 2, \dots, K\}$. Also let $\mathcal{X} = \bigtimes_{i=1}^{K} \mathcal{X}^i$ be the state space of the resultant MDP, with the assumption that there is no shared state in different arms.

(ii) Let $\mathcal{U}$ be a finite action space and $U(x) \subseteq \mathcal{U}$ a nonempty set of admissible actions at each state $x \in \mathcal{X}$. Given the current state $x_t \in \mathcal{X}$ at each step $t \in \mathbb{N}$, we execute an action $u_t = (u_t^1, u_t^2, \dots, u_t^K) \in U(x_t)$. Here u_t^i is the action applied to arm $i \in \mathcal{K}$, where, at step t, $u_t^i = 1$ denotes the activation of arm i and $u_t^i = 0$ denotes the absence of arm i from play. Actions that result in exactly one element of u_t equaling one are considered admissible actions.

(iii) The state of an activated arm changes at each step in a Markovian manner in accordance with the transition probabilities $Q^i(x, 1, y) := \mathbb{P}(x_{t+1}^i = y \mid x_t^i = x,\ u_t^i = 1)$, $i \in \mathcal{K}$, $x, y \in \mathcal{X}^i$, $t \in \mathbb{N}$, where x_t^i represents the state of arm i at step t. On the other hand, the state of a non-play arm stays the same, therefore for each $i \in \mathcal{K}$, $x, y \in \mathcal{X}^i$, $t \in \mathbb{N}$ we have $Q^i(x, 0, y) := \mathbb{P}(x_{t+1}^i = y \mid x_t^i = x,\ u_t^i = 0) = 1$ if $y = x$, 0 otherwise.

(iv) The cost incurred by playing arm i and its transition to the next state is represented by the cost function g^i, which is specified to be finite and non-positive for each arm $i \in \mathcal{K}$. Let us denote $c^i(x^i, u^i, y^i)$ as the sum of the costs associated with playing arm i and changing from the state $x^i \in \mathcal{X}^i$ to state $y^i \in \mathcal{X}^i$ under the action $u^i \in \{0, 1\}$. That is

$$c^i(x^i, u^i, y^i) := \begin{cases} g^i(x^i, y^i) & \text{if } u^i = 1, \\ 0 & \text{if } u^i = 0. \end{cases}$$

Also $c(x,u,y) = \sum_{i\in\mathcal{K}} c^i(x^i,u^i,y^i)$, $x = (x^1,\dots,x^K) \in \mathcal{X}$, $y = (y^1,\dots,y^K) \in \mathcal{X}$, and $u = (u^1,\dots,u^K) \in U(x)$.

(v) The switching costs, incurred by switching from one option (or arm) to another, are denoted as $s = (s^1, s^2, \dots, s^K)$. Each s^i represents the cost incurred when transitioning to arm $i \in \mathcal{K}$. These costs could be financial, time-based, or resource-oriented, reflecting the real-world implications of changing strategies or actions. In our model, these costs are finite and positive.

(vi) $\Pi = (\pi,\dots,\pi)$ denote a stationary Markov policy, with $\pi : \mathcal{X} \to \mathcal{U}$ as an action rule. Note that, we have $u_t = \pi(x_t)$, which means both represent the action to be conducted at state $x_t \in \mathcal{X}$ for $t \in \mathbb{N}$. We also define $\pi^i : \mathcal{X}^i \to \{0,1\}$ which indicates the decision to be taken for arm i in a given state.

In this work, the decision-maker must navigate not only the potential rewards and risks associated with each arm of the MAB problem but also the frequency and financial implications of switching between these options. These switching penalties are particularly pertinent in dynamic environments where adaptability is key, yet they introduce significant trade-offs. We model risk aversion through dynamic coherent risk measures and seek a stationary Markov policy that minimizes the total risk-averse discounted costs over an infinite horizon. As Ruszczyński (2010) have shown, such a policy exists for infinite horizon stationary MDPs with dynamic coherent risk measures and can be identified using value iteration or policy iteration algorithms. However, these methods can be computationally intensive and impractical for complex problems, and the resulting policy structures can be challenging to interpret in real-world scenarios. To overcome these hurdles, our research is directed towards developing suitable index-based heuristic strategies. These strategies aim to strike a balance between computational feasibility and interpretability, effectively addressing the intertwined challenges of risk aversion and switching costs in decision-making processes.

2.2. Preliminaries on dynamic coherent risk measures

We first describe dynamic risk measures and their characteristic before introducing our model. Take into account a probability space $(\Omega, \mathcal{F}, P)$, a filtration $\{\emptyset, \Omega\} = \mathcal{F}_1 \subset \dots \subset \mathcal{F}_{T+1} \subset \mathcal{F}$, and an adapted sequence of random variables $Z_t = c(x_{t-1}, u_{t-1}, x_t) \in \mathcal{Z}_t$, $t \in \{2,\dots,T+1\}$. Define the spaces $\mathcal{Z}_t$ of $\mathcal{F}_t$-measurable random variables on Ω, $t \in \{1,\dots,T+1\}$ and $\mathcal{Z}_{1,T+1} = \mathcal{Z}_1 \times \dots \times \mathcal{Z}_{T+1}$, where $\mathcal{Z}_1 = \mathbb{R}$. Given that each $\mathcal{F}_t$ in our case is finite, it is possible to specify the spaces $\mathcal{Z}_t$ with finite dimensional vector spaces.

Before giving a formal explanation of dynamic risk measures, we first define the one-step conditional risk measures that serve as its building blocks: One-step conditional risk measure, $\rho_t : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$, $t \in \{1,\dots,T\}$, meets the following axioms (for details see Riedel (2004) and Ruszczyński and Shapiro (2006a)):

(A1) $\rho_t(\delta Z + (1-\delta)W) \le \delta\rho_t(Z) + (1-\delta)\rho_t(W)$, $\forall \delta \in (0,1)$, $Z, W \in \mathcal{Z}_{t+1}$;
(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) $\rho_t(\delta Z) = \delta\rho_t(Z)$, $\forall Z \in \mathcal{Z}_{t+1}, \delta \ge 0$.

These axioms, known as convexity (A1), monotonicity (A2), translation invariance (A3), and positive homogeneity (A4), align with the principles of coherent risk measures as proposed by Artzner et al. (1999). Convexity, or Axiom (A1), upholds the diversification principle, suggesting that the risk of a mixed portfolio cannot exceed the weighted average risk of its components. Monotonicity (A2) posits that a portfolio consistently yielding better outcomes than another should not be deemed riskier, ensuring that risk assessment is aligned with intuitive expectations of portfolio performance. Translation invariance (A3) states that adding a certain amount to all outcomes of a portfolio uniformly adjusts its risk measure by the same amount, reflecting the direct impact of guaranteed gains or losses on the portfolio's risk profile. Lastly, positive homogeneity (A4) implies that scaling the outcomes of a portfolio by a positive factor proportionally affects its risk measure, emphasizing the direct correlation between the magnitude of investment actions and associated risk levels.

One popular and frequently applied coherent risk measure is first-order mean-semideviation. Here, we provide the conditional version of this risk measure (Ogryczak & Ruszczyński, 1999, 2001, 2002; Ruszczyński & Shapiro, 2006a, 2006b). The definition of the conditional first-order mean-semideviation risk measure is

$$\rho_t(Z_{t+1}) = \mathbb{E}[Z_{t+1}|\mathcal{F}_t] + \kappa\mathbb{E}[(Z_{t+1} - \mathbb{E}[Z_{t+1}|\mathcal{F}_t])_+|\mathcal{F}_t], \tag{1}$$

where $\kappa \in [0,1]$ and $(a)_+ := \max\{a, 0\}$ for $a \in \mathbb{R}$.

A dynamic risk measure is defined as a sequence of one-step conditional risk measures (Artzner et al., 2007; Calafiore & Dabbene, 2006; Cheridito et al., 2006; Föllmer & Penner, 2006; Ruszczyński, 2010). The dynamic risk measure $\varrho_{1,T} : \mathcal{Z}_{1,T+1} \mapsto \mathcal{Z}_1$ on the finite horizon with length T can be presented as:

$$\varrho_{1,T}(Z_2, Z_3, \dots, Z_{T+1}) := \rho_1\big(Z_2 + \rho_2\big(Z_3 + \rho_3\big(Z_4 + \dots + \rho_T\big(Z_{T+1}\big)\dots\big)\big)\big) \tag{2}$$

and in the discounted scenario with discount factor $\beta \in (0,1)$, it is given as:

$$\begin{aligned}\varrho^{\beta}_{1,T}(Z_2, \dots, Z_{T+1}) := \rho_1\big(Z_2 + \rho_2\big(\beta Z_3 + \rho_3\big(\beta^2 Z_4 + \dots \\ + \ \rho_T\big(\beta^{T-1} Z_{T+1}\big)\dots\big)\big)\big)\end{aligned} \tag{3}$$

Accordingly, the dynamic risk measure for an infinite horizon can be described as:

$$\varrho(Z_2, Z_3, Z_4, \dots) := \lim_{T\to\infty} \varrho_{1,T}(Z_2, Z_3, \dots, Z_{T+1}) \tag{4}$$

$$\varrho^{\beta}(Z_2, Z_3, Z_4, \dots) := \lim_{T\to\infty} \varrho^{\beta}_{1,T}(Z_2, Z_3, \dots, Z_{T+1}) \tag{5}$$

To clarify the concept of dynamic risk measures, the following example demonstrates the application of the first-order mean-semideviation as a dynamic risk measure over two periods. This example is particularly instructive as it concretely illustrates how risk assessments can evolve over time, capturing the essence of dynamic risk management.

Example 1 (*Malekipirbazari & Çavuş, 2024*). For dynamic risk measure of first-order mean-semideviation with two periods and a discount factor of β, we obtain

$$\begin{aligned}\varrho^{\beta}_{1,2}(Z_2, Z_3) &= \rho_1(Z_2 + \rho_2(\beta Z_3)) \\ &= \mathbb{E}\big[Z_2 + \rho_2(\beta Z_3)|\mathcal{F}_1\big] + \kappa\mathbb{E}\Big[\Big(\ Z_2 + \rho_2(\beta Z_3) \\ &\quad - \ \mathbb{E}\ \big[\ Z_2 + \rho_2(\beta Z_3)|\mathcal{F}_1\ \big]\Big)_+\Big|\mathcal{F}_1\ \Big],\end{aligned}$$

where $\rho_2(\beta Z_3) = \mathbb{E}[\beta Z_3|\mathcal{F}_2] + \kappa\mathbb{E}\big[\big(\beta Z_3 - \mathbb{E}[\beta Z_3|\mathcal{F}_2]\big)_+|\mathcal{F}_2\big]$.

2.3. Problem formulation for risk-averse MAB with switching penalties

Now, we explore the mathematical formulation of the risk-averse multiarmed bandit problem with switching penalties. Our focus is on assessing the risk of cost sequences under a given policy Π and initial state x_1, incorporating the dynamic risk measures as outlined in (5). The formulation is as follows:

$$\begin{aligned}R^{\Pi}(x_1) = \varrho^{\beta}\Big(c(x_1,u_1,x_2) + s^T u_1, c(x_2,u_2,x_3) + s^T u_2 \mathbb{1}_{u_2\neq u_1}, c(x_3,u_3,x_4) \\ + \ s^T u_3 \mathbb{1}_{u_3\neq u_2}, \dots\Big).\end{aligned} \tag{6}$$

The term $s^T u_t$ represents the inner product of two vectors, s and u_t, and is crucial for calculating the switching costs in our model. As previously mentioned, the vector u_t has a single non-zero element indicating

the arm currently in play. Consequently, $s^T u_t$ effectively yields the switching cost associated with the arm to which we are transitioning. It is important to note that this switching cost becomes relevant only when there is a change in the arm being played, as indicated by the condition $\mathbb{1}_{u_t \neq u_{t-1}} = 1$. In such cases, the cost $s^T u_t$ is incurred due to the transition. This mechanism ensures that switching costs are accounted for in the total risk-averse cost calculation only when a change in the active arm occurs, aligning with the realistic scenarios where switching decisions carry financial or operational implications.

