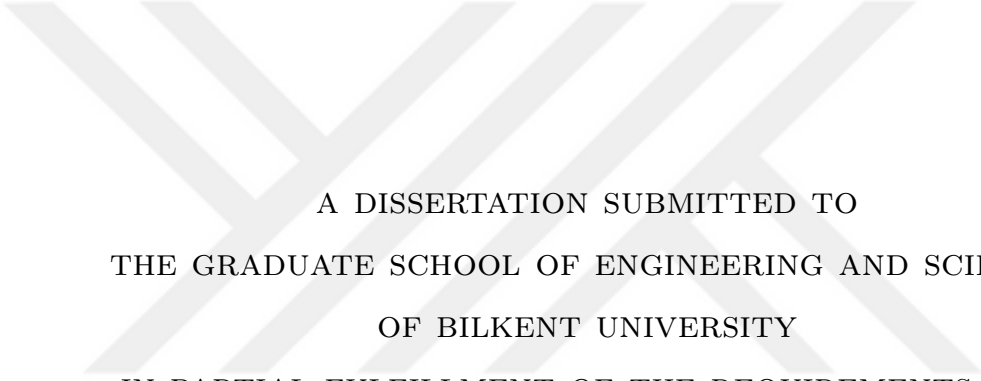


RISK-AVERSE MULTI-ARMED BANDIT PROBLEM



A DISSERTATION SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
INDUSTRIAL ENGINEERING

By
Milad Malekipirbazari
August 2021

RISK-AVERSE MULTI-ARMED BANDIT PROBLEM

By Milad Malekipirbazari

August 2021

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Özlem Çavuş İyigün(Advisor)

Sakine Batun

Çağın Ararat

Firdevs Ulus

Ali Devin Sezer

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

RISK-AVERSE MULTI-ARMED BANDIT PROBLEM

Milad Malekipirbazari
Ph.D. in Industrial Engineering
Advisor: Özlem Çavuş İyigün
August 2021

In classical multi-armed bandit problem, the aim is to find a policy maximizing the expected total reward, implicitly assuming that the decision maker is risk-neutral. On the other hand, the decision makers are risk-averse in some real life applications. In this study, we design a new setting for the classical multi-armed bandit problem (MAB) based on the concept of dynamic risk measures, where the aim is to find a policy with the best risk adjusted total discounted outcome. We provide theoretical analysis of MAB with respect to this novel setting, and propose two different priority-index heuristics giving risk-averse allocation indices with structures similar to Gittins index. The first proposed heuristic is based on Lagrangian duality and the indices are expressed as the Lagrangian multiplier corresponding to the activation constraint. In the second part, we present a theoretical analysis based on Whittle's retirement problem and propose a generalized version of restart-in-state formulation of the Gittins index to compute the proposed risk-averse allocation indices. Finally, as a practical application of the proposed methods, we focus on optimal design of clinical trials and we apply our risk-averse MAB approach to perform risk-averse treatment allocation based on a Bayesian Bernoulli model. We evaluate the performance of our approach against other allocation rules, including fixed randomization.

Keywords: Multi-armed bandit, Gittins index, dynamic risk-aversion, coherent risk measures, Markov decision process, clinical trials.

ÖZET

RİSKTEN KAÇINAN ÇOK KOLLU HAYDUT PROBLEM

Milad Malekipirbazarı
Endüstri Mühendisliği, Doktora
Tez Danışmanı: Özlem Çavuş İyigün
Ağustos 2021

Klasik çok kollu haydut probleminde amaç, karar vericinin dolaylı olarak riske duyarsız olduğu varsayımıyla, beklenen ödülü maksimize eden bir politika bulmaktır. Öte yandan, karar vericiler bazı gerçek yaşam uygulamalarında riskten kaçınırlar. Bu çalışmada, koşullu dinamik risk ölçütlerini kullanarak klasik çok kollu haydut problemleri (ÇKH) için, amacın riske göre ayarlanmış en iyi indirimli toplam çıktıyı veren bir politika bulmak olduğu, yeni bir yaklaşım öneriyoruz. Bu yeni yaklaşıma göre ÇKH'nın teorik analizini yapıyoruz ve Gittins endeksine benzer yapılara sahip riskten kaçınan tahsis endeksleri veren iki farklı öncelik endeksi öneriyoruz. Önerilen ilk yöntem, Lagrange çifteliğine dayanmakta ve endeksler aktivasyon kısıtına karşılık gelen Lagrange çarpanı olarak ifade edilmektedir. İkinci bölümde, Whittle'ın emeklilik sorununa dayanan teorik bir analiz sunuyoruz ve önerilen riskten kaçınan tahsis endekslerini hesaplamak için Gittins endeksinin başlangıç durumundan yeniden başlatma formülasyonunun genelleştirilmiş bir versiyonunu öneriyoruz. Son olarak, önerilen yöntemlerin pratik bir uygulaması olarak, klinik deneylerin optimal tasarımına odaklanıyoruz ve Bayesian Bernoulli modeline dayalı olarak riskten kaçınan tedavi tahsisini gerçekleştirmek için riskten kaçınan ÇKH yaklaşımımızı uyguluyoruz. Yaklaşımımızın performansını, sabit rastgeleleştirme dahil olmak üzere diğer tahsis kurallarına göre değerlendiriyoruz.

Anahtar sözcükler: Çok kollu haydut, Gittins endeksi, dinamik riskten kaçınma, tutarlı risk ölçütleri, Markov karar süreci, klinik denemeler.

Acknowledgement

Foremost, I would like to express my deepest gratitude to my advisor Asst. Prof. Özlem Çavuş İyigün for her guidance and encouragement throughout my study. I am forever grateful to her for believing in me and allowing me to work in a respectful and supportive environment. Without her invaluable and continuous support and understanding this dissertation would not have been possible.

I would like to express my gratitude to Prof. Selim Aktürk, our department chair, for trusting me and supporting me every step of the way, and allowing me to teach during my final year of Ph.D. studies. My sincere thanks also goes to Prof. Cem İyigün, Prof. Oya Karaşan, and Assoc. Prof. Vural Aksakallı for their unwavering support and guidance over these years.

I am grateful to Asst. Prof. Sakine Batun and Asst. Prof. Çağın Ararat for reading each part of this dissertation and providing precious suggestions during my study. I also want to express my gratitude to Asst. Prof. Firdevs Ulus and Prof. Ali Devin Sezer for accepting to be a member of my examination committee and for their valuable suggestions.

I am grateful to so many people for their unlimited support to me. First, I would like to thank my dear friends Kamyar Kargar, Meltem Peker, Muhammad Umer, and Nihal Berktaş for their endless support and encouragement. I also would like to thank Amir Mozammel Hokmabadi, Çağla Fatma Dursunoğlu, Deniz Emre, Farzad Shams, and Halil İbrahim Bayrak for their companionship during my graduate life. It was because of their friendship that my studies and life in Bilkent were so enjoyable.

Last but not least, I would like to thank Parinaz for always standing by my side, supporting me all through and ensuring me that I was set up for success no matter what. I would like to thank my mother Narges and my father Hormoz. I could not succeed without their endless love, encouragement, and support.

Contents

1	Introduction	1
2	Literature Review	5
2.1	Stochastic Multi-armed Bandit Problem	5
2.2	Risk-Aversion in Stochastic MAB	7
2.3	Markov Multi-armed Bandit Problem	11
2.3.1	Gittins Index	12
2.3.2	Calculation of Gittins Index	13
2.4	Risk-Aversion in Markovian MAB	16
2.5	Summary	17
3	Risk-averse Allocation Indices for Multi-armed Bandit Problem	19
3.1	Problem Description and Preliminaries	20
3.2	Relaxation and Decomposition	24
3.3	Indexability and Indices	27
3.4	Index Policy for Risk-averse One-armed Bandit Problem	32
3.5	Computation of the Risk-averse Indices	35
3.6	Computational Experiments	36
3.7	Conclusion	39
4	Index Policy for Multi-armed Bandit Problem with Dynamic Risk Measures	43
4.1	Problem Description and Preliminaries	44
4.2	Risk-averse Retirement Problem	49
4.3	Indexability and Indices	51
4.4	Computation of the Risk-averse Indices	58

4.5	Computational Experiments	61
4.6	Conclusion	68
5	Multi-armed Bandit Designs with Dynamic Risk Measures for Clinical Trials	71
5.1	Problem Description and Preliminaries	71
5.2	An Index Policy for Risk-averse MAB	77
5.3	Computation of Risk-averse Sequential Allocation Rules in Clinical Trials	78
5.4	Numerical Study	87
5.4.1	Allocation Index Values	88
5.4.2	Simulation Study	89
5.5	Conclusion	92
6	Conclusion	94
A	Allocation index values for risk measures of first-order mean-semideviation and mean-AVaR	103

List of Figures

3.1	Run time of obtaining the optimal policy and index policy for first-order mean-semideviation	42
4.1	Runtime of obtaining the optimal policy via solving risk-averse MDP along with executing risk-averse index heuristic versus the size of the risk-averse MAB for first-order mean-semideviation risk measure.	67
4.2	Runtime of obtaining the optimal policy via solving risk-averse MDP along with executing risk-averse index heuristic versus the size of the risk-averse MAB for Mean-AVaR risk measure.	68
5.1	Markov chain process of arm i	72
5.2	The (approximate) Gittins index values for $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures.	89
5.3	The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for first-order mean-semideviation risk measure with different parameters of risk aversion.	90
5.4	The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for mean-AVaR risk measure with different parameters of risk aversion.	91

List of Tables

3.1	Statistics of maximum suboptimality percentage of policies RN and RA for first-order mean-semideviation	40
3.2	Average similarity percentages of policies RN and RA for first-order mean-semideviation	40
3.3	Statistics of maximum suboptimality percentage of policies RN and RA for mean-AVaR	41
3.4	Average similarity percentages of policies RN and RA for mean-AVaR	41
4.1	Statistics of maximum suboptimality percentage of policies <i>RN</i> and <i>RA</i> for first-order mean-semideviation	63
4.2	Statistics of median suboptimality percentage of policies <i>RN</i> and <i>RA</i> for first-order mean-semideviation	64
4.3	Average similarity percentages of policies <i>RN</i> and <i>RA</i> for first-order mean-semideviation	64
4.4	Statistics of maximum suboptimality percentage of policies <i>RN</i> and <i>RA</i> for mean-AVaR	65
4.5	Statistics of median suboptimality percentage of policies <i>RN</i> and <i>RA</i> for mean-AVaR	66
4.6	Average similarity percentages of policies <i>RN</i> and <i>RA</i> for mean-AVaR	66
4.7	Description of problem sizes	67
4.8	Comparison of maximum suboptimality percentage of policies RA1 and RA2 for first-order mean-semideviation	69

4.9	Average similarity percentages of policies RA1 and RA2 for first-order mean-semideviation	69
4.10	Comparison of maximum suboptimality percentage of policies RA1 and RA2 for mean-AVaR	70
4.11	Average similarity percentages of policies RA1 and RA2 for mean-AVaR	70
5.1	Performance comparison of different designs for BB-MAB.	93
A.1	The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for first-order mean-semideviation risk measure where $\beta = 0.50$	104
A.2	The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for first-order mean-semideviation risk measure where $\beta = 0.75$	105
A.3	The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for first-order mean-semideviation risk measure where $\beta = 0.90$	106
A.4	The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for mean-AVaR risk measure where $\beta = 0.50$	107
A.5	The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for mean-AVaR risk measure where $\beta = 0.75$	108
A.6	The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for mean-AVaR risk measure where $\beta = 0.90$	109

Chapter 1

Introduction

The multi-armed bandit problem (MAB) is a problem in probability theory, in which a gambler is faced with a line of slot machines and must plot which machine to play at each step. MAB is a term that refers to any problem, in which there is a trade-off between resource exploration and exploitation. This problem is a formulation for an agent that is attempting to obtain new knowledge while optimizing their decisions based on existing information. In other words, MAB is the process of optimizing the policy return over time as we learn it. This problem arises in many practical applications, such as clinical trials, portfolio design, tuning parameters, internet advertisement, auction mechanisms, adaptive routing in networks, and project management.

Clearly, an optimal solution to the risk-neutral MAB could be inherently risky. However, in many applications of MAB, like clinical trials, we cannot afford high variability in the results even if it would yield the highest expected outcome. Motivated by this, in this work, we look for risk-averse models that can capture the preferences of a risk-averse decision maker. One way of incorporating the risk into the model is by using some utility functions. However, utility functions have been the subject of considerable criticism over the years. The greatest burden of this approach is that the decision maker needs to specify an appropriate utility function with respect to the degree of risk-aversion, which is not always

an easy task to perform. It is also argued that it is impossible to measure utility across individuals objectively. Moreover, the resulting solutions may be difficult to interpret. In order to handle these difficulties, we propose a different approach for risk incorporation by using coherent risk measures.

In this thesis, we incorporate risk into MAB with the help of dynamic risk measures and aim to find a policy with the best risk-adjusted total discounted outcome. However, similar to the risk-neutral case, the size of MAB increases exponentially as the number of arms increases, therefore solving large scale MAB with risk-averse dynamic programming approaches would be computationally expensive. In this regard, in this thesis, we introduce index-based solution heuristics for risk-averse multi-armed bandit models. We provide theoretical analysis of MAB with respect to this novel setting, and propose two different priority-index heuristics giving risk-averse allocation indices with a similar structure to Gittins index. We also provide computational experiments to evaluate the quality of the proposed methods. Empirical results express the excellence of these heuristics and show that with risk-averse allocation indices we can achieve optimal or near-optimal interpretable policies. Moreover, we focus on optimal design of clinical trials and apply our risk-averse MAB approach to perform risk-averse treatment allocation based on a Bayesian Bernoulli model.

The rest of the thesis is organized as follows. In Chapter 2, we present the related literature on risk-averse multi-armed bandit problems. In Chapter 3, we propose risk-averse allocation indices based on Lagrangian duality. These indices are expressed as the Lagrangian multiplier corresponding to the activation constraint. To achieve this, we first provide the characterization of an indexable and decomposable problem close to risk-averse MAB. Then, we introduce a definition of indices to obtain an optimal solution to this new problem. Afterwards, we present an algorithmic approach to efficiently compute these indices. At the end, we computationally test the performance of our index policy using two famous coherent risk measures of first-order mean-semideviation and mean-AVaR. These experiments indicate the excellence of our index policy in all evaluated criteria. Based on our results, the proposed index policy provides much lower suboptimality percentages and it is more similar to the optimal policy compared to the

policy derived from Gittins index. Our experiments comparing the computational complexity of our index heuristic to solving the corresponding MDP process reveals that our method is time efficient and the computation time grows linearly with the problem size, whereas the risk-averse MDP computation time shows an exponential growth.

In Chapter 4, we present a theoretical analysis based on Whittle’s retirement problem, where we also allow retiring from the process at each step with a fixed cost. We then propose a priority-index policy based on this retirement cost that reduces to Gittins index when the level of risk-aversion converges to zero. We generalize the restart-in-state formulation of the Gittins index to effectively compute these risk-averse allocation indices. Our comprehensive numerical experiments reveal that the efficiency of our risk-averse index policy can be very close to the optimal policy. Furthermore, our risk-averse allocation indices can achieve optimal or near-optimal policies which, in some instances, are easier to interpret compared to the optimal policy.

In Chapter 5, we focus on the design of clinical trials and we apply our risk-averse MAB approach to perform risk-averse treatment allocation based on a Bayesian Bernoulli model. In this study, each arm represents a treatment with an unknown success rate, each decision step is an incoming patient who will receive a treatment, and consequently each play will be assigning a treatment to a patient. At each time, playing an arm is considered as drawing a sample observation from a Bernoulli population with unknown success probability, with which we earn a reward that can be either 0 or 1. The Bayesian feature here is incorporated by allowing each unknown success probability to have a Beta prior. We evaluate the performance of our approach against other allocation rules, including fixed randomization. Based on our results, as the decision maker’s risk-aversion level increases, the decision maker will be more inclined to exploit rather than explore, which consequently would provide us with lower variability of the results compared to a risk-neutral setting.

Chapter 6 contains the concluding remarks and a discussion of the thesis’s contribution to the existing literature. In Chapter 6, we also discuss future research

directions and possible extensions of present work in more detail.



Chapter 2

Literature Review

In general, there are two different approaches towards multi-armed bandits. In the former approach, following [1], the aim is to minimize the rate of regret for the expected total reward. In the latter, following [2], we aim to maximize the expected discounted reward over an infinite horizon. In this chapter, we focus on these two approaches, and try to have a comprehensive survey in their risk-averse settings.

2.1 Stochastic Multi-armed Bandit Problem

In this problem, in each round, the gambler iteratively pulls one of the K arms in the infinite horizon or finite horizon of T rounds and monitors the associated reward X_{ks} which is the s^{th} i.i.d. random sample of arm k . The rewards delivered by each one of the arms can be considered coming from a set of real distributions (i.e., $X_{ks} \sim \nu_k$ for $k = 1, \dots, K$). Each one of these distributions has the mean value of μ_k and variance of σ_k^2 , and may be bounded in an interval (i.e., $X_{ks} \in [l, u]$).

While all arms and samples are considered independent, the most well-known performance measure for the multi-armed bandit problem is the total expected

regret. The expected regret over t rounds $\mathcal{R}_t$ is described as the expected difference between the total reward associated with the best arm and the total amount of the collected rewards:

$$\mathcal{R}_t = t\mu_{k^*} - \sum_{k=1}^K \mu_k n_{kt}, \quad (2.1)$$

where μ_{k^*} is the maximum reward mean, i.e. $\mu_{k^*} = \max_k \{\mu_k\}$, and n_{kt} denotes the number of times the k th arm has been selected up to time t .

Since in sequential learning the selection of arms should be performed on the basis of the empirical estimates, empirical mean and variance of arm k after t rounds are defined as stated in equations (2.2) and (2.3), respectively.

$$\hat{\mu}_{kt} = \frac{1}{n_{kt}} \sum_{s=1}^{n_{kt}} X_{ks}, \quad (2.2)$$

$$\hat{\sigma}_{kt}^2 = \frac{1}{n_{kt}} \sum_{s=1}^{n_{kt}} (X_{ks} - \hat{\mu}_{kt})^2. \quad (2.3)$$

Moreover, upon these measures, the empirical expected regret $\hat{\mathcal{R}}_t$ is introduced as

$$\hat{\mathcal{R}}_t = t\hat{\mu}_{k^*t} - \sum_{k=1}^K \hat{\mu}_{kt} n_{kt}, \quad (2.4)$$

where $\hat{\mu}_{k^*t} = \max_k \{\hat{\mu}_{kt}\}$.

A strategy in which the average regret per round ($\mathcal{R}_t/t$) approaches to zero while the number of rounds the gambler plays tends to infinity is called a *zero-regret strategy*. Intuitively, this kind of strategy is guaranteed to converge to an optimal strategy assuming that enough rounds are played. Regret grows at least logarithmically, i.e., $\mathcal{R}_t = O(\log t)$. And checking this lower bound is the criteria usually employed to check whether an algorithm can solve this problem or not. A statistical analysis and overview of bandit problems are offered in [3].

2.2 Risk-Aversion in Stochastic MAB

Although there exists a large literature on online learning, especially on the algorithms proposed for solving the bandits problem with guaranteed worst-case regret bounds, there has been limited attention to the risk incurred by such algorithms. One reason is that in the analysis of risk, we frequently assume we know the distributions of the underlying uncertainty, while in bandit models, this information may be unavailable. The other reason is that the complexity of risk estimation in sequential decision making rises notably as the number of decisions increases.

One of the first articles considering risk in online learning is the work presented in [4]. They try to only minimize the risk ignoring the expectation of the rewards. In that work, they present two algorithms for online variance minimization by assuming that in each step a covariance matrix is revealed. Ignoring the expectation of the rewards, mainly in applications related to finance, makes the results relatively immature for real problems. To fill this gap, authors in [5] consider the same problem in settings, in which there is a trade-off between rewards and risk by exploiting two accepted measures: the Sharpe ratio [6] and the mean-variance criterion [7]. A description of this criterion is presented in Definition 2.2.1.

Definition 2.2.1. The mean-variance criterion (MV): The mean-variance of arm k with coefficient of absolute risk tolerance φ is estimated as:

$$MV_k = \sigma_k^2 - \varphi \mu_k. \quad (2.5)$$

In [8], the authors try to model the MAB problem by competing against the arm with the best risk-return trade-off. The measure used in that analysis is the popular MV model. The optimal arm therefore is the arm with the minimum mean-variance. And, the empirical mean-variance measure for arm k after t rounds is defined as:

$$\widehat{MV}_{kt} = \hat{\sigma}_{kt}^2 - \varphi \hat{\mu}_{kt}. \quad (2.6)$$

They introduced two algorithms to find a feasible policy, MV-LCB and ExpExp. In MV-LCB, the aim is to minimize the MV of cumulative regret. In each step,

the effort is to select the arm which minimizes the following equation:

$$\widehat{MV}_{kt} - (5 + \varphi) \sqrt{\frac{\log \frac{1}{\delta}}{2n_{kt}}}, \quad (2.7)$$

where δ is a parameter regulated by the time horizon T . They conclude that this selection rule results in a small regret of order $O(K \frac{\log(T)}{T})$ for easy problems (i.e., where the mean-variance gaps are big with respect to T), and a constant worst-case regret. In ExpExp algorithm, the strategy is to launch all arms uniformly in the exploration phase, and select the arm with the minimum empirical MV in the exploitation phase. In this case, the regret bound of $O(KT^{-\frac{1}{3}})$ is achieved if we fix the exploration length to $K(\frac{T}{14})^{\frac{2}{3}}$. At the end, the authors confirm their theoretical results by numerical simulations. Detailed proofs and further experiments are addressed in [9].