Considering $R(x_1)$ as the optimal risk-averse total discounted cost with switching penalty starting from state x_1, and denoting $\mathbf{\Pi}$ as the class of stationary admissible policies for our problem, the objective is formalized as follows:

$$R(x_1) = \min_{\Pi \in \mathbf{\Pi}} R^{\Pi}(x_1). \tag{7}$$

In the context of switching costs, the state of the problem at any given time step t cannot be adequately captured by the state vector x_t alone, except at the initial step ($t = 1$). This is because the decision-making process is influenced not just by the current state but also by the history of actions, specifically which arm was last played. Therefore, to accurately account for this aspect, we extend our formulation to include the identity of the immediately played arm. Accordingly, the risk of cost sequences for a policy Π at time t, considering arm $i \in \mathcal{K}$ as the last played arm, is expressed as:

$$\begin{aligned} R^{\Pi}(x_t, i) = \varrho^{\beta} \big(& c(x_t, u_t, x_{t+1}) + s^T u_t \mathbb{1}_{u_t^i \neq 1}, c(x_{t+1}, u_{t+1}, x_{t+2}) \\ & + s^T u_{t+1} \mathbb{1}_{u_{t+1} \neq u_{t+2}}, \dots \big) . \end{aligned} \tag{8}$$

In this extended formulation, $R(x_t, i)$ represents the optimal risk-averse total discounted cost when starting from state x_t with arm i as the last played arm, incorporating the positive switching penalty. The optimization problem thus becomes:

$$R(x_t, i) = \min_{\Pi \in \mathbf{\Pi}} R^{\Pi}(x_t, i). \tag{9}$$

This formulation captures the essence of decision-making under risk and switching penalties, reflecting scenarios where each decision's consequences unfold over time. However, solving the optimization problems represented by (7) and (9), using risk-averse dynamic programming approaches introduced in Ruszczyński (2010), becomes increasingly impractical for larger-scale problems. This impracticality arises from the exponential growth in computational complexity. Moreover, while index policies are commonly employed for optimizing MAB problems, their effectiveness in certain contexts, including ours, can be limited. As highlighted by Banks and Sundaram (1994), there are scenarios in risk-neutral MAB problems with positive switching costs where optimal solutions are not index-based. This realization diminishes the incentive to seek optimal solutions, especially when considering the interpretability and practical applicability of these solutions in complex scenarios. In light of this, our approach aims to develop index-based heuristic strategies. These strategies are designed to offer a pragmatic balance, providing solutions that are not only computationally feasible but also maintain a level of interpretability and practical relevance. This involves creating a generalized form of the problem and developing calibrating functions for each arm. By focusing on one arm at a time and devising appropriate calibrating functions, we aim to simplify the problem and develop a solution that balances the need for practicality with the desire for theoretical rigor.

This study also contemplates a variant of the MAB problem that incorporates switching delays, an aspect equally pivotal as switching costs. The formulation and implications of MAB under switching delays are explored in Section 4.7. This additional discussion complements our primary focus on switching costs, providing a broader perspective on the dynamics of decision-making in risk-averse environments.

3. The dynamics of switching costs in risk-averse bandit models

Our analysis in this section is twofold: firstly, we aim to validate whether our specific approach to modeling switching costs is sufficiently general to encapsulate a broader range of scenarios. Secondly, we endeavor to investigate the existence (or lack thereof) of optimal index policies in such risk-averse MAB problems. This exploration is crucial for understanding the limitations and potential of index-based strategies in environments where both risk aversion and the costs of switching strategies play important roles.

3.1. Generalization of switching costs in a risk-averse framework

In this section, we examine the impact of incorporating switching costs within a risk-averse MAB setting. A key aspect of our analysis involves simplifying the model to make it more tractable without losing generality. Following the approach of Banks and Sundaram (1994), we adopt the assumption that any bandit problem involving costs for switching both "away from" and "to" an arm can be effectively transformed into an equivalent problem with only the cost of switching to an arm. This assumption is crucial for streamlining our analysis, and it raises an important question: Is considering a bandit problem with only a single type of switching cost comprehensive enough to capture the effects of switching in our risk-averse model?

To address this, Banks and Sundaram (1994) demonstrated that in the risk-neutral context, MAB problems with dual switching costs (both "from" and "to" an arm) can be equivalently represented by problems with only "to" switching costs. We extend this concept to the risk-averse setting, hypothesizing that a similar equivalence holds. The following theorem is critical in our discussion, asserting that every risk-averse bandit problem with costs for both switching away from and to an arm can be equivalently modeled as a problem with costs incurred only for switching to an arm.

Theorem 1. *Every risk-averse bandit problem in which switching away from and switching to an arm incurs costs has an equivalent risk-averse bandit problem in which only switching to an arm incurs costs.*

Proof. Consider two different risk-averse bandits with switching costs that possess the same state space and action space, thus the same policy space. Let their transition probabilities also be the same, which means in both problems, the same distribution is induced by a given policy on infinite histories. In the first bandit, the change in play for arm i includes both "switching to" cost $s_i^{(1)}$ and "switching from" cost $d_i^{(1)}$, yet in the second bandit only "switching to" an arm is costly with the cost of $s_i^{(2)} = s_i^{(1)} + d_i^{(1)}$. Moreover, let the cost of transition in arm i at time t be $c^i(x_t^i, 1, x_{t+1}^i)$ and $c^i(x_t^i, 1, x_{t+1}^i) - (1-\beta)d_i^{(1)}$ for the first and the second bandits, respectively. The aim is to show that these two bandits are equivalent. For that, let us consider a finite-horizon problem where the switch to arm i at step t_T is the last arm switch in the play. First, we show that the risk-adjusted total discounted cost for this arm during the periods of its continuous usage is the same in both problems. Suppose arm i is played for k_T consecutive steps prior to termination. The risk-adjusted total discounted value of these k_T periods for the first and the second bandits respectively are:

$$\begin{aligned} R^{(1)}_{\Delta k_T}(x_{t_T}) &= \varrho^{\beta}_{1,k_T+1} \Big(s_i^{(1)} + c^i(x^i_{t_T}, 1, x^i_{t_T+1}), c(x^i_{t_T+1}, 1, x^i_{t_T+2}), \dots, \\ & \quad c(x^i_{t_T+k_T-1}, 1, x^i_{t_T+k_T}), d_i^{(1)} \Big) \\ &= s_i^{(1)} + \beta^{k_T} d_i^{(1)} + \varrho^{\beta}_{1,k_T} \Big(c^i(x^i_{t_T}, 1, x^i_{t_T+1}), c(x^i_{t_T+1}, 1, x^i_{t_T+2}), \dots, \\ & \quad c(x^i_{t_T+k_T-1}, 1, x^i_{t_T+k_T}) \Big), \\ R^{(2)}_{\Delta k_T}(x_{t_T}) &= \varrho^{\beta}_{1,k_T} \Big(s_i^{(1)} + d_i^{(1)} + c^i(x^i_{t_T}, 1, x^i_{t_T+1}) - (1-\beta) d_i^{(1)}, \\ & \quad c(x^i_{t_T+1}, 1, x^i_{t_T+2}) - (1-\beta) d_i^{(1)}, \end{aligned}$$

$$\ldots, c(x^i_{t_T+k_T-1}, 1, x^i_{t_T+k_T}) - (1-\beta)d_i^{(1)}\Big)$$
$$= s_i^{(1)} + d_i^{(1)} - (1-\beta)d_i^{(1)} \sum_{t=0}^{k_T-1} \beta^t + \varrho^{\beta}_{1,k_T}\ \big(c^i(x^i_{t_T}, 1, x^i_{t_T+1}), \ldots,$$
$$c(x^i_{t_T+k_T-1}, 1, x^i_{t_T+k_T})\ \big)$$
$$= s_i^{(1)} + \beta^{k_T} d_i^{(1)} + \varrho^{\beta}_{1,k_T}\Big(c^i(x^i_{t_T}, 1, x^i_{t_T+1}), c(x^i_{t_T+1}, 1, x^i_{t_T+2}), \ldots,$$
$$c(x^i_{t_T+k_T-1}, 1, x^i_{t_T+k_T})\Big).$$

Thus, $R^{(1)}_{\Delta k_T}(x_{t_T}) = R^{(2)}_{\Delta k_T}(x_{t_T})$. Then, suppose at t_{T-1} before the last switch arm j is played for k_{T-1} consecutive steps. The risk-adjusted total discounted value from step t_{T-1} for the first and the second bandits respectively are:

$$R^{(1)}_{\Delta k_{T-1}}(x_{t_{T-1}})$$
$$= \varrho^{\beta}_{1,k_{T-1}+1}\Big(s_j^{(1)} + c^j(x^j_{t_{T-1}}, 1, x^j_{t_{T-1}+1}), \ldots, c(x^j_{t_{T-1}+k_{T-1}-1}, 1, x^j_{t_{T-1}+k_{T-1}}),$$
$$d_j^{(1)} + R^{(1)}_{\Delta k_T}(x_{t_T})\Big)$$
$$= s_j^{(1)} + \beta^{k_{T-1}} d_j^{(1)}$$
$$+ \varrho^{\beta}_{1,k_{T-1}+1}\Big(c^j(x^j_{t_T}, 1, x^j_{t_T+1}), \ldots, c(x^j_{t_T+k_{T-1}-1}, 1, x^j_{t_T+k_{T-1}}), R^{(1)}_{\Delta k_T}(x_{t_T})\Big),$$

$$R^{(2)}_{\Delta k_{T-1}}(x_{t_{T-1}})$$
$$= \varrho^{\beta}_{1,k_{T-1}+1}\Big(s_j^{(1)} + d_j^{(1)} + c^i(x^j_{t_T}, 1, x^j_{t_T+1})$$
$$- (1-\beta)d_j^{(1)}, c(x^j_{t_T+1}, 1, x^j_{t_T+2}) - (1-\beta)d_j^{(1)},$$
$$\ldots, c(x^j_{t_T+k_{T-1}-1}, 1, x^j_{t_T+k_{T-1}}) - (1-\beta)d_j^{(1)}, R^{(2)}_{\Delta k_T}(x_{t_T})\Big)$$
$$= s_j^{(1)} + \beta^{k_{T-1}} d_j^{(1)}$$
$$+ \varrho^{\beta}_{1,k_{T-1}+1}\Big(c^i(x^j_{t_T}, 1, x^j_{t_T+1}), \ldots, c(x^j_{t_T+k_{T-1}-1}, 1, x^j_{t_T+k_{T-1}}), R^{(2)}_{\Delta k_T}(x_{t_T})\Big),$$

implying $R^{(1)}_{\Delta k_{T-1}}(x_{t_{T-1}}) = R^{(2)}_{\Delta k_{T-1}}(x_{t_{T-1}})$. Iterating similarly down to the beginning of the play, the equivalence of the two bandits follows. Finally, when the horizon length goes to infinity, the assertion of the theorem follows. □

The proof of Theorem 1 while conceptually extending the framework established by Banks and Sundaram (1994), encounters unique challenges due to the incorporation of risk aversion into the MAB problem. Specifically, the nonlinear nature of the risk operator used to model risk aversion introduces certain complexity. Unlike in the risk-neutral setting, where costs are aggregated linearly, the risk-averse setting requires careful consideration of how risk is compounded over time and across decisions. This nonlinearity complicates the analysis, as it affects the valuation of future costs and the decision-making process itself.

This theorem not only simplifies our analytical framework but also ensures that our conclusions are broadly applicable, encompassing a wider range of real-world scenarios where switching costs are asymmetric or involve different types of costs for entering and exiting positions.

3.2. Non-existence of optimal index in risk-averse MAB with switching costs

Here, we discuss the feasibility of identifying an optimal index policy in risk-averse bandits that incorporate switching costs. Our objective is to demonstrate that, contrary to certain risk-neutral scenarios, an optimal index policy does not exist in this more complex risk-averse context with switching costs.

Theorem 2. *There is no optimal index for the risk-averse bandits with switching costs.*

The proof of Theorem 2 extends the arguments presented by Banks and Sundaram (1994) to the risk-averse setting, utilizing the first-order mean-semideviation measure with parameter κ to account for risk. Due to its foundational reliance on established results in the neutral case, we omit the detailed proof here, noting that the adaptation follows a similar logical structure with necessary modifications to incorporate risk aversion.

4. Indexability and indices with switching costs

Index policies play an important role in solving MAB problems. These policies, which rely on numerical indices to evaluate and select arms, must be carefully tailored to effectively balance expected rewards against the inherent uncertainties and costs of switching strategies. This section delves into the details of index policies in the context of risk-averse MAB problems, examining both their limitations and potential enhancements to accommodate switching costs. We begin by revisiting the traditional Gittins index and its recent extensions to risk-averse scenarios, then propose a novel index policy that explicitly incorporates switching costs, thereby offering a more comprehensive framework for decision-making in these complex environments. Further, we extend our discussion to encompass restricted environments, including deterministic and stochastic settings, providing insights into the optimal policy structures in these specific contexts. Lastly, we tackle the computational aspects of the new indices, proposing a method that synthesizes the algorithmic approaches used in both risk-neutral cases with switching costs and risk-averse scenarios without switching costs.

4.1. Evolving index policies: Addressing switching costs in risk-averse scenarios

Traditional Gittins index policies do not account for the additional complexities introduced by switching costs. Asawa and Teneketzis (1996) addressed this by defining a "switching index" to handle switching penalties, including both costs and delays. Malekipirbazari and Çavuş (2021) extended the Gittins index to incorporate risk aversion, showing that in risk-averse MAB problems with dynamic coherent risk measures, each arm is indexable. This adaptation allows for refined decision-making where both expected rewards and associated risks are considered. However, these advancements still do not fully address the challenges posed by switching costs. Our proposed methodology introduces a risk-averse switching index that integrates both potential rewards and switching costs, considering the decision-maker's risk preferences. This development builds on Malekipirbazari and Çavuş (2021), which introduced the "risk-averse allocation index" (RAI), extending the conceptual and computational framework to scenarios where both risk aversion and switching costs significantly influence decision-making processes.

4.2. Decomposition and index heuristic development

In pursuit of devising an index heuristic for the problem outlined in (9), we transition to an equivalent optimization challenge by broadening the scope of Π to encompass Π', a class of stationary policies that permit unrestricted action selection. This modification allows for the possibility of engaging no arm or multiple arms at any given step:

$$R(x_1, j) = \min_{\Pi\in\Pi'} \varrho^{\beta}\Big(c(x_1, u_1, x_2) + s^T u_1 \mathbb{1}_{u_1^j\neq 1}, c(x_2, u_2, x_3) + s^T u_2 \mathbb{1}_{u_2\neq u_1},$$
$$c(x_3, u_3, x_4) + s^T u_3 \mathbb{1}_{u_3\neq u_2}, \ldots\Big)$$
$$\text{s.t.} \sum_{i\in\mathcal{K}} u_t^i = 1,\ \forall t\in\mathbb{N}, \tag{10}$$
$$u_t^i \in \{0,1\},\ \forall i\in\mathcal{K},\ t\in\mathbb{N}. \tag{11}$$

The constraint (10) mandates the activation of precisely one arm at each step. This constraint is subsequently relaxed to:

$$\varrho^{\beta}\Big(\sum_{i\in\mathcal{K}} u_1^i, \sum_{i\in\mathcal{K}} u_2^i, \dots\Big) = \frac{1}{1-\beta}. \tag{12}$$

By substituting constraint (10) with (12) and applying a Lagrangian relaxation, we derive the following Lagrangian dual function:

$$\begin{aligned}\mathcal{L}_D(\nu, x_1, j) = \min_{\Pi\in\mathbf{\Pi}'} \ & \varrho^{\beta}\Big(c(x_1,u_1,x_2)+s^T u_1 \mathbb{1}_{u_1^j\neq 1}, c(x_2,u_2,x_3)+s^T u_2 \mathbb{1}_{u_2\neq u_1}, \\ & c(x_3,u_3,x_4)+s^T u_3 \mathbb{1}_{u_3\neq u_2}, \dots\Big) + \nu\varrho^{\beta}\left(\sum_{i\in\mathcal{K}} u_1^i, \sum_{i\in\mathcal{K}} u_2^i, \dots\right) - \nu\left(\frac{1}{1-\beta}\right) \\ & \text{s.t. } \ u_t^i \in \{0,1\}, \ \forall i\in\mathcal{K}, \ t\in\mathbb{N}.\end{aligned} \tag{13}$$

where $\nu \in \mathbb{R}$ represents the Lagrangian multiplier. Given the duality principle, $\mathcal{L}_D(\nu, x_1, j) \leq R(x_1, j)$ holds true for any $\nu \in \mathbb{R}$. Utilizing the subadditivity property of dynamic risk measures, discussed in Malekipirbazari and Çavuş (2021, Lemma III.1), we introduce a dual function $\mathcal{L}'_D(\nu, x_1, j)$ to approximate $R(x_1, j)$:

$$\begin{aligned}\mathcal{L}'_D(\nu, x_1, j) = \min_{\Pi\in\mathbf{\Pi}'} \sum_{i\in\mathcal{K}} & \varrho^{\beta}\Big(c^i(x_1^i,u_1^i,x_2^i)+s^i u_1^i \mathbb{1}_{i\neq j}, c^i(x_2^i,u_2^i,x_3^i)+s^i u_2^i \mathbb{1}_{u_2^i\neq u_1^i}, \\ & c^i(x_3^i,u_3^i,x_4^i) + s^i u_3^i \mathbb{1}_{u_3^i\neq u_2^i}, \dots\Big) + \nu\sum_{i\in\mathcal{K}} \varrho^{\beta}\left(u_1^i, u_2^i, \dots\right) - \nu\left(\frac{1}{1-\beta}\right) \\ & \text{s.t. } \ u_t^i \in \{0,1\}, \ \forall i\in\mathcal{K}, \ t\in\mathbb{N}.\end{aligned} \tag{14}$$

Although $\mathcal{L}_D(\nu, x_1, j)$ serves as a lower bound for $R(x_1, j)$, $\mathcal{L}'_D(\nu, x_1, j)$ may not necessarily be smaller than $R(x_1, j)$. Nonetheless, problem (14) facilitates an index-based policy, offering a feasible and interpretable strategy for the problem (9). We proceed to decompose problem (14) into K subproblems, each corresponding to individual arms, to derive this index-based policy and demonstrate its proximity to an optimal solution for (9) through computational analysis. Each subproblem for arm i, denoted as $\mathcal{L}'_{Di}(\nu, x_1^i, j)$, can be expressed and minimized independently:

$$\begin{aligned}\mathcal{L}'_{Di}(\nu, x_1^i, j) = \min_{\Pi^i\in\mathbf{\Pi}'^i} & \varrho^{\beta}\Big(c^i(x_1^i,u_1^i,x_2^i) \\ & + s^i u_1^i \mathbb{1}_{i\neq j}, c^i(x_2^i,u_2^i,x_3^i) + s^i u_2^i \mathbb{1}_{u_2^i\neq u_1^i}, \dots\Big) + \nu\varrho^{\beta}\left(u_1^i, u_2^i, \dots\right) \\ & \text{s.t. } \ u_t^i \in \{0,1\}, \ \forall t\in\mathbb{N},\end{aligned} \tag{15}$$

where $\mathbf{\Pi}'^i$ represents the set of all possible stationary policies for arm i, without any constraints on the actions at each step. The term $\nu/(1-\beta)$ present in (14) is omitted in subproblem (15) without affecting the optimal policies.

To elucidate the indexability of subproblem (15) and to outline the structure of the indices, we introduce a calibrating function for each arm. This function is pivotal for understanding the optimal policy structure for each arm within the broader MAB problem.

Lemma 1. *For arm $i \in \mathcal{K}$ and for all $t \in \mathbb{N}$, if the action "not play" is optimal for state $x_t^i \in \mathcal{X}^i$, then it remains optimal for state $x_l^i \in \mathcal{X}^i$, for all $l \geq t+1$.*

Proof. The optimality of the "not play" action for state x_t^i implies that, due to the stationary nature of the problem, the same action remains optimal for state x_l^i, for all $l \geq t+1$. This stationary property ensures consistency in the decision-making process across time steps. □

This lemma sets the stage for defining an optimal policy for subproblem (15), focusing on the individual dynamics of each arm. Accordingly, our analysis will focus on two distinct stationary policies: one that opts never to play and another that plays the arm continuously until a certain stopping time τ^i is reached. We proceed to derive an index-based policy that is not only interpretable but also closely approximates the optimal policy for the original problem (9). To achieve this, it is necessary to identify appropriate calibration functions for each arm. The resulted indices for arm i in state $x^i \in \mathcal{X}^i$ are determined as the values of ν at which the actions "play" and "not play" are deemed equally preferable, tailored to whether arm i was previously in play or not, respectively.

4.3. Risk-averse switching index (RASI)

Building on the conceptual framework outlined earlier, we now present the detailed mathematical formulation of the "risk-averse switching index" (RASI). RASI is calculated based on two scenarios: one where the arm is currently in play (denoted by $\mu^i(x_1^i, 1)$) and another where the arm is not currently in play (denoted by $\mu^i(x_1^i, 0)$). The formulation of RASI is based on the concept of dynamic coherent risk measures, which allows for a time-consistent evaluation of risk and reward.

Definition 1. The RASI for each state of arm i currently in play is given by:

$$\mu^i(x_1^i, 1) := \sup_{\tau^i>1} \frac{\varrho^{\beta}_{1,\tau^i-1}\left(c^i(x_1^i,1,x_2^i), c^i(x_2^i,1,x_3^i), \dots, c^i(x_{\tau^i-1}^i,1,x_{\tau^i}^i)\right)}{\varrho^{\beta}_{1,\tau^i-1}(-1,-1,\dots,-1)}. \tag{16}$$

Similarly, the RASI for each state of arm i not currently in play is:

$$\mu^i(x_1^i, 0) := \sup_{\tau^i>1} \frac{\varrho^{\beta}_{1,\tau^i-1}\left(s^i + c^i(x_1^i,1,x_2^i), c^i(x_2^i,1,x_3^i), \dots, c^i(x_{\tau^i-1}^i,1,x_{\tau^i}^i)\right)}{\varrho^{\beta}_{1,\tau^i-1}(-1,-1,\dots,-1)}. \tag{17}$$

Building upon the mathematical foundation of the RASI, we introduce a pragmatic and effective decision-making strategy, termed the "RASI policy". This policy is an index-based heuristic that leverages the computed RASI values to guide arm selection in a risk-averse MAB environment, particularly under conditions where switching costs are significant. It operationalizes the concept of RASI by systematically selecting the arm with the highest index value at each decision step.

For the indices $\mu^i(x_1^i, 1)$ and $\mu^i(x_1^i, 0)$ to effectively guide decision-making, they must establish a consistent state ordering. Specifically, if choosing not to play arm i in state x_1^i is optimal for a given index value, then not playing should remain optimal for any ν exceeding that index value. To formalize this, we introduce $\Theta_1^i(\nu)$ and $\Theta_0^i(\nu)$ as the sets of states for arm i where the optimal action is "not play", given specific ν values. This distinction is made based on whether arm i was previously active (1) or inactive (0). As per Whittle's indexability concept, arm i is indexable if, for every $\nu \in \mathbb{R}$, there exists an optimal policy for the subproblem (15) such that the size of $\Theta_1^i(\nu)$ and $\Theta_0^i(\nu)$ increases monotonically with ν. This criterion ensures a coherent policy structure that adapts based on the arm's current state of play. Accordingly, the indices are defined as

$$\mu^i(x_1^i, 1) = \inf\{\nu \in \mathbb{R} : x^i \in \Theta_1^i(\nu)\}$$

for states where the arm was active, and

$$\mu^i(x_1^i, 0) = \inf\{\nu \in \mathbb{R} : x^i \in \Theta_0^i(\nu)\}$$

for states where the arm was inactive. This framework ensures that the indices $\mu^i(x_1^i, 1)$ and $\mu^i(x_1^i, 0)$ provide a meaningful and actionable basis for decision-making in the context of risk-averse MAB problems with switching costs. Accordingly, for a given $\nu \in \mathbb{R}$, the stopping times $\tau_1^i(x_1^i)$ and $\tau_0^i(x_1^i)$ can be computed as:

$$\tau_1^i(x_1^i) := \inf\{t > 1 : x_t^i \in \Theta_1^i(\nu)\} \tag{18}$$

$$\tau_0^i(x_1^i) := \inf\{t > 1 : x_t^i \in \Theta_0^i(\nu)\} \tag{19}$$

Building on Lemma 1 and incorporating (18) and (19), it follows that for all $t \geq \tau_1^i(x_1^i)$, the state x_t^i falls within $\Theta_1^i(\nu)$, and similarly, for all $t \geq \tau_0^i(x_1^i)$, x_t^i is included in $\Theta_0^i(\nu)$. This observation simplifies the task of identifying an optimal policy to determining appropriate stopping times, as delineated by the indices introduced in Definition 1. For the sake of clarity, and in reference to equation (6) in Ruszczyński (2010), we adopt a notation where $\varrho^\beta_{1,\tau-1}(Z_2, Z_3, \dots, Z_\tau)$ is equated to $\varrho^\beta(Z_2, Z_3, \dots, Z_\tau, 0, 0, \dots)$, albeit with a broader interpretation to accommodate our context.

The following theorem asserts the indexability of each arm within the framework of RASI, thereby establishing a direct link between the RASI and the optimal solution strategy for (15). It demonstrates that our approach aligns with the foundational principles of index-based decision-making in the context of risk-averse multiarmed bandit problems with switching costs.

Theorem 3. *Each arm $i \in \mathcal{K}$ is indexable with respect to the RASI introduced in Definition 1.*

Proof. For arm $i \neq j$ and a given constant $\nu \in \mathbb{R}$, we examine the subproblem (15), considering two distinct policy approaches: either engaging the arm until a predetermined stopping time τ^i or opting not to engage it at all.

Engaging arm i from its initial state $x_1^i \in \mathcal{X}^i$ up to the stopping time τ^i yields the objective function value as follows:

$$\varrho^\beta_{1,\tau^i-1}\left(s^i + c^i(x_1^i, 1, x_2^i), \dots, c^i(x^i_{\tau^i-1}, 1, x^i_{\tau^i})\right) + \nu\varrho^\beta(\underbrace{1, \dots, 1}_{\tau^i-1}, 0, 0, \dots).$$

Conversely, if arm i is not engaged from state x_1^i, leveraging Lemma 1, the objective function value is determined as zero. Thus, the action to "play" is unequivocally optimal for x_1^i when there exists a stopping policy τ^i such that the above value is less than the objective value of "not play", specifically:

$$\varrho^\beta_{1,\tau^i-1}\left(s^i + c^i(x_1^i, 1, x_2^i), \dots, c^i(x^i_{\tau^i-1}, 1, x^i_{\tau^i})\right) + \nu\varrho^\beta_{1,\tau^i-1}(1, \dots, 1) < 0.$$

Reorganizing this inequality, we find:

$$\varrho^\beta_{1,\tau^i-1}\left(s^i + c^i(x_1^i, 1, x_2^i), \dots, c^i(x^i_{\tau^i-1}, 1, x^i_{\tau^i})\right) < \nu\varrho^\beta_{1,\tau^i-1}(-1, \dots, -1).$$

Therefore, the action "play" is optimal for arm $i \in \mathcal{K}$ at the initial state x_1^i, whenever there exists τ^i such that

$$\frac{\varrho^\beta_{1,\tau^i-1}\left(s^i + c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x^i_{\tau^i-1}, 1, x^i_{\tau^i})\right)}{\varrho^\beta_{1,\tau^i-1}(-1, \dots, -1)} > \nu,$$

leading to $\mu^i(x_1^i, 0) > \nu$.