In [10], inspired by [8], the authors consider the problem of searching for the maximally risk-averse arm. Firstly, they provide a different interpretation of the regret, which is claimed to be more natural and easier to interpret, where it is just defined based on the variability of reward distributions, and helps to control the tail of the regret:

$$-\frac{1}{\lambda} \log \left(\mathbb{E}(e^{-\lambda X_{ks}}) \right). \quad (2.8)$$

They justify this risk measure by arguing that risk-aversion is generally avoiding too low rewards. So, there should be a control on the tails, especially on the lower-tail. Afterwards, they propose an algorithm, called *RA-UCB*, together with a high probability bound on the resulted regret.

In [11], it is first argued that incorporating risk in machine learning problems is a vital task. Then, they investigate the way this incorporation may affect the underlying theory with regard to the classical expected-reward setting. They analyze the complexity gap between the single-period and the multi-period versions of risk-averse bandit problems. They also show that the multi-period version has a high complexity even with the independence assumption between the arms. Finally, as their major contribution, they propose an arm-selection algorithm (the CuRisk algorithm). In their algorithms, they consider risk minimization by some risk notions like value-at-risk (VaR), a well-known risk measure widely used in

finance. Definition 2.2.2 provides a technical description of this risk measure.

Definition 2.2.2. Value at Risk (VaR): Having α as the confidence level ($0 < \alpha < 1$) and X representing the reward, $VaR_\alpha(X)$ is the smallest number x such that the probability that the underlying X exceeds $-x$ is at least α :

$$VaR_\alpha(X) = \inf\{x \in \mathbb{R} : \mathbb{P}(X + x < 0) \leq 1 - \alpha\} = \inf\{x \in \mathbb{R} : 1 - F_X(-x) \geq \alpha\}.$$

In [12], the authors propose the multi-armed risk-aware bandit (MaRaB) algorithm with the aim of improving the exploration phase of previous algorithms by restricting the exploration of risky arms. MaRaB selects an arm with maximal *conditional value-at-risk* (CVaR). Definition 2.2.3 provides a technical description of this risk measure.

Definition 2.2.3. Conditional value-at-risk (CVaR): CVaR considers the quantiles of the distribution for each reward. Defining α as the target quantile level ($0 < \alpha < 1$), the associated quantile value v_α is the value such that $\mathbb{P}(X < v_\alpha) = \alpha$, where X represents the random rewards with continuous distribution. The conditional value-at-risk α , denoted by $CVaR_\alpha$, is then defined as $CVaR_\alpha[X] = \mathbb{E}[X|X < v_\alpha]$.

Then, by defining the estimate of the conditional value-at-risk α of arm i as

$$\widehat{CVaR}_{\alpha k} = \frac{1}{\lceil \alpha n_{kt} \rceil} \sum_{s=1}^{\lceil \alpha n_{kt} \rceil} X_{ks} \quad (2.9)$$

in each step, the effort is to select the arm which minimizes the following equation:

$$\widehat{CVaR}_{\alpha k} - C \sqrt{\frac{\log(\lceil t\alpha \rceil)}{\lceil \alpha n_{kt} \rceil}}, \quad (2.10)$$

where C is a positive parameter controlling the exploration vs exploitation trade-off. When the risk level approaches to 0, MaRaB is almost equivalent to the MIN multi-armed bandit algorithm, in which the arm with maximal minimal value is selected. It is shown that the MIN algorithm attains logarithmic regret under mild assumptions [13]. Furthermore, with empirical validation, they demonstrate

that, comparing to the algorithms presented in [8], MaRaB has less exploration of the arms with low distribution tails.

Authors in [14] also study risk-averse multi-armed bandits under mean-variance measures. They propose two risk mitigation models targeted for different applications. The first model takes the risk as the variations in the reward values of different times and aims to minimize the mean-variance of the observed rewards. This arrangement is useful for some applications like clinical trials where it is advantageous to avoid high variations in the treatment outcomes of different patients. This risk-averse algorithm, referred to as MV-UCB, at each round t , plays the arm with the smallest amount of

$$\widehat{MV}_{k,t} - C\sqrt{\frac{\log t}{n_{k,t}}}. \quad (2.11)$$

They show that in this way a regret order of $O(\log T)$ is achieved in contrast to the $O(\sqrt{T})$ regret order given in [8]. On the other hand, the second model, referred to as MV-DSEE, attempts to minimize the mean-variance of the total reward at the end of the time horizon, which is of particular interest in economic and financial applications. In this policy, total rounds of play is divided into an exploration sequence and an exploitation sequence. In the exploration sequence, the player plays all arms in a round-robin fashion. In the exploitation sequence, the player plays the arm which has the smallest sample mean-variance. Under both models, they establish asymptotic and finite-time lower bounds on regret and propose online learning algorithms that attain the lower bounds.

In [15], a third risk mitigation model to solve risk-averse MAB problem is presented. In this model, same as the second model in [14], the player is only focused on the total reward at the end of the time horizon (rather than each individual reward value achieved throughout the process). The aim is to maximize the VaR of the total reward. This model is more helpful in financial applications like investments where we desire a high probability guarantee on the outcome of the investment.

As discussed, the previous works try to define a proper measure of risk based on the intended application and analyze the corresponding regret accordingly. This

is because there is no universally agreed notion of a good risk measure, and the correct notion varies from one problem to another. In this way, the applicability of the results would become restricted to a specific application. That's why, authors in [16] characterize the risk-averse MAB with arbitrary (but fixed) risk measure and the corresponding regret. The risk measure is defined as a function of the mean and the variance (generalization of risk measure in [8]). They consider the notion of regret similar to [10], and first propose an algorithm, called φ -LCB, and prove logarithmic risk-averse regret bounds for all continuous functions. In this algorithm, in each step, we select the arm with the minimum value of

$$f(\hat{\mu}_{k,n_{k(t-1)}}, \hat{\sigma}_{k,n_{k(t-1)}}) - \varphi\left(6\sqrt{\frac{\ln \frac{1}{\delta}}{2n_{k(t-1)}}}\right), \quad (2.12)$$

where φ is a continuous function with specific characterizations. Finally, they propose the second algorithm, namely, φ -LCB2, that makes learning feasible with the mild assumption that no arm hits the discontinuity points.

Authors in [17] address the issue of risk in MAB and develop some results under the measure of mean-variance. They show that the model-specific regret as well as the model-independent regret are lower bounded by $O(\log T)$ and $O(T^{2/3})$, respectively. Then, they show *MV-UCB* and *MV-DSEE* policies achieve these lower bounds.

2.3 Markov Multi-armed Bandit Problem

There exists another formulation of MAB in which each arm represents an independent Markov machine. Whenever a specific arm is played, the state of that arm changes to a new one, chosen with respect to the Markov state transition probabilities, earning an income (possibly negative) which depends on the current state of the arm and sometimes the state to which transition takes place. The arms that are not played maintain their current states. Each step of this process contains one transition, and the typical goal is to maximize the expectation of the discounted total reward with discount factor $\beta \in (0, 1)$.

With K available arms and considering arm $i \in \{1, 2, \dots, K\}$ as a Markov process with a finite space $\mathcal{X}^i$, X_t^i is the state of arm i at time t , where $t \in \{0, 1, 2, \dots\}$. Let $u_t = (u_t^1, \dots, u_t^K)$ denote the decision that the gambler takes at time t . Then u_t^i is equal to

$$u_t^i = \begin{cases} 1 & \text{gambler plays arm } i \\ 0 & \text{otherwise,} \end{cases}$$

which means u_t has just one non-zero element. In short, whenever $u_t^i = 1$, X_t^i evolves in a Markovian manner with respect to the transition probabilities $P_{mn}^i(1) = \mathbb{P}(x_{t+1}^i = n \mid x_t^i = m, u_t^i = 1)$, $m, n \in \mathcal{X}^i$, otherwise it remains frozen, that is, for all $m, n \in \mathcal{X}^i$ we have

$$P_{mn}^i(0) = \mathbb{P}(x_{t+1}^i = n \mid x_t^i = m, u_t^i = 0) = \begin{cases} 1, & \text{if } m = n \\ 0, & \text{otherwise.} \end{cases}$$

When the gambler plays arm i , he receives a reward $r^i(X_t^i)$, and as described the aim is to choose a decision strategy π in order to maximize the expected total discounted reward:

$$\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \beta^t \sum_{i=1}^K r^i(X_t^i) u_t^i \mid X_0 = x_0 \right], \quad (2.13)$$

where $x_0 = (x_0^1, \dots, x_0^K)$ is the initial state of the process.

2.3.1 Gittins Index

In order to solve MAB with Markovian structure, we may consider the problem as a Markov decision process (MDP) and apply Markov decision theory to find the best policy. However, the size of bandit problem increases exponentially as the number of arms increases, therefore solving large-scale MABs would be computationally challenging. In this regard, Gittins [18] provides a decomposition for the problem, by proposing an index policy approach, where these indices are called *dynamic allocation indices*, afterwards referred to as *Gittins indices*. Regarding the specific setting of that problem, he shows that an index can be

allocated to each state with respect to the data of the arm so that, in each step, by playing the arm having the highest index, the maximum expected total discounted reward is obtained. These indices, which afterwards are referred to as *Gittins indices*, for arm i and state $x^i \in \mathcal{X}^i$ can be computed as:

$$\nu^i(x^i) = \max_{\tau > 0} \frac{\mathbb{E} \left[\sum_{t=0}^{\tau} \beta^t r^i(X_t) | X_0^i = x^i \right]}{\mathbb{E} \left[\sum_{t=0}^{\tau} \beta^t | X_0^i = x^i \right]}, \quad (2.14)$$

where τ is the stopping time.

Generalizations and different proofs of this key idea are discussed vastly in [19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]. These generalizations usually are based on new frameworks proposed as the consequence of relaxing some assumptions provided in the original study [18]. For instance, Whittle [31] introduces *Whittle indices* to solve restless multi-armed bandit problem in which states of the arms not being played also change. He expresses this index as the Lagrangian multiplier corresponding to the activation constraint, which specifies the number of arms activated at each step.

2.3.2 Calculation of Gittins Index

Algorithms to compute the Gittins index can be classified into two groups of *offline* and *online*. The latter is more complex and more convenient for bandits with large or infinite state spaces.