Similarly, the action "not play" is strictly optimal for x_1^i if for all $\tau^i > 1$:

$$\frac{\varrho^\beta_{1,\tau^i-1}\left(s^i + c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x^i_{\tau^i-1}, 1, x^i_{\tau^i})\right)}{\varrho^\beta_{1,\tau^i-1}(-1, \dots, -1)} \leq \mu^i(x_1^i, 0) < \nu.$$

In essence, the action "play" is optimal in state x_1^i if and only if $\mu^i(x_1^i, 0) \geq \nu$, and the action "not play" is optimal if and only if $\mu^i(x_1^i, 0) \leq \nu$. This reasoning, along with a similar argument for the case where $i = j$, indicates the presence of a family of optimal policies with associated inactive sets $\Theta_1^i(\nu)$ and $\Theta_0^i(\nu)$, $i \in \mathcal{K}$, that expand nondecreasingly with ν. This establishes the indexability of each arm $i \in \mathcal{K}$ as per Definition 1. □

By decomposing the problem and focusing on the indexability of individual arms, we provide a pathway towards developing efficient and interpretable strategies that are both theoretically sound and practically viable. This approach not only enhances our understanding of the optimal decision-making process in complex environments but also offers a scalable solution to the challenges posed by risk aversion and switching penalties in MAB problems.

4.4. Analyzing the dynamics of RASI in MAB environments

The introduction of the RASI brings a new layer of complexity and strategic depth to the decision-making process in our problem. To understand the practical implications of RASI, we first recall the stopping times $\tau_1^i(x_1^i)$ and $\tau_0^i(x_1^i)$, which represent the points at which the supremum values in (16) and (17) are achieved for a given state $x_1^i \in \mathcal{X}^i$. To simplify the notation, we often omit the dependence of τ_1^i and τ_0^i on x_1^i throughout this section. The following propositions offer insights into the behavior of these indices and their impact on decision-making.

Proposition 1. *For $i \in \mathcal{K}$, for each state $x_1^i \in \mathcal{X}^i$, $\mu^i(x_1^i, 1) - \mu^i(x_1^i, 0)$ is a positive and increasing function with respect to the switching cost of arm i.*

Proof. According to (16) and (17) and the definitions of τ_1^i and τ_0^i, we have

$$\begin{aligned}
\mu^i(x_1^i, 1) &= \frac{\varrho^\beta_{1,\tau_1^i-1}\left(c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x^i_{\tau_1^i-1}, 1, x^i_{\tau_1^i})\right)}{\varrho^\beta_{1,\tau_1^i-1}(-1, -1, \dots, -1)} \\
&\geq \frac{\varrho^\beta_{1,\tau_0^i-1}\left(c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x^i_{\tau_0^i-1}, 1, x^i_{\tau_0^i})\right)}{\varrho^\beta_{1,\tau_0^i-1}(-1, -1, \dots, -1)} \\
&> \frac{\varrho^\beta_{1,\tau_0^i-1}\left(s^i + c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x^i_{\tau_0^i-1}, 1, x^i_{\tau_0^i})\right)}{\varrho^\beta_{1,\tau_0^i-1}(-1, -1, \dots, -1)} \\
&= \mu^i(x_1^i, 0),
\end{aligned}$$

where the second inequality is due to the positivity of switching costs along with axiom (A3). This establishes the positivity property.

Now assume to the contrary that for some switching costs s_1^i and s_2^i where $s_1^i < s_2^i$, we have

$$\mu^i(x_1^i, 1) - \sup_{\tau^i > 1} \frac{\varrho^\beta_{1,\tau^i-1}\left(s_1^i + c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x^i_{\tau^i-1}, 1, x^i_{\tau^i})\right)}{\varrho^\beta_{1,\tau^i-1}(-1, -1, \dots, -1)} \geq$$
$$\mu^i(x_1^i, 1) - \sup_{\tau^i > 1} \frac{\varrho^\beta_{1,\tau^i-1}\left(s_2^i + c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x^i_{\tau^i-1}, 1, x^i_{\tau^i})\right)}{\varrho^\beta_{1,\tau^i-1}(-1, -1, \dots, -1)}.$$

This implies that for any $\hat{\tau} > 1$, we have

$$\frac{\varrho^\beta_{1,\hat{\tau}-1}\left(s_1^i + c^i(x_1^i, 1, x_2^i), \dots, c^i(x^i_{\hat{\tau}-1}, 1, x^i_{\hat{\tau}})\right)}{\varrho^\beta_{1,\hat{\tau}-1}(-1, -1, \dots, -1)} \leq$$
$$\sup_{\tau^i > 1} \frac{\varrho^\beta_{1,\tau^i-1}\left(s_2^i + c^i(x_1^i, 1, x_2^i), \dots, c^i(x^i_{\tau^i-1}, 1, x^i_{\tau^i})\right)}{\varrho^\beta_{1,\tau^i-1}(-1, -1, \dots, -1)},$$

from which and by letting τ_0^i as the stopping time achieving the supremum value on the right-hand side, we can write

$$\frac{\varrho^\beta_{1,\tau_0^i-1}\left(s_1^i + c^i(x_1^i, 1, x_2^i), \dots, c^i(x^i_{\tau_0^i-1}, 1, x^i_{\tau_0^i})\right)}{\varrho^\beta_{1,\tau_0^i-1}(-1, -1, \dots, -1)} \leq$$
$$\frac{\varrho^\beta_{1,\tau_0^i-1}\left(s_2^i + c^i(x_1^i, 1, x_2^i), \dots, c^i(x^i_{\tau_0^i-1}, 1, x^i_{\tau_0^i})\right)}{\varrho^\beta_{1,\tau_0^i-1}(-1, -1, \dots, -1)}.$$

This implies that $s_1^i \geq s_2^i$ which contradicts our assumption that $s_1^i < s_2^i$ and thus establishes the increasing property. □

This proposition suggests that the decision to continue playing an arm that is already active becomes increasingly favorable as the switching cost rises. This outcome aligns with intuitive expectations, as higher switching costs naturally discourage frequent transitions between arms.

The computation of RASI, similar to the Gittins index, involves optimization across a range of stopping times. The next proposition explores the detailed relationship between these stopping times, revealing how they impact the decision-making process within the RASI framework.

Proposition 2. *For $i \in \mathcal{K}$, for each state $x_1^i \in \mathcal{X}^i$, we have $\tau_1^i(x_1^i) \leq \tau_0^i(x_1^i)$.*

Proof. Assume to the contrary that for some state x_1^i, we have $\tau_1^i(x_1^i) > \tau_0^i(x_1^i)$. According to the definition of RASI for states in the immediately played arm given in (16), we have

$$\frac{\varrho^{\beta}_{1,\tau_1^i-1}\left(c^i(x_1^i,1,x_2^i),\dots,c^i(x^i_{\tau_1^i-1},1,x^i_{\tau_1^i})\right)}{\varrho^{\beta}_{1,\tau_1^i-1}(-1,-1,\dots,-1)} \geq \frac{\varrho^{\beta}_{1,\tau_0^i-1}\left(c^i(x_1^i,1,x_2^i),\dots,c^i(x^i_{\tau_0^i-1},1,x^i_{\tau_0^i})\right)}{\varrho^{\beta}_{1,\tau_0^i-1}(-1,-1,\dots,-1)}.$$

Based on positivity of s^i values and the assumption of $\tau_1^i(x_1^i) > \tau_0^i(x_1^i)$, we also have

$$\frac{s^i}{\varrho^{\beta}_{1,\tau_0^i-1}(-1,-1,\dots,-1)} < \frac{s^i}{\varrho^{\beta}_{1,\tau_1^i-1}(-1,-1,\dots,-1)}.$$

The above two inequalities imply that

$$\frac{\varrho^{\beta}_{1,\tau_1^i-1}\left(s^i+c^i(x_1^i,1,x_2^i),\dots,c^i(x^i_{\tau_1^i-1},1,x^i_{\tau_1^i})\right)}{\varrho^{\beta}_{1,\tau_1^i-1}(-1,-1,\dots,-1)} > \frac{\varrho^{\beta}_{1,\tau_0^i-1}\left(s^i+c^i(x_1^i,1,x_2^i),\dots,c^i(x^i_{\tau_0^i-1},1,x^i_{\tau_0^i})\right)}{\varrho^{\beta}_{1,\tau_0^i-1}(-1,-1,\dots,-1)},$$

which contradicts the definition of RASI for the states in the non-immediately played arms given in (17) and thus completes the proof. □

The forthcoming Proposition 3 will establish the boundedness of the RASI, ensuring its practical applicability and theoretical soundness within our framework.

Proposition 3. *RASI is guaranteed to be a finite term for any arm $i \in \mathcal{K}$, with specific bounds determined by the arm's minimum and maximum playing costs, C_L^i and C_U^i, and the switching cost s^i. Specifically, for each state $x_1^i \in \mathcal{X}^i$, the RASI values are constrained within the following range:*

$$-(1-\beta)s^i - C_U^i \leq \mu^i(x_1^i,0) < \mu^i(x_1^i,1) \leq -C_L^i$$

Proof. Let us consider any arm $i \in \mathcal{K}$, such that for all $x^i, y^i \in \mathcal{X}^i$, we have $C_L^i \leq c^i(x^i,1,y^i) \leq C_U^i$. First, by normalizing the RASI calculation for the scenario where the arm is currently in play (16) against $-C_L^i$, we derive:

$$\begin{aligned}\frac{\mu^i(x_1^i,1)}{-C_L^i} &= \sup_{\tau^i>1}\frac{\varrho^{\beta}_{1,\tau^i-1}\left(c^i(x_1^i,1,x_2^i),c^i(x_2^i,1,x_3^i),\dots,c^i(x^i_{\tau^i-1},1,x^i_{\tau^i})\right)}{-C_L^i\ \varrho^{\beta}_{1,\tau^i-1}(-1,-1,\dots,-1)}\\ &= \sup_{\tau^i>1}\frac{\varrho^{\beta}_{1,\tau^i-1}\left(c^i(x_1^i,1,x_2^i),c^i(x_2^i,1,x_3^i),\dots,c^i(x^i_{\tau^i-1},1,x^i_{\tau^i})\right)}{\varrho^{\beta}_{1,\tau^i-1}(C_L^i,C_L^i,\dots,C_L^i)} \leq 1,\end{aligned}$$

where the equality follows from the positive homogeneity axiom 4, and the inequality is justified by the monotonicity property (A2) of dynamic risk measures.

Next, by normalizing the RASI calculation for the scenario where the arm is not currently in play (17) against $-C_U^i$, we find:

$$\begin{aligned}\frac{\mu^i(x_1^i,0)}{-C_U^i} &= \sup_{\tau^i>1}\frac{\varrho^{\beta}_{1,\tau^i-1}\left(s^i+c^i(x_1^i,1,x_2^i),c^i(x_2^i,1,x_3^i),\dots,c^i(x^i_{\tau^i-1},1,x^i_{\tau^i})\right)}{-C_U^i\ \varrho^{\beta}_{1,\tau^i-1}(-1,-1,\dots,-1)}\\ &= \sup_{\tau^i>1}\frac{\varrho^{\beta}_{1,\tau^i-1}\left(s^i+c^i(x_1^i,1,x_2^i),c^i(x_2^i,1,x_3^i),\dots,c^i(x^i_{\tau^i-1},1,x^i_{\tau^i})\right)}{\varrho^{\beta}_{1,\tau^i-1}(C_U^i,C_U^i,\dots,C_U^i)}\\ &\geq 1+\sup_{\tau^i>1}\frac{s^i}{\varrho^{\beta}_{1,\tau^i-1}(C_U^i,C_U^i,\dots,C_U^i)} = 1+(1-\beta)s^i/C_U^i,\end{aligned}$$

where the addition of $(1-\beta)s^i/C_U^i$ to 1 reflects the inclusion of switching costs in the numerator.

Given the bounds above and in light of Proposition 1, along with the bounded nature of costs in our model, we can guarantee the finiteness of the index. □

RASI is thus assured to be finite, sharing a structural resemblance and interpretative parallel with the Gittins index. It can be viewed as an extended version of the Gittins index, adeptly incorporating dynamic risk aversion and the cost of switching between arms into its calculations. RASI demonstrates notable adaptability under varying conditions. As risk aversion diminishes to zero, RASI aligns with the switching index from Asawa and Teneketzis (1994), effectively transitioning to a risk-neutral framework. Conversely, when switching costs become negligible, RASI evolves into the RAI detailed in Malekipirbazari and Çavuş (2021), emphasizing its foundation in dynamic risk measures. These convergence properties highlight RASI's robustness, bridging risk-neutral and risk-averse strategies while accommodating the complexities introduced by switching costs.

As we assess the complexities and potential of the RASI policy in addressing the challenges of our MAB problem, it becomes clear that an accurate estimation of switching costs is essential. Switching costs can vary significantly across different scenarios and environments, and any inaccuracies in their estimation can substantially reduce the effectiveness of the RASI policy. This underscores the importance of robust and context-sensitive methodologies in assessing and incorporating switching costs into the RASI framework. Ensuring precise integration of these costs is crucial for the policy to accurately reflect the real-world trade-offs and benefits associated with switching decisions.