One of the properties of Gittins index is that it induces a total order on $\mathcal{X}^i$, i.e., for $a, b \in \mathcal{X}^i$ as two distinct states of arm i , we have $a \preceq b \iff \nu^i(a) \leq \nu^i(b)$. Offline algorithms exploit this property and split the state space into two disjoint sets of *continuation set* (C) and *stopping set* (S):

$$C(a) = \{b \in \mathcal{X}^i : a \prec b\},$$

$$S(a) = \{b \in \mathcal{X}^i : b \preceq a\}.$$

As this terminology suggests, once the gambler is playing arm i in state a it would be optimal to continue playing in all states of $C(a)$ since $\nu^i(a) < \nu^i(b)$ for all $b \in C(a)$. Similarly, the stopping time which corresponds to the initial state a is referred to the moment the process hits any state which belongs to $S(a)$ at time $T(S(a))$. Employing this characterization, Gittins index for each state of arm i is represented as:

$$\nu^i(x^i) = \max_{S(x^i) \subseteq X^i} \frac{\mathbb{E} \left[\sum_{t=0}^{T(S(x^i))} \beta^t r^i(X_t) | X_0^i = x^i \right]}{\mathbb{E} \left[\sum_{t=0}^{T(S(x^i))} \beta^t | X_0^i = x^i \right]}. \quad (2.15)$$

Two well-known techniques for computing Gittins index are presented here. The first is an offline method called *largest-remaining-index algorithm*, while the second is an online method called *restart formulation*.

2.3.2.1 Largest-Remaining-Index Algorithm

Introduced in [32], this method is based on finding the Gittins indices according to the decreasing order of states in $\mathcal{X}^i$. Let α_j denote the state which is the j^{th} largest Gittins index in arm i , then we have:

$$\alpha_1 \succeq \alpha_2 \succeq \cdots \succeq \alpha_{|\mathcal{X}^i|}.$$

Let I denote the identity matrix. Algorithm 1 describes the computation of Gittins indices for arm i with one step transition probability matrix $P^i(1)$ and reward r^i . In this algorithm, first we identify the state with the highest Gittins index which is the state with the highest reward. The Gittins index of this state is equal to the reward that we receive by visiting this state. Then, in order to find the state with the second highest index value, we construct the continuation set as $C(\alpha_2) = \{\alpha_1\}$ and define the $|\mathcal{X}^i| \times |\mathcal{X}^i|$ matrix $Q^{(2)}$ by $\forall a, b \in \mathcal{X}^i$:

$$Q_{ab}^{(2)} := \begin{cases} P_{ab}^i(1), & \text{if } b \in C(\alpha_2) \\ 0, & \text{otherwise,} \end{cases}$$

Algorithm 1 Largest-Remaining-Index Algorithm

```
1: Initialization  $\alpha_1 = \arg \max_{a \in \mathcal{X}^i} r_a$  with  $\nu(\alpha_1) = r_{\alpha_1}$ 
2: for  $j = 2$  to  $|\mathcal{X}^i|$  do
3:   for  $a, b \in \mathcal{X}^i$  do
4:     if  $b \in C(\alpha_j)$  then
5:        $Q_{ab}^{(j)} = P_{ab}^i(1)$ 
6:     else
7:        $Q_{ab}^{(j)} = 0$ 
8:     end if
9:   end for
10:   $d^{(j)} = [I - \beta Q^{(j)}]^{-1} r^i$ 
11:   $b^{(j)} = [I - \beta Q^{(j)}]^{-1} \mathbf{1}$ 
12:   $\alpha_j = \arg \max_{a \in C(\alpha_j)} \{d_a^{(j)} / b_a^{(j)}\}$ 
13:   $\nu(\alpha_j) = d_{\alpha_j}^{(j)} / b_{\alpha_j}^{(j)}$ 
14: end for
15: return  $\nu$ 
```

and define the $|\mathcal{X}^i| \times 1$ vectors:

$$d^{(2)} = [I - \beta Q^{(2)}]^{-1} r^i,$$

$$b^{(2)} = [I - \beta Q^{(2)}]^{-1} \mathbf{1}.$$

Then, we find the second highest index state by $\alpha_2 = \arg \max_{a \in C(\alpha_2)} \{d_a^{(2)} / b_a^{(2)}\}$, and the corresponding index value as $\nu(\alpha_2) = \{d_{\alpha_2}^{(2)} / b_{\alpha_2}^{(2)}\}$. Afterwards, the next highest index is computed by repeating the same computation on a new continuation set obtained by adding α_2 to the previous continuation set, that is, $C(\alpha_3) = \{\alpha_1, \alpha_2\}$. The next steps employ this idea and progress in similar fashion.

2.3.2.2 Restart Formulation

Introduced in [23] and based on the approach given in [19], *restart formulation* is an online approach for computing Gittins indices. As illustrated in [19], Gittins indices can be computed by introducing a new process. In this process, there is an extra option, which is at any step the gambler can receive a retirement reward and quit the game. With this approach, Gittins index of state $a \in \mathcal{X}^i$

Algorithm 2 Restart Formulation

```
1: for  $j = 1$  to  $|\mathcal{X}^i|$  do
2:   Restart Option:
3:      $r^{(0)} = r^i(j)\mathbf{1}$ 
4:     for  $x, y \in \mathcal{X}^i$  do
5:        $Q_{x,y}^{(0)} = P_{jy}^i(1)$ 
6:     end for
7:   Continuation Option:
8:      $r^{(1)} = r^i$ 
9:      $Q^{(1)} = P^i(1)$ 
10:  Dynamic Program:
11:     $v = \max \{r^{(0)} + \beta Q^{(0)}v, r^{(1)} + \beta Q^{(1)}v\}$ 
12:     $\nu(j) = (1 - \beta)v_j$ 
13: end for
14: return  $\nu$ 
```

will be the retirement reward that the gambler is indifferent between receiving it or continuing the game when a is the initial state. In [23], they show that it is possible to consider this retirement reward and restarting in that state as two equivalent options. So, we now have a restart problem by substituting these two equivalent options, and the problem is the gambler has two options of continue the bandit process or instantaneously start in state a . According to this process, Gittins indices can be computed as described in Algorithm 2.

2.4 Risk-Aversion in Markovian MAB

Contrary to the risk-neutral version, there are few works on the incorporation of risk in Markov multi-armed bandits. Authors of [33] study risk-aversion in route choosing, modeling it as a one-armed bandit problem under different information regimes. The choice is between a random or a safe selection. Risk-neutral travelers tend to choose the risky route, while risk-averse travelers choose the safe route with a higher frequency. In this research, the risk preference is incorporated in the model using utility functions. With the same perspective towards risk incorporation, same authors in [34] consider a situation where a decision maker is dealing with an uncertain event, however, she can costly change her knowledge of

the uncertainties. They focus on the effect of risk-aversion on the decision maker's search behavior. They model the problem as a one-armed bandit (one risky and one safe arm), and show how risk-aversion changes the sequence of choosing arms. As a result, they confirm that the more concave the utility function, the more chance exists for the decision maker to choose the safe arm.

The most detailed and theoretical study in this area is performed by [35]. They introduce the concept of risk in the Markovian setting with an analytic approach employing concave utility functions. In that setting, stopping times are replaced with a simple recursion to allocate to state k a utility $\bar{r}(k)$ and an amplification $a(k)$ of future utility that helps us to perform pairwise comparison: state k is said to dominate state j if

$$\bar{r}(k) + a(k)\bar{r}(j) > \bar{r}(j) + a(j)\bar{r}(k). \quad (2.16)$$

Regarding this definition, they introduce a ranking of the states of the arms. They illustrate that it is optimal to play the arm with the highest rank state. They perform a comprehensible analysis that encloses three models; first with risk-averse exponential utility, then with risk-seeking exponential utility, and finally with linear utility. More specifically, for risk-averse case, they define the utility function as $u(x) = -e^{-\lambda x}$, and construct the index of state k in arm i as:

$$\rho(k) = \frac{1 - a(k)}{\bar{r}(k)} = \frac{1 - \sum_{j \in \mathcal{X}^i} P_{kj}^i(1) e^{-\lambda r^i(k,j)}}{-P_{kT}^i(1) e^{-\lambda r^i(k,T)}}, \quad (2.17)$$

where T is the state representing termination of play and $r^i(k, j)$ is the reward of transitioning from state k to j . Improvements and modification of this approach are presented in [36].

2.5 Summary

As described above, unlike stochastic MAB, existing studies on selection of the arm with the best risk-return trade-off in Markovian MAB is limited to incorporating the risk by using some utility functions. However, utility functions have

been the subject of considerable criticism over the years. The greatest burden of this approach is that the decision maker needs to specify an appropriate utility function with respect to the degree of risk-aversion, which is not always an easy task to perform. It is also argued that it is impossible to measure utility across individuals objectively. Moreover, the resulting solutions may be difficult to interpret [37]. In order to handle these difficulties, in this thesis, we propose a different approach for risk incorporation by using coherent risk measures [38]. To model our problem, we employ the idea of the study [39], where discounted risk-averse discrete-time Markov models with infinite horizon are formulated using dynamic coherent measures of risk, and value and policy iteration algorithms are proposed to find an optimal policy.

Chapter 3

Risk-averse Allocation Indices for Multi-armed Bandit Problem

In this chapter, we design a new setting based on the concept of dynamic risk measures where the aim is to find a policy with the best risk-adjusted total discounted outcome. We provide a theoretical analysis of multi-armed bandit problem with respect to this novel setting, and propose a priority-index heuristic which gives *risk-averse allocation indices* having a structure similar to Gittins index. Although an optimal policy is shown not always to have index-based form, empirical results express the excellence of this heuristic and show that with *risk-averse allocation indices* we can achieve optimal or near-optimal interpretable policies.

This chapter aims to develop the classical work of Gittins [18] and provide index-based solutions for risk-averse MAB with dynamic coherent risk measures. In this regard, we seek to achieve the following: (1) the characterization of an indexable and decomposable problem close to risk-averse MAB, (2) definition of indices to obtain an optimal solution to this new problem, (3) an algorithm to compute these indices, (4) an experimental approach to check how good these indices are for the original problem. To the best of our knowledge, this version of MAB is new, and no study exists on this setting.

The results of this chapter are published in IEEE Transactions on Automatic Control [40].