4.5. RASI policy in restricted environments

In this section, we evaluate the efficacy of the RASI policy in addressing special bandit problems that entail switching costs. Our focus is on environments with at least one single-state arm, as these scenarios offer insightful perspectives on the RASI values in a risk-averse MAB problem with switching costs. Given that, it can be informative to examine the RASI values associated with such arms in a risk-averse MAB problem that involves switching costs. We explore this in the subsequent lemma.

Lemma 2. *The RASI values for a single-state arm x are $\mu(x,1) = -c$ and $\mu(x,0) = -c-(1-\beta)s$, where c is the cost of playing the state of the arm, and s is the switching cost to this arm.*

Proof. The RASI of a single state x with respect to (16) is

$$\begin{aligned}\mu(x,1) &= \sup_{\tau>1}\frac{\varrho^{\beta}_{1,\tau-1}(c,c,\dots,c)}{\varrho^{\beta}_{1,\tau-1}(-1,-1,\dots,-1)}\\ &= \sup_{\tau>1}\frac{-c\ \varrho^{\beta}_{1,\tau-1}(-1,-1,\dots,-1)}{\varrho^{\beta}_{1,\tau-1}(-1,-1,\dots,-1)} = -c,\end{aligned}$$

where the second equality is due to axiom (A4). Similarly, the RASI of single state x with respect to (17) is

$$\mu(x,0) = \sup_{\tau>1} \frac{\varrho^{\beta}_{1,\tau-1}(s+c,c,\dots,c)}{\varrho^{\beta}_{1,\tau-1}(-1,-1,\dots,-1)} = -c - s(1-\beta)\inf_{\tau>1}\frac{1}{1-\beta^{\tau-1}} = -c - s(1-\beta),$$

where the second equality is derived by applying axioms (A3) and (A4). □

This lemma illustrates the implications of Proposition 1, where $\mu(x,1)-\mu(x,0) = (1-\beta)s$ is a positive and strictly monotone function of the switching cost. Additionally, for a single-state arm x, we observe that $\tau_1(x) = 2$ while $\tau_0(x) = \infty$, aligning with Proposition 2.

4.5.1. Single-state arms with switching costs.

Consider a MAB problem where each arm consists of a single state, and playing a state in arm i incurs a cost c^i, $i \in \mathcal{K}$. Theorem 4 demonstrates the optimality of the RASI policy for such bandits, highlighting its adaptability and effectiveness in environments characterized by single-state arms and switching costs.

Theorem 4. *In risk-averse multiarmed bandits comprised of single-state arms with switching costs, the RASI policy is optimal.*

4.5.2. Risk-averse one-armed bandit problem with switching costs.

We also assess the optimal policy structure for a special case of two-armed bandits with switching costs, where one arm has a single state. This scenario, known as the one-armed bandit problem, features one stochastic arm with a Markovian reward structure. The following theorem asserts the optimality of the RASI policy in this context, further reinforcing its utility in a broad range of risk-averse decision-making scenarios.

Theorem 5. *In risk-averse one-armed bandits with switching costs, the RASI policy is optimal.*

4.6. Applicability of RASI in deterministic and stochastic settings

Our exploration of the RASI policy's performance extends to both deterministic and stochastic MAB problems with switching costs, drawing upon the seminal work of Asawa and Teneketzis (1996).

4.6.1. Deterministic two-armed bandit problem

Asawa and Teneketzis (1996) established a critical theorem for deterministic two-armed bandits with switching costs, asserting that optimal scheduling policies require decisions only at specific time instants where an appropriate index is achieved. This finding is particularly relevant to our risk-averse scenario in deterministic cases, as risk aversion aligns with the risk-neutral perspective in such settings.

4.6.2. Stochastic multiarmed bandit problem

Theorem 3.1 in Asawa and Teneketzis (1996) extends to stochastic MAB problems with switching costs, suggesting that optimal scheduling decisions are made at stopping times achieving an appropriate index. This theorem implies that if the index policy in a risk-neutral case advises continuing with the currently played arm, such action is optimal. However, it is crucial to note that the applicability of Asawa and Teneketzis's approach to risk-averse scenarios is limited. Their work leverages the fact that the Gittins index provides an optimal policy for risk-neutral cases without switching costs. In contrast, the risk-averse case without switching costs does not exhibit this property, thus precluding a direct application of their approach to general risk-averse scenarios. Thus, while the RASI policy demonstrates effectiveness in various settings as evidenced by our numerical experiments, a comprehensive analytical investigation of its application in general risk-averse MAB problems remains a valuable direction for future research.

4.7. Risk-averse MAB problems with switching delays

In addition to the scenario involving switching costs, we extend our analysis to encompass the case of switching delays, which we consider equally critical. The MAB problem with switching delays parallels the problem with switching costs, with the primary distinction being the nature of the incurred penalty when transitioning between arms. By incorporating switching delays, we acknowledge the real-world scenario where transitions between tasks or projects are not instantaneous and carry inherent time-based costs. Specifically, for each arm i, a switching (setup) delay d^i is experienced when the decision-maker moves from one arm to another. During this delay interval, no cost or reward is accumulated. We assume that the delay d^i is a nonnegative integer-valued random variable with a known distribution, satisfying $0 < \mathbb{E}[d^i] < \infty$, and is independent of the machine dynamics.

The objective in the MAB problem with switching delays is to identify a policy Π that minimizes the following risk-adjusted cost function:

$$R'^{\Pi}(x_1) = \beta^{d^T u_1}\rho_1\Big(c(x_1,u_1,x_2) + \beta^{1+d^T u_2 \mathbf{1}_{u_2\neq u_1}}\rho_2\Big(c(x_2,u_2,x_3) + \beta^{1+d^T u_3 \mathbf{1}_{u_3\neq u_2}}\rho_3(\dots)\Big)\Big) \tag{20}$$

To address this variant, we introduce the Risk-Averse Switching Delay Index (RASDI), defined as follows:

Definition 2. The RASDI for each state of arm i currently in play is given by:

$$\psi^i(x_1^i,1) := \sup_{\tau^i>1}\frac{\varrho^{\beta}_{1,\tau^i-1}\Big(c^i(x_1^i,1,x_2^i),\dots,c^i(x^i_{\tau^i-1},1,x^i_{\tau^i})\Big)}{\varrho^{\beta}_{1,\tau^i-1}(-1,-1,\dots,-1)}. \tag{21}$$

Similarly, the RASDI for each state of arm i not currently in play is:

$$\psi^i(x_1^i,0) := \sup_{\tau^i>1}\frac{\beta^{d^i}\varrho^{\beta}_{1,\tau^i-1}\Big(c^i(x_1^i,1,x_2^i),\dots,c^i(x^i_{\tau^i-1},1,x^i_{\tau^i})\Big)}{\varrho^{\beta}_{1,\tau^i+d^i-1}(-1,-1,\dots,-1)}. \tag{22}$$

Similar to the case with switching costs, the index rule is not optimal for the problem with switching delays. However, analogous results to those presented in the previous sections are applicable to the delay scenario. Specifically, the RASDI provides a heuristic for decision-making in the context of switching delays, balancing the trade-off between immediate rewards and the potential long-term impact of delays.

4.8. Computation of the RASI values

The computation of the RASI values presents unique challenges, especially when compared to the computation of the Gittins index in risk-neutral scenarios. In this subsection, we propose a method to calculate RASI, inspired by Asawa and Teneketzis (1996)'s algorithm for the risk-neutral case with switching costs and Malekipirbazari and Çavuş (2021)'s approach for the risk-averse scenario without switching costs.

The core of our proposed algorithm lies in solving single-arm optimal stopping problems, similar to the approach used for computing Gittins indices. Given an arm i in a risk-averse MAB problem, let x denote a generic state of this arm. We consider the evolution of this arm under the "play" action, starting from state x. For each arm i and state x, we define a stopping set Φ^i_k. This stopping set Φ^i_k includes all states of arm i, excluding the $k-1$ states with the highest indices. The stopping time $\tau^i_x(\Phi^i_k)$ is then the first time the state of arm i enters the set Φ^i_k, starting from x. The indices for a state x in arm i, given that it has the kth highest index value, is computed as follows:

$$N^i_x(\Phi^i_k) = \frac{\varrho^{\beta}_{1,\tau^i_x(\Phi^i_k)-1}\Big(c^i(x,1,x_2^i),\dots,c^i(x^i_{\tau^i_x(\Phi^i_k)-1},1,x^i_{\tau^i_x(\Phi^i_k)})\Big)}{\varrho^{\beta}_{1,\tau_x(\Phi^i_k)-1}(-1,\dots,-1)}.$$

To adapt this algorithm for the computation of RASI, we construct a new Markov chain with an expanded state space $\{1, 2, \dots, M, 1', 2', \dots, M'\}$, where $M = |\mathcal{X}^i|$. The transition probability matrix $\hat{Q}^i$ and the associated costs $\hat{c}^i$ are defined as follows:

$$\hat{Q}^i(x,1,y) = Q^i(x,1,y) \text{ and } \hat{c}^i(x,1,y) = c^i(x,1,y), \text{ for } x, y \in \{1, 2, \dots, M\}$$
$$\hat{Q}^i(x,1,y) = 0 \text{ and } \hat{c}^i(x,1,y) = 0, \text{ for } x \in \{1, 2, \dots, M\}, y \in \{1', 2', \dots, M'\}$$
$$\hat{Q}^i(x,1,y) = Q^i(x,1,y) \text{ and } \hat{c}^i(x,1,y) = c^i(x,1,y) + s^i, \text{ for } x \in \{1', 2', \dots, M'\}, y \in \{1, 2, \dots, M\}$$
$$\hat{Q}^i(x,1,y) = 0 \text{ and } \hat{c}^i(x,1,y) = 0, \text{ for } x, y \in \{1', 2', \dots, M'\}$$

This expanded Markov chain and the associated cost structure enable the computation of RASI by considering both the risk measures and the switching costs. The algorithm iteratively identifies the index values for each state in a given arm, starting with the state with the highest index and progressively recalculating the index values after excluding the highest index state from the previous iteration. This process continues until all states are ranked according to their index values, effectively integrating the complexities of risk aversion and switching costs into the decision-making framework of the MAB problem.

It is important to note that the RASI values for each state are determined based on their respective subsets in the expanded state space. Specifically, $\mu^i(x,1)$, the RASI value when the arm is currently in play, will correspond to the computed index in the state subset $\{1, 2, \dots, M\}$. Conversely, $\mu^i(x,0)$, the RASI value when the arm is not currently in play, will correspond to the computed index in the state subset $\{1', 2', \dots, M'\}$. This distinction is crucial for accurately reflecting the impact of switching costs in the computation of RASI values, ensuring that the indices provide a comprehensive representation of the strategic choices available in our MAB problem.

It is also worth noting that Malekipirbazari and Çavuş (2021) compared the computational complexity of their index heuristic to solving the corresponding MDP. They revealed that their heuristic is time-efficient, with computation time growing linearly with the problem size, whereas the risk-averse MDP computation time shows exponential growth. Our proposed method to calculate RASI is inspired by their approach for the risk-averse scenario without switching costs. The difference is that our proposed method jointly computes the index of an arm with $2M$ states, resulting in an increase in arithmetic operations relative to those in Malekipirbazari and Çavuş (2021). Despite this increase in arithmetic operations, the complexity of our method remains manageable since it scales linearly with the number of arms, compared to the exponential growth of solving the MDP. Therefore, while the computation of RASI values involves additional complexity due to the expanded state space, the linear scalability of our method ensures that it remains practical for larger instances. This makes the RASI policy a feasible and effective strategy for risk-averse decision-making in MAB problems with switching costs.

5. Numerical experiments

This section is dedicated to assessing the efficiency of our proposed index-based policy through a series of computational experiments. Through these experiments, we aim to provide a comprehensive evaluation of the proposed policy under varying conditions, offering insights into their practical applicability and performance in risk-averse settings.

5.1. Setup

We explore the following policy implementations for each test scenario:

(1) The optimal policy via risk-averse value iteration algorithm, detailed in Ruszczyński (2010).
(2) The risk-neutral switching index policy (RN), proposed by Asawa and Teneketzis (1996).
(3) The RAI policy, proposed by Malekipirbazari and Çavuş (2021).
(4) The RASI policy, introduced in this paper.

Our initial test bed comprises a bandit problem with three arms, each containing four states, resulting in a total state space of 64. For each test case, the transition probabilities under the play action for each arm are randomly generated from a uniform distribution and normalized to ensure row-wise probability summation to one. The costs associated with state transitions are drawn from a truncated normal distribution with mean values uniformly distributed between −6 and −5, and standard deviations set at $\{0.01, 0.5, 1\}$. The switching costs are sampled from a truncated normal distribution $N(s, 0.1s)$, where s varies over $\{0, 2, 4\}$. The experiments are conducted with discount factors β set at $\{0.50, 0.75, 0.90\}$, and the first-order mean-semideviation risk measure is employed as defined in (1), with κ values ranging from 0 to 1 in increments of 0.25.