3.1 Problem Description and Preliminaries

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

- (i) With the assumption that there exists no common state in distinct arms, let $\mathcal{X} = \times_{i=1}^K \mathcal{X}^i$ be the *state space* of the resulting MDP, where each arm i is a Markov chain with a finite state space $\mathcal{X}^i$, $i \in \mathcal{K} = \{1, 2, \dots, K\}$.
- (ii) At step $t \in \mathbb{N}$, an *action* $u_t = \{u_t^1, u_t^2, \dots, u_t^K\}$ is applied, where $u_t^i \in \{-1, 0\}$ is the action applied to arm $i \in \mathcal{K}$. Here, $u_t^i = 0$ denotes that arm i is played at step t , on the other hand, $u_t^i = -1$ represents that arm i is not played. The *admissible actions* are the ones for which exactly one element of u_t equals to zero at each step $t \in \mathbb{N}$. That is, at each step, under any choice of admissible actions, the state of one and only one arm evolves independently from the states of other arms in a Markovian manner, and the states of the remaining arms are frozen.
- (iii) The state of the arm we play at each step changes in a Markovian fashion with respect to the transition probabilities $P_{mn}^i(0) = \mathbb{P}(x_{t+1}^i = n \mid x_t^i = m, u_t^i = 0)$, $i \in \mathcal{K}$, $m, n \in \mathcal{X}^i$, $t \in \mathbb{N}$. However, the state of an arm remains unchanged if it is not played, that is, for all $i \in \mathcal{K}$, $m, n \in \mathcal{X}^i$, $t \in \mathbb{N}$ we have

$$P_{mn}^i(-1) = \mathbb{P}(x_{t+1}^i = n \mid x_t^i = m, u_t^i = -1) = \begin{cases} 1, & \text{if } m = n \\ 0, & \text{otherwise.} \end{cases}$$

(iv) For each $i \in \mathcal{K}$, the *cost function* $g^i : \mathcal{X}^i \rightarrow \mathbb{R}_-$ is the finite and non-positive cost of playing arm i at state x^i . We define $c^i(x^i, u^i)$ as the cost incurred by arm i at state $x^i \in \mathcal{X}^i$ under action $u^i \in \{0, -1\}$, $i \in \mathcal{K}$:

$$c^i(x^i, u^i) := \begin{cases} g^i(x^i), & \text{if } u^i = 0 \\ 0, & \text{if } u^i = -1. \end{cases} \quad (3.1)$$

We also define $c(x, u)$ representing the cost over all arms, that is $c(x, u) = \sum_{i \in \mathcal{K}} c^i(x^i, u^i)$, $x = (x^1, \dots, x^K) \in \mathcal{X}$, and $u = (u^1, \dots, u^K)$ where $u^i \in \{-1, 0\}$, $i \in \mathcal{K}$.

(v) We denote a stationary (Markov) *policy* as $\pi : \mathcal{X} \rightarrow \{0, -1\}^K$. Note that, with respect to our notation, $u_t = \pi(x_t)$, that is, both denote the decision to be taken at state $x_t \in \mathcal{X}$ for $t \in \mathbb{N}$. We also define $\pi^i : \mathcal{X}^i \rightarrow \{0, -1\}$ which prescribes the action to be taken for arm $i \in \mathcal{K}$ at a particular state, that is, $\pi(x_t) = (\pi^1(x_t^1), \pi^2(x_t^2), \dots, \pi^K(x_t^K))$.

Here, we look for a stationary Markov policy in order to minimize the total risk-averse discounted cost incurred over an infinite horizon. With respect to the theory of [39], an infinite horizon stationary MDP with dynamic risk measures has a stationary optimal policy and it is possible to find this policy using *value iteration* or *policy iteration* algorithms. However, these algorithms may be computationally expensive or even practically infeasible for large-scale real-life problems. Furthermore, in some cases, the structure of optimal policy makes it difficult to interpret, which creates complication while applying them in real-life problems. Motivated to overcome these issues, we seek good index-based heuristic policies.

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

In this context, $\rho_t : \mathcal{Z}_{t+1} \rightarrow \mathcal{Z}_t$, $t \in \{1, \dots, T\}$ satisfying the following axioms is defined as a *one-step conditional risk measure* (see [41, 42]):

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

Conditions (A1)-(A4) are called *convexity*, *monotonicity*, *translation invariance* and *positive homogeneity*, respectively. They are analogous to the axioms introduced by [38] for coherent measures of risk.

Two important coherent risk measures, which have been extensively studied in the literature, are *first-order mean-semideviation* [43] and *mean-average value-at-risk* (mean-AVaR) [44]. We now provide the conditional versions of these two risk measures. Conditional first-order mean-semideviation risk measure is defined as

$$\rho_t(Z_{t+1}) = \mathbb{E}[Z_{t+1} | \mathcal{F}_t] + \kappa \mathbb{E}[(Z_{t+1} - \mathbb{E}[Z_{t+1} | \mathcal{F}_t])_+ | \mathcal{F}_t], \quad (3.2)$$

where $\kappa \in [0, 1]$ and $(a)_+ := \max\{a, 0\}$ for $a \in \mathbb{R}$. Given two parameters of $\alpha \in (0, 1)$ and $\lambda \in [0, 1]$, conditional mean-AVaR risk measure is defined as

$$\rho_t(Z_{t+1}) = \lambda \mathbb{E}[Z_{t+1} | \mathcal{F}_t] + (1 - \lambda) \text{AVaR}_\alpha(Z_{t+1} | \mathcal{F}_t), \quad (3.3)$$

where conditional AVaR_α can be represented as

$$\text{AVaR}_\alpha(Z_{t+1} | \mathcal{F}_t) = \min_{\eta \in \mathcal{Z}_t} \left\{ \eta + \frac{1}{1 - \alpha} \mathbb{E}[(Z_{t+1} - \eta)_+ | \mathcal{F}_t] \right\}.$$

A *dynamic risk measure* is a sequence of one-step conditional risk measures [39, and references therein]. Thereby, dynamic risk measure $\varrho_{1,T}^\beta : \mathcal{Z}_{1,T} \mapsto \mathcal{Z}_1$ on a finite horizon of length T with discount factor $\beta \in (0, 1)$ is defined as follows:

$$\varrho_{1,T}^\beta(Z_1, Z_2, \dots, Z_T) := Z_1 + \rho_1(\beta Z_2 + \rho_2(\beta^2 Z_3 + \dots + \rho_{T-1}(\beta^{T-1} Z_T) \dots)). \quad (3.4)$$

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

$$\varrho^\beta(Z_1, Z_2, Z_3, \dots) = \lim_{T \rightarrow \infty} \varrho_{1,T}^\beta(Z_1, Z_2, \dots, Z_T). \quad (3.5)$$

Applying the dynamic risk measure (3.5), we can now evaluate the risk of the cost sequences $c(x_t, u_t) \in \mathcal{Z}_t$, $t \in \mathbb{N}$ resulted by policy π . Let $R(x_1)$ be the optimal risk-averse total discounted cost under the constraint that one arm will be played at each step:

$$\begin{aligned} R(x_1) &= \min_{\pi \in \Pi} \varrho^\beta(c(x_1, u_1), c(x_2, u_2), \dots) \\ &= \min_{\pi \in \Pi} \sum_{i \in \mathcal{K}} c^i(x_1^i, u_1^i) + \rho_1 \left(\beta \sum_{i \in \mathcal{K}} c^i(x_2^i, u_2^i) + \rho_2 \left(\beta^2 \sum_{i \in \mathcal{K}} c^i(x_3^i, u_3^i) + \dots \right) \right), \end{aligned} \quad (3.6)$$

where Π is the class of stationary admissible policies for our problem, that is, at each step we are only allowed to play one arm. Note that second equality is due to definition of $c(x_t, u_t)$. Since the additivity of risk measure is one main assumption in [18] to guarantee the optimality of an index policy, it follows that index policies may not be optimal for the risk-averse problem, as coherent risk measures are subadditive. Therefore, optimal policy can only be achieved by performing the far more difficult task of solving the risk-averse dynamic programming equations. At this point, following the classic works of [18] and [31], and based on the theory and presentation of [45], we aim to find a heuristic solution of the problem in the form of index policies. Therefore, in order to find proper actions in each state, we need to design a calibrating function for each arm that can provide an index value for each state of the arm [18, 31, 45]. To come up with such indices, we need to perform a decomposition of our problem into K separate problems, one for each arm. And, with the help of these separate problems, we can acquire the appropriate calibrating functions.

3.2 Relaxation and Decomposition

With the aim of attaining an index heuristic for problem (3.6), we obtain an equivalent optimization problem by extending Π to the class Π' of stationary policies, which allows a free choice of action, i.e., no arm or more than one arm can also be played at a step:

$$R(x_1) = \min_{\pi \in \Pi'} \sum_{i \in \mathcal{K}} c^i(x_1^i, u_1^i) + \rho_1 \left(\beta \sum_{i \in \mathcal{K}} c^i(x_2^i, u_2^i) + \rho_2 \left(\beta^2 \sum_{i \in \mathcal{K}} c^i(x_3^i, u_3^i) + \dots \right) \right) \quad (3.7)$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{K}} u_t^i = 1 - K, \quad \forall t \in \mathbb{N}, \quad (3.8)$$

$$u_t^i \in \{-1, 0\}, \quad \forall i \in \mathcal{K}, \quad t \in \mathbb{N}. \quad (3.9)$$

Here, constraint (3.8) implies that playing exactly one arm is allowed at each step. Further, for mathematical convenience, we relax this constraint by replacing it with

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

which can also be written as follows using definitions (3.4) and (3.5):

$$\sum_{i \in \mathcal{K}} u_1^i + \rho_1 \left(\beta \sum_{i \in \mathcal{K}} u_2^i + \rho_2 \left(\beta^2 \sum_{i \in \mathcal{K}} u_3^i + \dots \right) \right) = \frac{1 - K}{1 - \beta}. \quad (3.10)$$

Now, replacing constraint (3.8) with (3.10) and relaxing it in Lagrangian manner, we obtain the following Lagrangian dual function:

$$\begin{aligned} \mathcal{L}_D(\nu, x_1) &= \min_{\pi \in \Pi'} \mathcal{L}(\nu, x_1, \pi) \\ \text{s.t.} \quad &u_t^i \in \{-1, 0\}, \quad \forall i \in \mathcal{K}, \quad t \in \mathbb{N}, \end{aligned} \quad (3.11)$$

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

$$\begin{aligned} \mathcal{L}(\nu, x_1, \pi) &= \sum_{i \in \mathcal{K}} c^i(x_1^i, u_1^i) + \rho_1 \left(\beta \sum_{i \in \mathcal{K}} c^i(x_2^i, u_2^i) + \rho_2 \left(\beta^2 \sum_{i \in \mathcal{K}} c^i(x_3^i, u_3^i) + \dots \right) \right) \\ &+ \nu \left(\sum_{i \in \mathcal{K}} u_1^i + \rho_1 \left(\beta \sum_{i \in \mathcal{K}} u_2^i + \rho_2 \left(\beta^2 \sum_{i \in \mathcal{K}} u_3^i + \dots \right) \right) \right) - \nu \left(\frac{1 - K}{1 - \beta} \right). \end{aligned} \quad (3.12)$$

Note that, by duality, we have:

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

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

Lemma 3.2.1. *Consider the dynamic risk measure $\varrho_{1,T}^\beta$ in (3.4) and random costs $Z_t^i \in \mathcal{Z}_t$, $t \in \{1, \dots, T\}$, $i \in \mathcal{K}$. Then we have:*

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

Proof. From axioms (A1) and (A4), we get

$$\rho_{T-1} \left(\beta^{T-1} \sum_{i \in \mathcal{K}} Z_T^i \right) \leq \sum_{i \in \mathcal{K}} \rho_{T-1} (\beta^{T-1} Z_T^i). \quad (3.13)$$

Then, we obtain

$$\begin{aligned} & \rho_{T-2} \left(\beta^{T-2} \sum_{i \in \mathcal{K}} Z_{T-1}^i + \rho_{T-1} \left(\beta^{T-1} \sum_{i \in \mathcal{K}} Z_T^i \right) \right) \\ & \leq \rho_{T-2} \left(\sum_{i \in \mathcal{K}} \left(\beta^{T-2} Z_{T-1}^i + \rho_{T-1} (\beta^{T-1} Z_T^i) \right) \right) \\ & \leq \sum_{i \in \mathcal{K}} \rho_{T-2} \left(\beta^{T-2} Z_{T-1}^i + \rho_{T-1} (\beta^{T-1} Z_T^i) \right), \end{aligned}$$

where the first inequality follows from monotonicity axiom (A2) and inequality (3.13), and the second inequality is due to (A1) and (A4). Iterating similarly down to $t = 1$, the assertion of the lemma follows. $\square$

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

$$\begin{aligned} \mathcal{L}'(\nu, x_1, \pi) = & \sum_{i \in \mathcal{K}} \left(c^i(x_1^i, u_1^i) + \rho_1(\beta c^i(x_2^i, u_2^i) + \rho_2(\beta^2 c^i(x_3^i, u_3^i) + \dots)) + \nu(u_1^i \right. \\ & \left. + \rho_1(\beta u_2^i + \rho_2(\beta^2 u_3^i + \dots))) \right) - \nu \left(\frac{1-K}{1-\beta} \right), \end{aligned} \quad (3.14)$$

where $\nu \in \mathbb{R}_+$ and $x_1 \in \mathcal{X}$. Notice that, for $\nu \in \mathbb{R}_+$, using Lemma 3.2.1 and (3.5), we get

$$\mathcal{L}(\nu, x_1, \pi) \leq \mathcal{L}'(\nu, x_1, \pi). \quad (3.15)$$

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

$$\begin{aligned} \mathcal{L}'_D(\nu, x_1) = & \min_{\pi \in \Pi'} \mathcal{L}'(\nu, x_1, \pi) \\ \text{s.t. } & u_t^i \in \{-1, 0\}, \forall i \in \mathcal{K}, t \in \mathbb{N}, \end{aligned} \quad (3.16)$$

where $\nu \in \mathbb{R}_+$. Note that, from (3.15), for $\nu \in \mathbb{R}_+$, we obtain

$$\mathcal{L}_D(\nu, x_1) \leq \mathcal{L}'_D(\nu, x_1). \quad (3.17)$$

Although $\mathcal{L}_D(\nu, x_1)$ is a lower bound for $R(x_1)$, $\mathcal{L}'_D(\nu, x_1)$ might not be smaller than $R(x_1)$. On the other hand, problem (3.16) provides us with an index-based policy, which is an interpretable feasible policy for the risk-averse MAB problem (3.6). For the rest of the study, we will derive this index-based policy and computationally show that it is close to an optimal policy of (3.6).

Note that problem (3.16) can be decomposed into K subproblems, each of which relates to different arms. For each arm $i \in \mathcal{K}$, we have:

$$\begin{aligned} \mathcal{L}'_{Di}(\nu, x_1^i) = & \min_{\pi^i \in \Pi'^i} c^i(x_1^i, u_1^i) + \rho_1(\beta c^i(x_2^i, u_2^i) + \rho_2(\beta^2 c^i(x_3^i, u_3^i) + \dots)) + \\ & \nu \left(u_1^i + \rho_1(\beta u_2^i + \rho_2(\beta^2 u_3^i + \dots)) \right) \\ \text{s.t. } & u_t^i \in \{-1, 0\}, \forall t \in \mathbb{N}, \end{aligned} \quad (3.18)$$

where Π'^i is a class of stationary policies for arm i which has no restriction regarding the action to be taken at each step. Also, since the control variables don't appear in the term $\nu(\frac{1-K}{1-\beta})$ in (3.14), it was dropped in (3.18) without changing the optimal policies.

In the next section, we discuss the indexability of subproblem (3.18) and explain the structure of the indices by developing a calibrating function for each arm.

3.3 Indexability and Indices

In this step, in search of the structure of optimal policy for subproblem (3.18), we need to follow [31] and [45]. Regarding Whittle's discussion in [31], the indexability argument is related with the class of bandit problems centering on individual arms. Therefore, we focus our attention on subproblem (3.18), which we call as arm i , and provide below a lemma describing a key property of its optimal policy.

Lemma 3.3.1. *Consider the subproblem (3.18). For $i \in \mathcal{K}$, if action “not play” is optimal for state $x_t^i \in \mathcal{X}^i$, then it is also optimal for state $x_{t+1}^i \in \mathcal{X}^i$, $\forall t \in \mathbb{N}$.*

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

Now, in order to find an optimal policy of (3.18), we will consider the evolution of arm $i \in \mathcal{K}$ under some policy $\pi^i \in \Pi^i$ starting from the initial state x_1^i , and from Lemma 3.3.1, we will restrict our attention to two specific stationary policies of never playing as well as playing the arm from the beginning until some random time $\tau^i > 1$, which for brevity will be referred to as τ^i policy. To do so, we need to find a proper calibration function $\nu^i : \mathcal{X}^i \rightarrow \mathbb{R}_+$ for arm $i \in \mathcal{K}$. The index $\nu^i(x^i)$ of arm i when in state $x^i \in \mathcal{X}^i$ is defined as the value of ν which makes two actions of “play” and “not play” equally attractive.

However, for $\nu^i(x^i)$ to be meaningful, it requires to induce a consistent ordering of the states, such that if arm i being at state x^i is not played under $\nu^i(x^i)$, it will also not be played under $\nu > \nu^i(x^i)$. For this, we define Θ_ν^i as the set of states of arm i , for which the action is “not play” for a given value of ν . According to [31], arm $i \in \mathcal{K}$ is indexable if there exists a family of policies in which each policy, corresponding to a specific $\nu \in \mathbb{R}_+$, is optimal for subproblem (3.18) and the cardinality of Θ_ν^i is non-decreasing as ν increases. Under these conditions,

the corresponding index would be

$$\nu^i(x^i) = \inf\{\nu \in \mathbb{R}_+ : x^i \in \Theta_\nu^i\}. \quad (3.19)$$

As an interpretation of (3.19), $\nu^i(x^i)$ is a fair charge we pay to go from “not play” to “play” when arm i is in state x^i . For any $\nu \in \mathbb{R}_+$, if $\nu \leq \nu^i(x^i)$, then playing arm i is optimal, otherwise the optimal action will be not playing it. In this case, for a given $\nu \in \mathbb{R}_+$, the stopping time τ^i can be computed as

$$\tau^i = \inf\{t \geq 2 : x_t^i \in \Theta_\nu^i\}. \quad (3.20)$$

From Lemma 3.3.1 and using (3.20), we have $x_t^i \in \Theta_\nu^i, \forall t \geq \tau^i$. Therefore, finding an optimal policy boils down to searching for some stopping time, which would be obtained by employing the indices described in Definition 3.3.1. Here, with a slight abuse of notation and referring to (6) in [39], we consider $\varrho_{1,\tau-1}^\beta(Z_1, Z_2, \dots, Z_{\tau-1}) = \varrho^\beta(Z_1, Z_2, \dots, Z_{\tau-1}, 0, 0, \dots)$.

Definition 3.3.1. For $i \in \mathcal{K}$, the *risk-averse allocation index* (RAI) for each initial state $x_1^i \in \mathcal{X}^i$ is given by:

$$\nu^i(x_1^i) := \sup_{\tau^i \geq 2} \frac{\varrho_{1,\tau^i-1}^\beta(c^i(x_1^i, 0), c^i(x_2^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0))}{\varrho_{1,\tau^i-1}^\beta(-1, -1, \dots, -1)}. \quad (3.21)$$

Note that based on (3.19) and (3.20), the stopping time in which we attain the supremum in (3.21) is $\tau^{i*} = \inf\{t \geq 2 : x_t^i \in \Theta_{\nu^i(x_1^i)}^i\}$.

Lemma 3.3.2. *The risk-averse allocation index of Definition 3.3.1 is guaranteed to be a finite term.*

Proof. For $i \in \mathcal{K}$, let $c_{min}^i := \min_{x \in \mathcal{X}^i} c^i(x, 0)$. Multiplying both sides of (3.21) with $-1/c_{min}^i$, we will have

$$\begin{aligned} \frac{\nu^i(x_1^i)}{-c_{min}^i} &= \sup_{\tau^i \geq 2} \frac{\varrho_{1,\tau^i-1}^\beta(c^i(x_1^i, 0), c^i(x_2^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0))}{-c_{min}^i \varrho_{1,\tau^i-1}^\beta(-1, -1, \dots, -1)} \\ &= \sup_{\tau^i \geq 2} \frac{\varrho_{1,\tau^i-1}^\beta(c^i(x_1^i, 0), c^i(x_2^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0))}{\varrho_{1,\tau^i-1}^\beta(c_{min}^i, c_{min}^i, \dots, c_{min}^i)} \\ &\leq 1, \end{aligned}$$

where the second equality follows from axiom (A4) and the inequality is the result of monotonicity property (A2). So, $\nu^i(x_1^i) \leq -c_{min}^i$, and with respect to the bounded costs in our setting, we conclude that this index is guaranteed to be finite. $\square$

Remark. The *risk-averse indices* of Definition 3.3.1 have the similar structure and interpretation as the Gittins index, and may be considered as a generalized form of Gittins index that is capable of taking dynamic risk-aversion into account.

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

Theorem 3.3.1. *Each risk-averse arm $i \in \mathcal{K}$ is indexable with the risk-averse allocation index introduced in Definition 3.3.1.*

Proof. Consider the subproblem (3.18) for arm $i \in \mathcal{K}$ and for a given constant $\nu \in \mathbb{R}_+$. Given the previous discussion, we have two policies to consider: either playing the arm until some stopping time or not playing it at all. Playing this arm from the initial state $x_1^i \in \mathcal{X}^i$ until stopping time $\tau^i \in \mathcal{T}^i$ results in the following objective function value:

$$\varrho^\beta(c^i(x_1^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0), 0, 0, \dots) + \nu \varrho^\beta(\underbrace{0, 0, \dots, 0}_{\tau^i-1}, -1, -1, \dots), \quad (3.22)$$

for which referring to (6) in [39] we have

$$\varrho^\beta(c^i(x_1^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0), 0, 0, \dots) = \varrho_{1, \tau^i-1}^\beta(c^i(x_1^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0)).$$

Similarly, if the arm is not played in state x_1^i , then from Lemma 3.3.1, the objective function value is

$$\varrho^\beta(0, 0, \dots) + \nu \varrho^\beta(-1, -1, \dots), \quad (3.23)$$

where $\varrho^\beta(0, 0, \dots) = 0$ as a result of axiom (A4). Therefore, the “play” action is strictly optimal for x_1^i when there exists a stopping policy $\tau^i \in \mathcal{T}^i$ for which the

quantity in (3.22) falls behind the quantity in (3.23), that is:

$$\varrho_{1,\tau^i-1}^\beta(c^i(x_1^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0)) + \nu \varrho^\beta(\underbrace{0, \dots, 0}_{\tau^i-1}, -1, -1, \dots) < \nu \varrho^\beta(-1, -1, \dots). \quad (3.24)$$

Organizing this inequality, we obtain

$$\begin{aligned} \varrho_{1,\tau^i-1}^\beta(c^i(x_1^i, 0), c^i(x_2^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0)) \\ < \nu(\varrho^\beta(-1, -1, \dots) - \varrho^\beta(\underbrace{0, \dots, 0}_{\tau^i-1}, -1, -1, \dots)) \\ = \nu\left(-\sum_{t=0}^{\tau^i-2} \beta^t\right) = \nu \varrho_{1,\tau^i-1}^\beta(-1, -1, \dots, -1), \end{aligned}$$

which results in

$$\varrho_{1,\tau^i-1}^\beta(c^i(x_1^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0)) < \nu \varrho_{1,\tau^i-1}^\beta(-1, \dots, -1).$$

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

$$\frac{\varrho_{1,\tau^i-1}^\beta(c^i(x_1^i, 0), c^i(x_2^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0))}{\varrho_{1,\tau^i-1}^\beta(-1, -1, \dots, -1)} > \nu,$$

and, from equation (3.21), we have $\nu^i(x_1^i) > \nu$.