For each parameter combination, we generate 1000 random test instances. In each instance, we compare the performance of the RN, RAI, and RASI policies using two key metrics:

(1) the suboptimality percentage of policy $\Pi \in \{\text{RN}, \text{RAI}, \text{RASI}\}$ 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 Π and switching cost s,
(2) optimality percentage of policy $\Pi \in \{\text{RN}, \text{RAI}, \text{RASI}\}$. It is computed as $100 \times \sum_{x \in \mathcal{X}} \mathbb{1}_{\text{subopt}}(x)/|\mathcal{X}|$, where $\mathbb{1}_{\text{subopt}}(x)$ is an indicator function that is equal to one if the suboptimality of the decision at state x under policy Π and switching cost of s is zero (that is, $R^{\Pi}(x) = R(x)$), and zero otherwise.

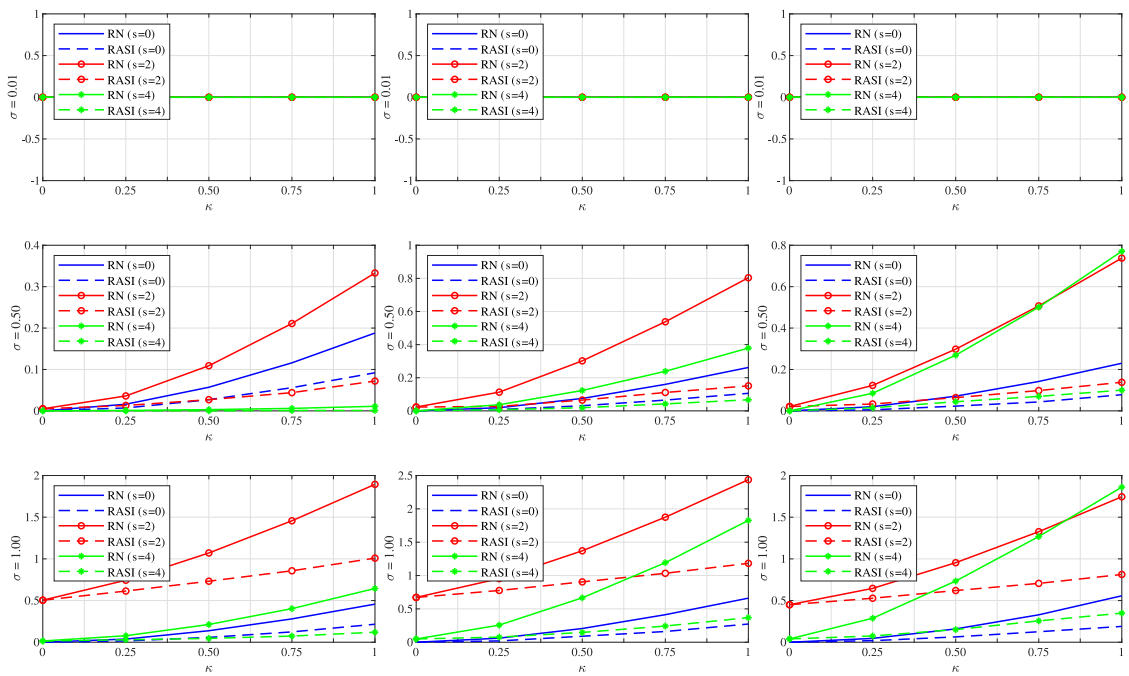
5.2. Discussion of the numerical results

In each of our generated instances, we calculated both the median and maximum suboptimality percentages across all states. Figs. 1 and 4 showcase the average maximum suboptimality percentages for policies RASI vs. RN and RASI vs. RAI, respectively. These figures aggregate results from all sets of 1000 test instances for each discount factor: 0.50, 0.75, and 0.90. Similarly, Figs. 2 and 5 illustrate the average median suboptimality percentages for the same policy comparisons and discount factors, also averaged across 1000 test instances. Furthermore, Figs. 3 and 6 present the average optimality percentage values for policies RASI vs. RN and RASI vs. RAI across the mentioned discount factors.

The experiments uncover fascinating patterns and trends, providing significant insights into the performance of risk-neutral and risk-averse policies under various conditions. Initially, we focus on examining the dynamics between the policies of RASI and RN, highlighting their behavior across different scenarios. Subsequently, our analysis shifts to explore the performances of RASI versus RAI, delving into how these interactions are influenced by varying levels of cost variability (σ), switching costs (s), and risk aversion parameters (κ).

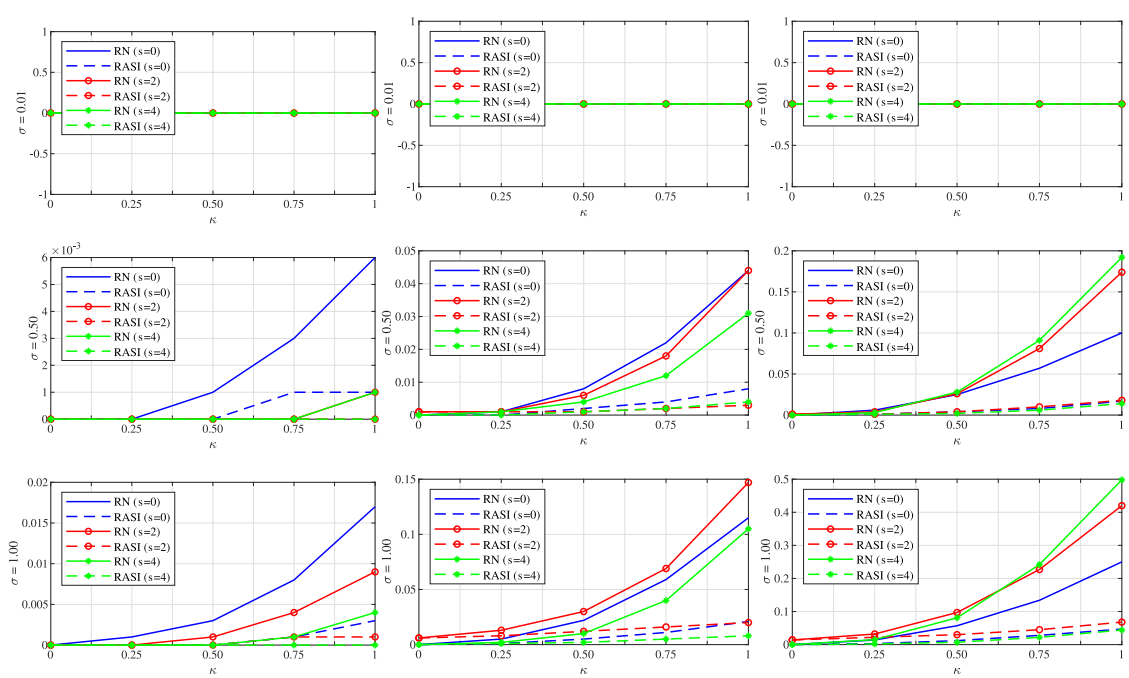
5.2.1. Evaluating RASI against RN

In scenarios with low-cost variability (specifically, $\sigma = 0.01$), both the RN and RASI policies demonstrate negligible suboptimality, with maximum and median suboptimality percentages consistently close to zero across all levels of κ. This trend suggests that in environments where costs are predictable, both policies are highly effective. However, as the cost variability increased, the RASI policy consistently outperformed the RN policy in terms of both maximum and median suboptimality percentages along with the average optimality percentages. For instance, Fig. 3 exhibits that with $\sigma = 0.50$ and $\beta = 0.50$, the RASI policy outperformed the RN policy, particularly at higher risk aversion levels ($\kappa = 1$), showing an average optimality percentage improvement from 91.91% to 96.98%. This trend was even more



(a) $\beta = 0.50$ (b) $\beta = 0.75$ (c) $\beta = 0.90$

Fig. 1. Average of maximum suboptimality percentages for policies RN and RASI using first-order mean-semideviation risk measure.

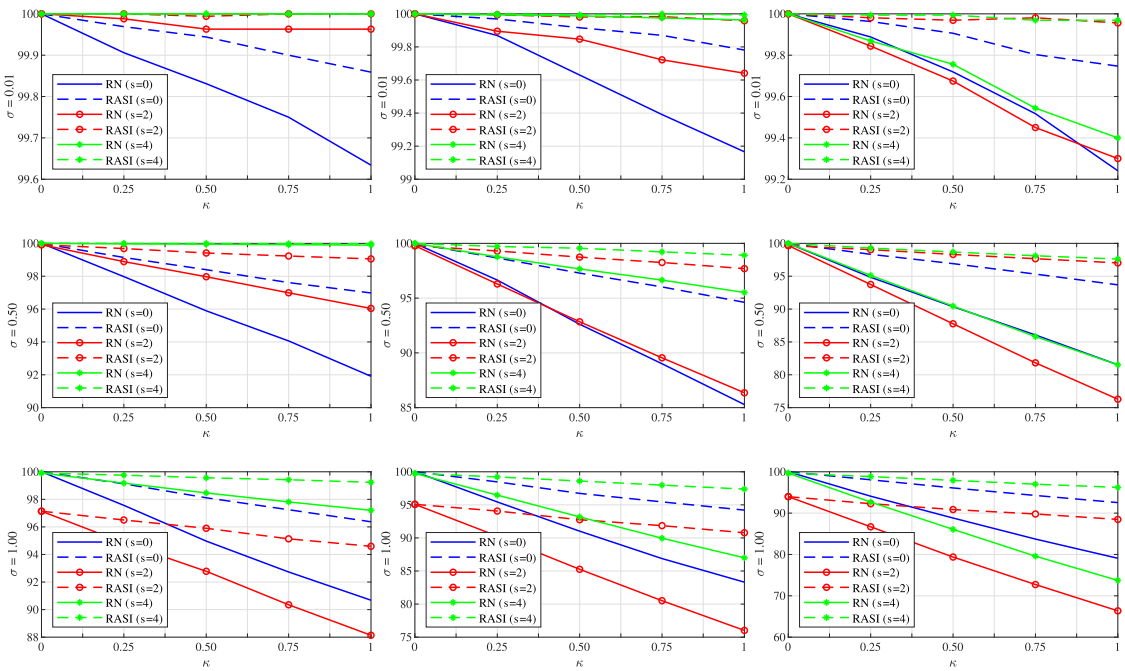


(a) $\beta = 0.50$ (b) $\beta = 0.75$ (c) $\beta = 0.90$

Fig. 2. Average of median suboptimality percentages for policies RN and RASI using first-order mean-semideviation risk measure.

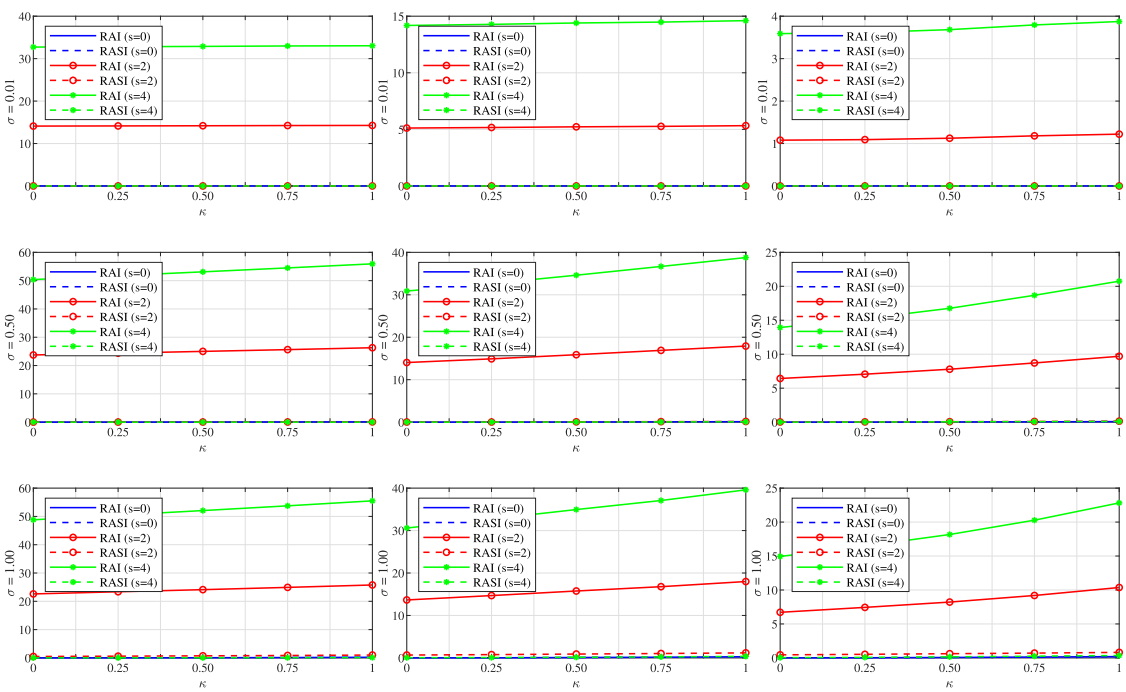
pronounced with $\sigma = 1$ and $\beta = 0.90$, where the RASI policy's average optimality percentage was significantly higher than the RN's (92.55% vs. 79.08%), highlighting the RASI policy's superior adaptability in high variability and risk-averse settings. The maximum and median suboptimality percentages also reflect this trend, with the RASI policy maintaining lower values, indicating more consistent performance near the optimal policy (see Figs. 1 and 2).