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

$$\frac{\varrho_{1,\tau^i-1}^\beta(c^i(x_1^i, 0), c^i(x_2^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0))}{\varrho_{1,\tau^i-1}^\beta(-1, -1, \dots, -1)} \leq \nu^i(x_1^i) < \nu.$$

In conclusion, action “play” is optimal in state x_1^i if and only if $\nu^i(x_1^i) \geq \nu$. Also, action “not play” is optimal in this state if and only if $\nu^i(x_1^i) \leq \nu$. From this argument, we can infer the existence of a family of optimal policies whose associated inactive sets Θ_ν^i , $i \in \mathcal{K}$ are nondecreasing in ν . This argument establishes the indexability of each arm $i \in \mathcal{K}$ with respect to Definition 3.3.1. $\square$

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

Proposition 3.3.1. *For each arm $i \in \mathcal{K}$, starting from $x_1^i \in \mathcal{X}^i$, a positive stopping time $\tau^i \in \mathcal{T}^i$ achieves the supremum in Definition 3.3.1.*

Proof. For a fixed non-negative value of ν , with respect to the previous proof, we know that it is optimal to play at step t if and only if $\nu^i(x_t^i) \geq \nu$. It is also optimal not to play if and only if $\nu^i(x_t^i) \leq \nu$. Particularly, it is optimal to both play and not to play at step one if $\nu^i(x_1^i) = \nu$. If we take action “play” at step one, then it would be optimal to continue taking this action before step τ^i if $\nu^i(x_s^i) \geq \nu$ for $1 \leq s \leq \tau^i - 1$. In this case, action “not play” will be optimal afterwards if $\nu^i(x_{\tau^i}^i) \leq \nu$. Thus, with respect to the rested nature of our problem and by assuming that decisions are taken in a stationary manner, only the sequences of actions where we either do not play or we play throughout $[1, \tau^i)$, where $\tau^i \in \mathcal{T}^i$, are optimal. Equating the objective values of these two policies for subproblem (3.18), provides us the indifference between these alternative actions for state x_1^i . The result of this, which is calculated by using equality in (3.24), concludes that

$$\nu^i(x_1^i) = \frac{\varrho_{1, \tau^i-1}^\beta(c^i(x_1^i, 0), c^i(x_2^i, 0), \dots, c^i(x_{\tau^i-1}^i, 0))}{\varrho_{1, \tau^i-1}^\beta(-1, -1, \dots, -1)},$$

which establishes the statement of proposition. $\square$

Note that RAI policy is a priority index policy, that is, a policy based on playing the arm with the highest index value at each decision step. It is an optimal solution to problem (3.16) but does not necessarily provide optimal policies for our risk-averse MAB. The reason lies in the fact that problem (3.16) is the result of several relaxations of the original problem (3.6), including the relaxation of activation constraint in a Lagrangian manner as well as taking the summations over all arms out in the objective function with respect to subadditivity and monotonicity properties of dynamic risk measures. Therefore, RAI policy is a feasible policy for our original problem (3.6) and provides an upper bound to the risk-averse value function $R(x_1)$ given in (3.6).

An important exception to the discussion above is *one-armed bandit problem*, as it is indexable with RAI, which we discuss in the next section.

3.4 Index Policy for Risk-averse One-armed Bandit Problem

In this section, we characterize optimal policies for the risk-averse one-armed bandit problem. One-armed bandit problem is also known as the two-armed bandit problem, where one arm contains a single state and the other one is a Markovian arm with finite number of states. This problem is of interest from both applied and theoretical viewpoints. The reader is referred to [46], [33], and [34] for some interesting applications of one-armed bandit problem.

In our notation, we consider that the deterministic arm is indexed with 1, and the stochastic one with 2. We intend to show that, unlike multi-armed bandits, the structure of optimal policies can be obtained by RAI presented in Definition 3.3.1. In this regard, first, we need to provide the optimal value of this problem, which is given in the following lemma.

Lemma 3.4.1. *In one-armed bandit problem, starting from state $x_1 \in \mathcal{X}$, the optimal risk-averse total discounted cost is*

$$R(x_1) = \frac{c^1}{1 - \beta} + \min_{\tau^2 \geq 2} \left\{ \varrho_{1, \tau^2-1}^\beta (c^2(x_1^2, 0), \dots, c^2(x_{\tau^2-1}^2, 0)) \right. \\ \left. + c^1 \varrho_{1, \tau^2-1}^\beta (-1, -1, \dots, -1) \right\},$$

where τ^2 is as in (3.20) and $c^1 = c^1(x_1^1, 0)$.

Proof. Let I_t be an indicator random variable which takes value of one if the stochastic arm is played at time $t \in \mathbb{N}$ and zero otherwise. Therefore, according

to (3.6), we have

$$\begin{aligned}
R(x_1) &= \min_{\pi \in \Pi} \varrho^\beta \left(c(x_1, u_1), c(x_2, u_2), c(x_3, u_3), \dots \right) \\
&= \min_{I_t \in \{0,1\}} \varrho^\beta \left(I_1 c^2(x_1^2, 0) + (1 - I_1) c^1, \right. \\
&\quad \left. I_2 c^2(x_2^2, 0) + (1 - I_2) c^1, I_3 c^2(x_3^2, 0) + (1 - I_3) c^1, \dots \right) \\
&= \min_{I_t \in \{0,1\}} \varrho^\beta \left(c^1 + I_1 (c^2(x_1^2, 0) - c^1), \right. \\
&\quad \left. c^1 + I_2 (c^2(x_2^2, 0) - c^1), c^1 + I_3 (c^2(x_3^2, 0) - c^1), \dots \right) \\
&= \frac{c^1}{1 - \beta} + \min_{I_t \in \{0,1\}} \varrho^\beta \left(I_1 (c^2(x_1^2, 0) - c^1), \right. \\
&\quad \left. I_2 (c^2(x_2^2, 0) - c^1), I_3 (c^2(x_3^2, 0) - c^1), \dots \right),
\end{aligned}$$

where the last equality follows from axiom (A3) and the fact that c^1 is deterministic.

According to the rested nature of our bandit problem, at the moment we play the deterministic arm, we will continue playing it for the rest of the process. For this reason, our playing policy is either always playing the deterministic arm or playing the stochastic arm for some time and then playing the deterministic one for the rest of the process. So, as there exists a stopping time τ^2 in playing our stochastic bandit, which we consider it to be infinity in case we always play the stochastic bandit, we will have

$$\begin{aligned}
R(x_1) &= \frac{c^1}{1 - \beta} + \min_{\tau^2 \geq 2} \varrho_{1, \tau^2-1}^\beta (c^2(x_1^2, 0) - c^1, \dots, c^2(x_{\tau^2-1}^2, 0) - c^1) \\
&= \frac{c^1}{1 - \beta} + \min_{\tau^2 \geq 2} \left\{ \varrho_{1, \tau^2-1}^\beta (c^2(x_1^2, 0), \dots, c^2(x_{\tau^2-1}^2, 0)) + c^1 \sum_{t=0}^{\tau^2-2} -\beta^t \right\} \\
&= \frac{c^1}{1 - \beta} + \min_{\tau^2 \geq 2} \left\{ \varrho_{1, \tau^2-1}^\beta (c^2(x_1^2, 0), \dots, c^2(x_{\tau^2-1}^2, 0)) \right. \\
&\quad \left. + c^1 \varrho_{1, \tau^2-1}^\beta (-1, -1, \dots, -1) \right\},
\end{aligned}$$

where the second equality is from axiom (A3) and the third equality is from the definition of dynamic risk measure. $\square$

Now, defining RAI as given in Definition 3.3.1, it is clear that the risk-averse

allocation index of the deterministic arm is $\nu^1(x_1^1) = -c^1$. In the next theorem, we show that the policy based on RAI is also an optimal policy.

Theorem 3.4.1. *One-armed bandit problem is indexable with the risk-averse allocation indices introduced in Definition 3.3.1.*

Proof. Based on Lemma 3.4.1, playing the one-armed bandit under the policy having the stopping time $\tau^2 \in \mathcal{T}^2$ would result in the following risk-adjusted total discounted cost for initial state x_1 :

$$\frac{c^1}{1-\beta} + \varrho_{1,\tau^2-1}^\beta(c^2(x_1^2, 0), \dots, c^2(x_{\tau^2-1}^2, 0)) + c^1 \varrho_{1,\tau^2-1}^\beta(-1, -1, \dots, -1). \quad (3.25)$$

Regarding the previous discussion, playing the stochastic arm will be strictly optimal if there exists a stopping time for which the quantity in (3.25) falls behind the cost of always playing the deterministic arm, that is, $\frac{c^1}{1-\beta}$. Therefore, when there exists $\tau^2 \in \mathcal{T}^2$ for which the following holds, playing the stochastic arm will be optimal:

$$\frac{c^1}{1-\beta} + \varrho_{1,\tau^2-1}^\beta(c^2(x_1^2, 0), \dots, c^2(x_{\tau^2-1}^2, 0)) + c^1 \varrho_{1,\tau^2-1}^\beta(-1, -1, \dots, -1) < \frac{c^1}{1-\beta},$$

which is equivalent to

$$\frac{\varrho_{1,\tau^2-1}^\beta(c^2(x_1^2, 0), \dots, c^2(x_{\tau^2-1}^2, 0))}{\varrho_{1,\tau^2-1}^\beta(-1, -1, \dots, -1)} > -c^1,$$

implying that $\nu^2(x_1^2) > \nu^1(x_1^1)$. Therefore, in this case, the policy based on risk-averse allocation indices dictates playing the stochastic arm at state x_1 , which is also an optimal policy of risk-averse one-armed bandit problem.