The addition of switching costs introduces a significant layer of complexity to the decision-making process, distinctly affecting the performance of the two policies under consideration. Initially, let us examine the scenario where the decision-maker exhibits no risk aversion, indicated by $\kappa = 0$. In this case, both policies yield identical results. However, as depicted in Figs. 1–3, particularly in situations of high-cost variability, an interesting pattern emerges with the variation in switching costs. When switching costs increase from none to a



(a) $\beta = 0.50$ (b) $\beta = 0.75$ (c) $\beta = 0.90$

Fig. 3. Average optimality percentages for policies RN and RASI using first-order mean-semideviation risk measure.

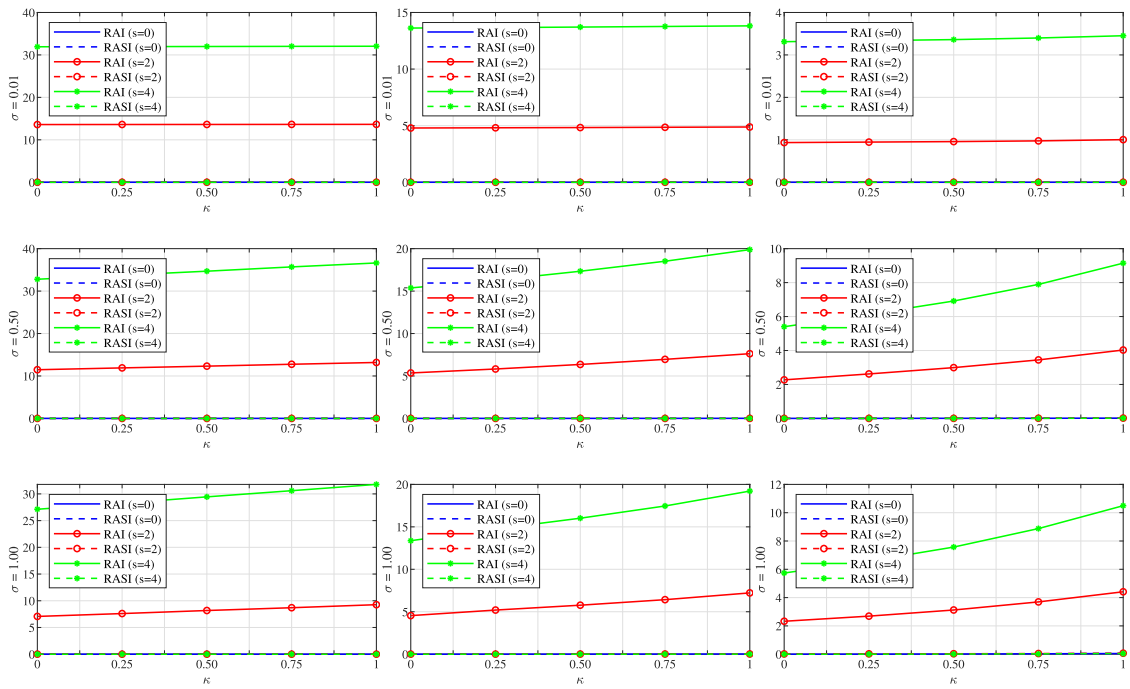


(a) $\beta = 0.50$ (b) $\beta = 0.75$ (c) $\beta = 0.90$

Fig. 4. Average of maximum suboptimality percentages for policies RAI and RASI using first-order mean-semideviation risk measure.

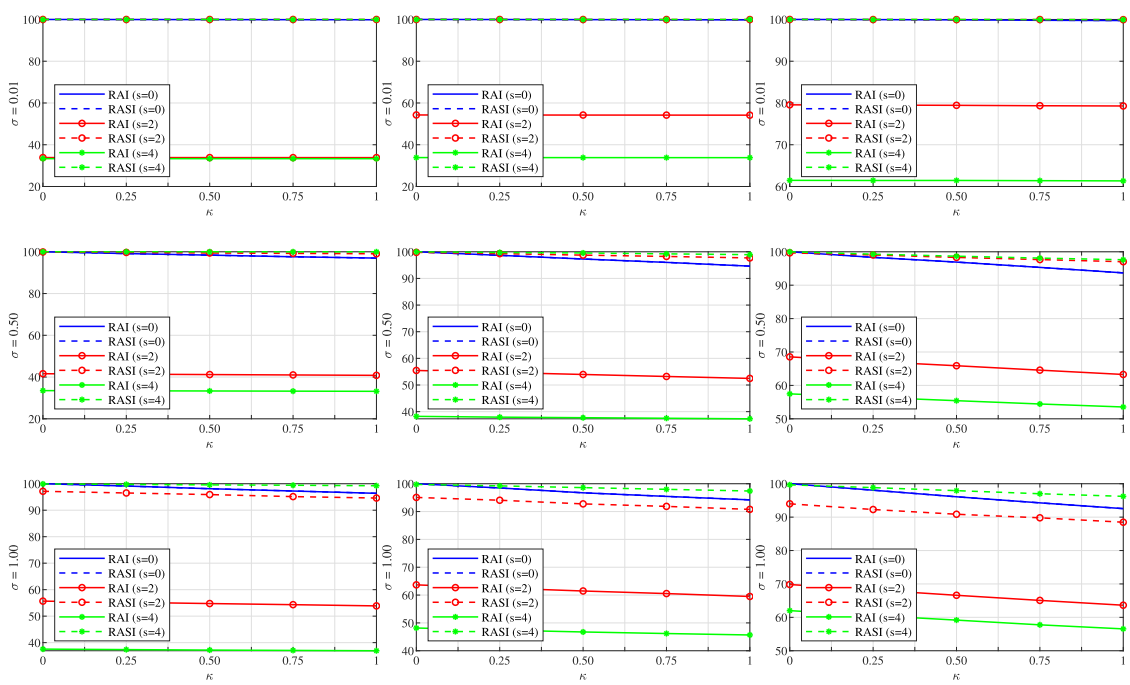
moderate level (i.e., from $s = 0$ to $s = 2$), there is a noticeable impact: the average maximum suboptimality rises by 0.45%, and the average similarity to the optimal policy decreases by over 6%. Interestingly, further elevating the switching costs from moderate to high (i.e., from $s = 2$ to $s = 4$) leads to an improvement in policy efficiency. This trend suggests that while escalating switching costs initially diminish the effectiveness of the index policy, this effect only persists up to a certain threshold. Beyond this point, as the optimal policy increasingly favors maintaining the current choice (due to higher costs of switching), the index policy regains its strength, closely approximating optimal performance. This improvement in policy efficiency, especially in scenarios with high switching costs, confirms the fact that the index policy is particularly optimal in prescribing the action of staying.

Now, let us examine how both policies perform when a risk-averse agent is involved, particularly in scenarios with switching costs. The RASI policy consistently aligns more closely with the optimal policy at higher levels of risk aversion. This alignment is evident in its higher average optimality percentages and lower suboptimality percentages



(a) $\beta = 0.50$ (b) $\beta = 0.75$ (c) $\beta = 0.90$

Fig. 5. Average of median suboptimality percentages for policies RAI and RASI using first-order mean-semideviation risk measure.



(a) $\beta = 0.50$ (b) $\beta = 0.75$ (c) $\beta = 0.90$

Fig. 6. Average optimality percentages for policies RAI and RASI using first-order mean-semideviation risk measure.

compared to the RN policy. For example, as depicted in Fig. 3, in a scenario with $\sigma = 1$, $\beta = 0.90$, and $\kappa = 1$, introducing a switching cost of $s = 4$ results in the RASI policy maintaining a high average optimality percentage of 96.22%, while the RN policy shows a significant drop to 73.73%. The RASI policy's robust performance in the face of switching costs highlights its effectiveness in risk-averse environments where switching decisions carry substantial financial weight.

Furthermore, when examining the effects of increasing switching costs, particularly in scenarios with high-cost variability, we observe distinct responses from the RN and RASI policies to changes in risk aversion levels. The RN policy demonstrates a heightened sensitivity to these changes, whereas the RASI policy exhibits a diminishing sensitivity. For example, consider a scenario with $\beta = 0.75$ and $\sigma = 1$. As we adjust the level of risk aversion from 0 to 1, the shift in switching costs from none to moderate (i.e., from $s = 0$ to $s = 2$) leads to a notable divergence in policy performance. The RN policy's average optimality shows an increase in the difference from around 5% to 7.3%. In contrast, the RASI policy's average optimality experiences a

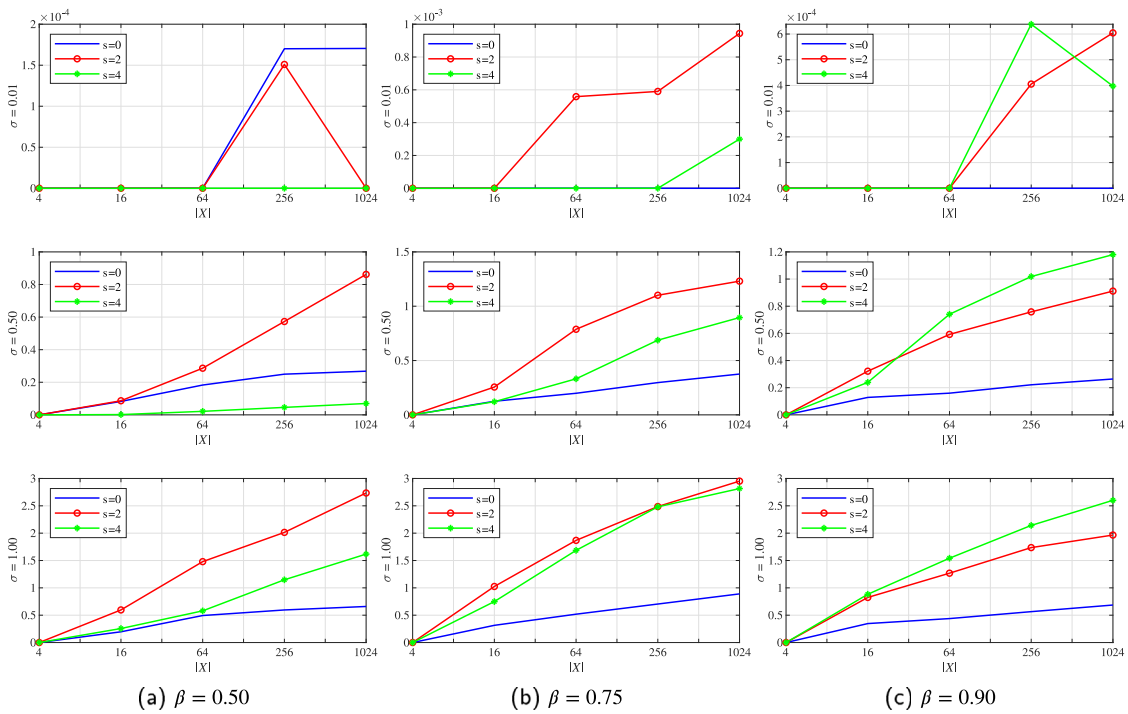


Fig. 7. Average of maximum performance gap percentages for policy RN with respect to RASI policy using first-order mean-semideviation risk measure.

decrease in the difference, moving from approximately 5% to 3.4%. This observation underscores the subtle differences in how each policy responds to variations in risk aversion and switching costs.

In scenarios with high switching costs (i.e., $s = 4$), particularly in risk-averse settings, the RASI policy generally exhibits enhanced performance compared to situations with lower switching costs. This is evident in terms of both the average median suboptimality percentages and the average optimality percentages. It is noteworthy that the RASI policy's least effective performance under high switching costs occurs in conditions of maximum cost variability, discount factor, and risk aversion level (specifically, $\sigma = 1$, $\beta = 0.90$, and $\kappa = 1$). Even in this challenging scenario, the RASI policy maintains a mere 0.04% average in median suboptimality percentages, as illustrated in Fig. 2, and achieves an impressive 96.22% average optimality percentage, as shown in Fig. 3. These findings reinforce the notion that the RASI policy not only excels in risk-averse environments but also tends to recommend the optimal or near-optimal action of maintaining the current choice, especially when faced with significant switching costs.

5.2.2. Evaluating RASI against RAI

In the context of our analysis, the performance of the RAI policy, when compared to the more robust RN and RASI policies, is notably weaker across various settings. This observation holds across different levels of cost variability, switching costs, risk aversion parameters, and discount factors. Specifically, the adaptability of the RAI policy to fluctuating cost environments diminishes as cost variability increases, a trend that is particularly pronounced when comparing low ($\sigma = 0.01$) to moderate ($\sigma = 0.50$) cost variability scenarios (see Fig. 4). This issue of adaptability is further exacerbated by the introduction of switching costs, where the RAI policy's effectiveness in managing the trade-offs between staying and switching actions under uncertainty is significantly challenged.

Moreover, our analysis, as detailed in Figs. 4–6, reveals that the influence of risk aversion on the RAI policy's performance is noteworthy. With increasing κ, indicating a higher aversion to risk, the RAI policy's performance tends to decline, suggesting that it struggles more than the RN and RASI policies to balance the risk-return trade-off, especially in environments characterized by high-cost variability and switching costs. Additionally, the discount factor plays a crucial role in shaping the RAI policy's performance dynamics. Higher discount factors, which place greater emphasis on future rewards, tend to improve the RAI policy's performance.

In summary, while the RAI policy provides a baseline for risk-averse decision-making, its performance is significantly outmatched by the RN and RASI policies across a range of scenarios. The RAI policy's challenges in managing switching costs effectively highlight its limitations as a strategy for risk-averse decision-making with noticeable switching penalties.

5.2.3. Concluding remarks on rasi's performance

Overall, our experiments highlight the RASI policy's superior performance in risk-averse MAB settings, especially under conditions of high-cost variability and significant switching costs. The RASI policy's consistent outperformance of the RN and RAI policies in these challenging scenarios highlights its potential as a more effective strategy for risk-averse decision-making in real-world applications.

5.3. Extended numerical experiments

To gain deeper insights into the practical applicability of our proposed policies in larger state spaces, we conduct additional experiments on MAB problems with increased complexity. In this extended setup, we evaluate the RN, RAI, and RASI policies on MAB instances with the range of one to five arms, each comprising four states. This results in MDPs with the number of states ranging from 4 to 1024. Transition probabilities, costs, and switching costs are generated as described in Section 5.1. We maintain the same discount factors but fix the risk aversion parameter κ to 1. Moreover, due to the impracticality of computing the optimal policy for the larger instances, we use the "performance gap percentage" for a policy with respect to RASI policy. This percentage evaluates how far a policy $\Pi \in \{\mathrm{RN}, \mathrm{RAI}\}$ deviates from the RASI policy as a reference point and is computed as $100 \times (R^{\mathrm{RASI}}(x) - R^{\Pi}(x))/R^{\mathrm{RASI}}(x)$ for each initial state $x \in \mathcal{X}$. Figs. 7 and 8 present the average maximum performance gap percentage for RN and

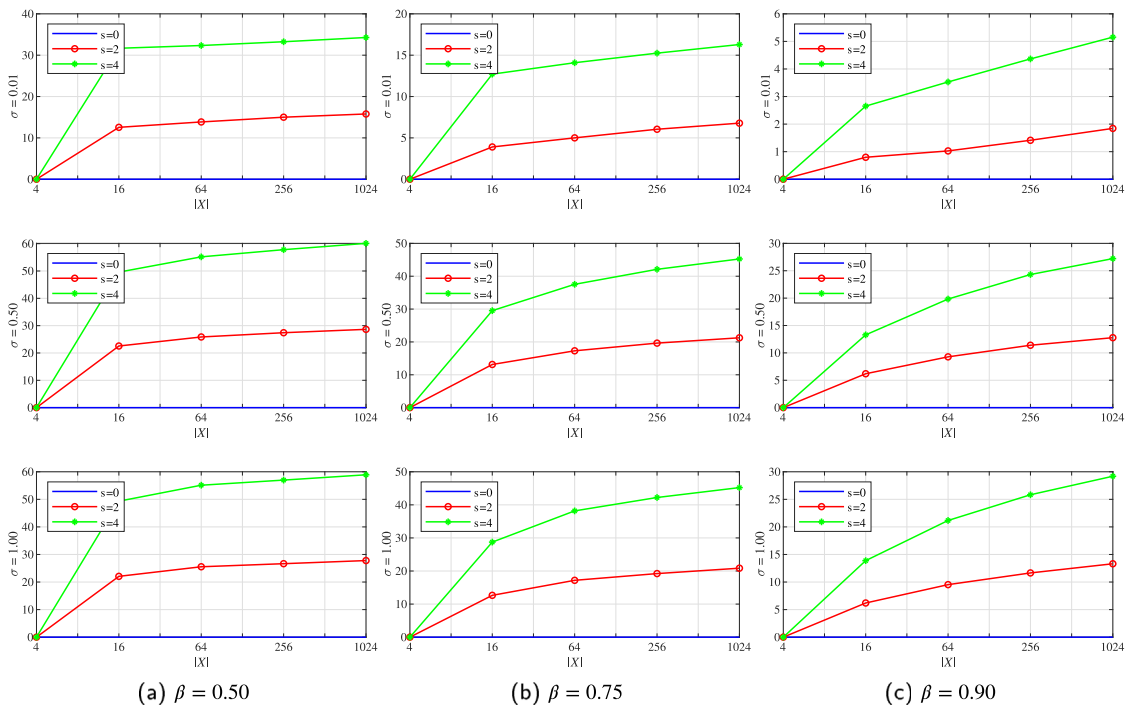


Fig. 8. Average of maximum suboptimality percentages for policy RAI with respect to RASI policy using first-order mean-semideviation risk measure.

RAI policies with respect to RASI, respectively, for varying numbers of arms and different discount factors. These figures aggregate results from 200 test instances for each configuration.

From these results, we observe several key trends. In scenarios with low-cost variability ($\sigma = 0.01$), RN policy demonstrates minimal performance gaps, suggesting that in environments where costs are predictable, the benefit of risk-aversion is less notable. However, as the cost variability increases, the RASI policy consistently outperforms the RN policy, particularly for larger instance sizes. This is evident from the increasing performance gap percentages as the number of arms grows, indicating that the RASI policy's consideration of risk provides a substantial advantage in more volatile environments. Specifically, in Fig. 7, the performance gap is more noticeable for higher levels of cost variability and larger instance sizes. For example, with $\sigma = 1$, $\beta = 0.75$, and $s = 2$, the performance gap between RN and RASI increases to 3% as the number of states increases to 1024. Conversely, as shown in Fig. 8, the RAI policy shows a notable deviation from the RASI policy, especially in scenarios with higher cost variability and larger instance sizes. Higher switching costs ($s = 2$ and $s = 4$) show a more significant gap, highlighting the RAI policy's difficulty in managing these costs as effectively as the RASI policy. These findings suggest that the relevance of risk-aversion remains significant even as the instance size increases, particularly in high variability settings. The extended experiments highlight the robustness and adaptability of the RASI policy in the face of increasing complexity due to a larger number of states.

6. Conclusions

This study explores the complexities of risk-averse MAB problems, emphasizing the influence of switching penalties. By integrating risk considerations and addressing the challenge of switching costs, we provide a comprehensive framework that reflects the complexities of real-world decision-making scenarios in diverse domains. We introduce the RASI policy, which effectively addresses these dual challenges. The RASI policy provides two sets of indices: one for arms that are immediately played and another for the remaining arms, allowing for a refined approach to decision-making. Despite the added complexity of computing these dual indices, the RASI policy remains computationally efficient, leveraging dynamic risk measures to balance risk and switching costs effectively.

Through extensive numerical experiments, the RASI policy demonstrates superior performance, particularly in scenarios characterized by high-cost variability and significant switching costs. The resilience of the RASI policy in navigating these complexities highlights its potential applicability in various real-world settings where switching decisions carry substantial financial implications. Our experiments also reveal that the relevance of risk aversion does not diminish with larger state spaces; instead, it becomes more critical, especially in high variability environments.

As future research, one promising direction is the development of more sophisticated algorithms that further optimize decision-making in risk-averse MAB problems with switching penalties. The other direction would be to investigate different risk measures and their impact on the performance of MAB algorithms. Additionally, an interesting area for future exploration is the application of our methodologies to the general setting of restless bandits. By extending our approach to restless bandits, one can uncover novel strategies and insights for managing the dynamic complexities inherent in these environments, thereby broadening the scope of risk-aware decision-making in operations research.

CRediT authorship contribution statement

Milad Malekipirbazari: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Acknowledgments

This research is supported by the Swedish National Science Foundation Project "Information, Fairness and Socially Beneficial Artificial Intelligence". Omitted proofs and extended details of the numerical results are available in the Supplementary Materials.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejor.2024.09.023.

References

Amir, I., Azov, G., Koren, T., & Livni, R. (2022). Better best of both worlds bounds for bandits with switching costs. *Advances in Neural Information Processing Systems*, *35*, 15800–15810.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203–228.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., & Ku, H. (2007). Coherent multiperiod risk adjusted values and bellman's principle. *Annals of Operations Research*, *152*(1), 5–22.

Asawa, M., & Teneketzis, D. (1994). Multi-armed bandits with switching costs. In *Decision and control, 1994. proceedings of the 33rd IEEE conference on, volume 1* (pp. 168–173). IEEE.

Asawa, M., & Teneketzis, D. (1996). Multi-armed bandits with switching penalties. *IEEE Transactions on Automatic Control*, *41*(3), 328–348.

Banks, J., & Sundaram, R. (1994). Switching costs and the gittins index. *Econometrica: Journal of the Econometric Society*, *68*, 7–694.

Bertsimas, D., & Mersereau, A. (2007). A learning approach for interactive marketing to a customer segment. *Operations Research*, *55*(6), 1120–1135.

Calafiore, G., & Dabbene, F. (2006). *Probabilistic and randomized methods for design under uncertainty*. Springer.

Caro, F., & Gallien, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, *53*(2), 276–292.

Chancelier, J.-P., De Lara, M., & De Palma, A. (2007). Risk aversion, road choice, and the one-armed bandit problem. *Transportation Science*, *41*(1), 1–14.

Cheridito, P., Delbaen, F., & Kupper, M. (2006). Dynamic monetary risk measures for bounded discrete-time processes. *Electronic Journal of Probability*, *11*, 57–106.

Cohen, S., & Treetanthiploet, T. (2019). Gittins' theorem under uncertainty. arXiv preprint arXiv:1907.05689.

Denardo, E., Feinberg, E., & Rothblum, U. (2013). The multi-armed bandit, with constraints. *Annals of Operations Research*, *208*(1), 37–62.

Denardo, E., Park, H., & Rothblum, U. (2007). Risk-sensitive and risk-neutral multiarmed bandits. *Mathematics of Operations Research*, *32*(2), 374–394.

El Karoui, N., & Karatzas, I. (1994). Dynamic allocation problems in continuous time. *The Annals of Applied Probability*, 255–286.

Föllmer, H., & Penner, I. (2006). Convex risk measures and the dynamics of their penalty functions. *Statistics & Decisions*, *24*(1/2006), 61–96.

Gittins, J. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *41*(2), 148–164.

Gittins, J., & Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. *Progress in Statistics*, 241–266.

Grechuk, B., & Zabarankin, M. (2016). Inverse portfolio problem with coherent risk measures. *European Journal of Operational Research*, *249*(2), 740–750.

Jun, T. (2004). A survey on the bandit problem with switching costs. *de Economist*, *152*, 513–541.

Kumar, U., & Saranga, H. (2010). Optimal selection of obsolescence mitigation strategies using a restless bandit model. *European Journal of Operational Research*, *200*(1), 170–180.

Malekipirbazari, M., & Çavuş, Ö. (2021). Risk-averse allocation indices for multiarmed bandit problem. *IEEE Transactions on Automatic Control*, *66*(11), 5522–5529.

Malekipirbazari, M., & Çavuş, Ö. (2024). Index policy for multiarmed bandit problem with dynamic risk measures. *European Journal of Operational Research*, *312*(2), 627–640.

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy*, *60*(2), 151–158.

Mwai, N., Malekipirbazari, M., & Johansson, F. (2024). Batched fixed-confidence pure exploration for bandits with switching constraints. In *ICML 2024 workshop: aligning reinforcement learning experimentalists and theorists*.

Niño-Mora, J. (2008). A faster index algorithm and a computational study for bandits with switching costs. *INFORMS Journal on Computing*, *20*(2), 255–269.

Niño-Mora, J. (2010). Computing an index policy for bandits with switching penalties. In *1st international ICST workshop on tools for solving structured Markov chains*.

Ogryczak, W., & Ruszczyński, A. (1999). From stochastic dominance to mean-risk models: Semideviations as risk measures. *European Journal of Operational Research*, *116*(1), 33–50.

Ogryczak, W., & Ruszczyński, A. (2001). On consistency of stochastic dominance and mean–semideviation models. *Mathematical Programming*, *89*(2), 217–232.

Ogryczak, W., & Ruszczyński, A. (2002). Dual stochastic dominance and related mean-risk models. *SIAM Journal on Optimization*, *13*(1), 60–78.

Powell, W. (2019). A unified framework for stochastic optimization. *European Journal of Operational Research*, *275*(3), 795–821.

Riedel, F. (2004). Dynamic coherent risk measures. *Stochastic Processes and their Applications*, *112*(2), 185–200.

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

Rouyer, C., Seldin, Y., & Cesa-Bianchi, N. (2021). An algorithm for stochastic and adversarial bandits with switching costs. In *International conference on machine learning* (pp. 9127–9135). PMLR.

Ruszczyński, A. (2010). Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, *125*(2), 235–261.

Ruszczyński, A., & Shapiro, A. (2006a). Conditional risk mappings. *Mathematics of Operations Research*, *31*(3), 544–561.

Ruszczyński, A., & Shapiro, A. (2006b). Optimization of convex risk functions. *Mathematics of Operations Research*, *31*(3), 433–452.

Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2009). *Lectures on stochastic programming: modeling and theory*. SIAM.

Van Oyen, M., Pandelis, D., & Teneketzis, D. (1992). Optimality of index policies for stochastic scheduling with switching penalties. *Journal of Applied Probability*, *29*(4), 957–966.

Villar, S., Bowden, J., & Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *30*(2), 199.

Washburn, R. (2008). Application of multi-armed bandits to sensor management. In *Foundations and applications of sensor management* (pp. 153–175). Springer.

Weber, R. (1992). On the Gittins index for multiarmed bandits. *The Annals of Applied Probability*, 1024–1033.

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 143–149.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, *25*(A), 287–298.