On the other hand, if $\nu^2(x_1^2) < \nu^1(x_1^1)$, then for all $\tau^2 \geq 2$

$$\frac{\varrho_{1,\tau^2-1}^\beta(c^2(x_1^2, 0), \dots, c^2(x_{\tau^2-1}^2, 0))}{\varrho_{1,\tau^2-1}^\beta(-1, -1, \dots, -1)} < -c^1,$$

from which it can be observed that playing the deterministic arm is strictly optimal for state x_1 . In conclusion, playing our stochastic arm is optimal in state x_1 if and only if $\nu^2(x_1^2) \geq -c^1$. Also, playing the deterministic arm is optimal in this state if and only if $\nu^2(x_1^2) \leq -c^1$. This argument establishes the indexability of one-armed bandit problem with respect to Definition 3.3.1. $\square$

3.5 Computation of the Risk-averse Indices

So far, we presented the structure of RAI for risk-averse multi-armed bandit problem. Now, we need to provide an algorithm for the computation of these indices. As stated, RAI has a similar structure to Gittins index, thereby its computation is not that much more difficult than the Gittins index calculation. The approach we suggest here is a modified version of the algorithm presented by [22] for computing Gittins indices. For both cases, we solve single-arm optimal stopping problems. Let x^i be a generic state of arm i and consider the arm evolving under “play” action starting from x^i . In contrast to Section 3.3, where the set of states with “not play” action is provided with respect to a given ν as Θ_ν^i , here we define Φ_k^i as the *stopping set* related to the k^{th} highest index state. Therefore, Φ_k^i has the same interpretation as $\Theta_{\nu^i(x^i)}^i$ provided that x^i has the k^{th} highest index value. In other words, Φ_k^i is the set of all states of arm $i \in \mathcal{K}$ excluding the best $k - 1$ index states. In this regard, we define $\tau_{x^i}^i(\Phi_k^i)$ as the first time that the state of arm i lies in subset Φ_k^i starting from x^i . Then, $\tau_{x^i}^i(\Phi_k^i)$ provides the supremum in (3.21) given that x^i has the k^{th} highest index value. Also, for each set $\Phi_k^i \subseteq \mathcal{X}^i$ and $x^i \in \mathcal{X}^i$, we define $N_{x^i}^i(\Phi_k^i)$ as

$$N_{x^i}^i(\Phi_k^i) = \frac{\varrho_{1, \tau_{x^i}^i(\Phi_k^i)-1}^\beta(c^i(x^i, 0), c^i(x_2^i, 0), \dots, c^i(x_{\tau_{x^i}^i(\Phi_k^i)-1}^i, 0))}{\varrho_{1, \tau_{x^i}^i(\Phi_k^i)-1}^\beta(-1, -1, \dots, -1)}, \quad (3.26)$$

which yields the optimal value in (3.21) as long as x^i is the k^{th} highest index state.

Algorithm 3 computes the RAI for arm $i \in \mathcal{K}$, with the inputs as the costs, discount factor and the parameters of the risk measure. The output of this algorithm, integer state identifiers, $\nu^i(k)$, $k \in \{1, \dots, |\mathcal{X}^i|\}$, provide the RAI of all states of arm i , numbered in increasing order of their index values, that is, state 1 having the highest and state $|\mathcal{X}^i|$ having the lowest RAI. In this algorithm, first we identify the highest index state, where all the states in arm i are members of the stopping set Φ_1^i , imposing the occurrence of the supremum in (3.21) in $\tau^i = 2$ almost surely. Afterwards, the second-highest index is computed by repeating the same computation on a new stopping set obtained by removing the state with

Algorithm 3 Computing RAI for Arm i

```
1: for  $k = 1$  to  $|\mathcal{X}^i|$  do
2:   if  $k = 1$  then
3:      $\Phi_k^i = \mathcal{X}^i$ 
4:   else
5:      $\Phi_k^i = \Phi_{k-1}^i \setminus \alpha_{k-1}^i$ 
6:   end if
7:    $\alpha_k^i = \arg \max_{x^i \in \Phi_k^i} N_{x^i}^i(\Phi_k^i)$ 
8:    $\nu^i(k) = N_{\alpha_k^i}^i(\Phi_k^i)$ 
9: end for
10: return  $\nu^i(k), k \in \{1, \dots, |\mathcal{X}^i|\}$ 
```

the highest index from previous stopping set. The next steps employ this idea and progress in similar fashion, until the last iteration, where $\Phi_{|\mathcal{X}^i|}^i$ is left with only one state which has the smallest index value.

3.6 Computational Experiments

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

- (1) an optimal policy of our problem, which is obtained by using the *risk-averse policy iteration algorithm with convex optimization method* (see, [47]).
- (2) the risk-averse index policy (RA) obtained via *risk-averse allocation indices*.
- (3) the risk-neutral policy (RN) obtained via Gittins indices.

We consider a problem of three-armed bandit, where each arm contains four states and randomly generate 1000 test instances. In each test instance, the transition probabilities of each arm under the play action are sampled from $U(0, 1)$, where U represents uniform distribution, and the resulting transition matrix is

normalized across the rows, i.e., entries of each row is divided by its sum. Following [48], the costs in each state are drawn from a truncated normal distribution with mean value generated from $U(-6, -5)$ and standard deviation σ taking values from the set $\{0.01, 0.5, 1\}$. The experiments are conducted with discount factor $\beta \in \{0.9, 0.95\}$.

As risk measures, we employ first-order mean-semideviation, given in (3.2), with $\kappa \in \{0.25, 0.50, 0.75, 1\}$, as well as mean-AVaR, given in (3.3), with parameters of $\alpha \in \{0.80, 0.90, 0.95\}$ and $\lambda \in \{0, 0.5\}$. For each test instance, we compare RA and RN policies based on two different performance measures:

- (1) the suboptimality percentage of policy $\pi \in \{\text{RA}, \text{RN}\}$ for each initial state $x \in \mathcal{X}$. This percentage is computed as $100 \times (R(x) - R^\pi(x))/R(x)$, where $R^\pi(x)$ denotes the value of objective function in (3.6) under policy π ,
- (2) similarity percentage of policy $\pi \in \{\text{RA}, \text{RN}\}$ compared to optimal policy π^* . It is computed as $100 \times \sum_{x \in \mathcal{X}} \mathbb{1}_{(\pi, \pi^*)}(x)/|\mathcal{X}|$, where $\mathbb{1}_{(\pi, \pi^*)}(x)$ is equal to one if $\pi^*(x) = \pi(x)$ and zero otherwise.

For each test instance and risk measure, we compute maximum of suboptimality percentages over all states. The maximum suboptimality percentages for first-order mean-semideviation risk measure are summarized in Table 3.1. In the third column, we report mean and maximum of these percentages over 1000 test instances. It can be seen that RA policy is much better than RN policy regarding this performance measure. For instance, for $\sigma = 1$, $\beta = 0.90$, and $\kappa = 1$, the maximum value for maximum suboptimality percentage for RN is 11.81% while it is 3.813% for RA. As indicated in Table 3.1, by increasing the variability of cost values, that is σ , we observe higher maximum suboptimality percentages for both policies RN and RA. However, this increase is more noticeable for RN. Moreover, increasing the level of risk-aversion (increasing the value of κ) results in higher maximum suboptimality percentages for both policies, yet higher increase belongs to RN. Concerning the discount factor β , no consistent behavior is observed. For instance, when $\kappa = 1$, the worst suboptimality percentage of

policy RA corresponding to $\sigma = 0.01$ decreases as β increases, on the other hand it increases for $\sigma = 1$.

The average similarity percentage for this risk measure is reported in Table 3.2. With respect to these results, the performance of our risk-averse index policy RA is exceptionally well. For instance, for $\sigma = 1$, $\beta = 0.95$, $\kappa = 1$, there is more than 6% difference in the average similarity percentages of RN and RA.

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

The results related to maximum suboptimality percentage for mean-AVaR risk measure are summarized in Table 3.3. Similar to first-order mean-semideviation, increase in cost value variability σ results in increase in maximum suboptimality percentage of RN and RA, where the increase in the former is more notable. And, no consistent behavior is observed with respect to the discount factor β . However, increasing the level of risk-aversion has a diverse effect in RN and RA. By increasing the level of risk-aversion, i.e., increasing the value of α or reducing λ , the performance of RN decreases in general. On the other hand, as α increases, the performance of RA in general increases. Additionally, for $\alpha \in \{0.80, 0.90\}$, as λ decreases the performance of RA decreases, however, it increases for $\alpha = 0.95$. Our RA policy is superior to RN policy. It can be seen that, when $\sigma = 1$, $\beta = 0.90$, $\lambda = 0$ and $\alpha = 0.90$, although the worst suboptimality percentage of RN is 39.692%, it is 2.761% for RA.

The average similarity percentages for this risk measure are summarized in Table 3.4. In all the performance measures, RA is superior to RN. While the average similarity percentage is more than 99% for RA, it can drop to around 85% for RN.

In general, the performance of our index policy RA is more superior to

RN when mean-AVaR is used as the risk measure instead of first-order mean-semideviation. This superiority is more clear when we focus on the maximum values in Table 3.3 for $\sigma = 1$, where in general the suboptimality percentage of RN is more than 10 times than that of RA.

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

Moreover, in a separate experiment, in order to visualize the time efficiency of our index policy compared to solving the corresponding MDP optimally, we record the CPU time of the algorithms for 12 problem sizes for first-order mean-semideviation. These problem sizes start from problem size 1 where we have 2 arms with 3 states each and ends with problem size 12 where we have 3 arms with 5 states each. The results are presented in Fig. 3.1 where the CPU time is in logarithmic scale to provide better comparison. It is seen that the computation of our risk-averse index heuristic takes significantly less time than that of finding optimal policy. By increasing the problem size, computation time of our risk-averse indices seems to grow linearly while the computation time of solving the risk-averse MDP optimally shows an exponential growth. The computation time of the same instances for mean-AVaR is longer but provides a similar pattern.

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

3.7 Conclusion

In this chapter, we introduce a new approach in incorporating risk into multi-armed bandit problem by using dynamic coherent measures of risk and propose

Table 3.1: Statistics of maximum suboptimality percentage of policies RN and RA for first-order mean-semideviation

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

Table 3.2: Average similarity percentages of policies RN and RA for first-order mean-semideviation

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

a priority-index heuristic, analogous to the structure of Gittins index. The experiments we conduct to test the performance of our index policy indicate the excellence of the policy. Compared to the policy derived from Gittins index, the proposed index policy provides much lower suboptimality percentages and it is more similar to the optimal policy. Our experiments comparing the computational complexity of our index heuristic to solving the corresponding MDP reveals that our method is time efficient and the computation time grows linearly with the problem size, whereas the risk-averse MDP computation time shows an exponential growth.

As future research, one direction would be to perform a regret analysis of

Table 3.3: Statistics of maximum suboptimality percentage of policies RN and RA for mean-AVaR

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

Table 3.4: Average similarity percentages of policies RN and RA for mean-AVaR

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

the risk-averse allocation index heuristic for this multi-armed bandit problem, which would further solidify the current computational results indicating that the RAI provides near-optimal policies. The other direction would be to extend this framework to the restless setting or to the settings with imperfect information.

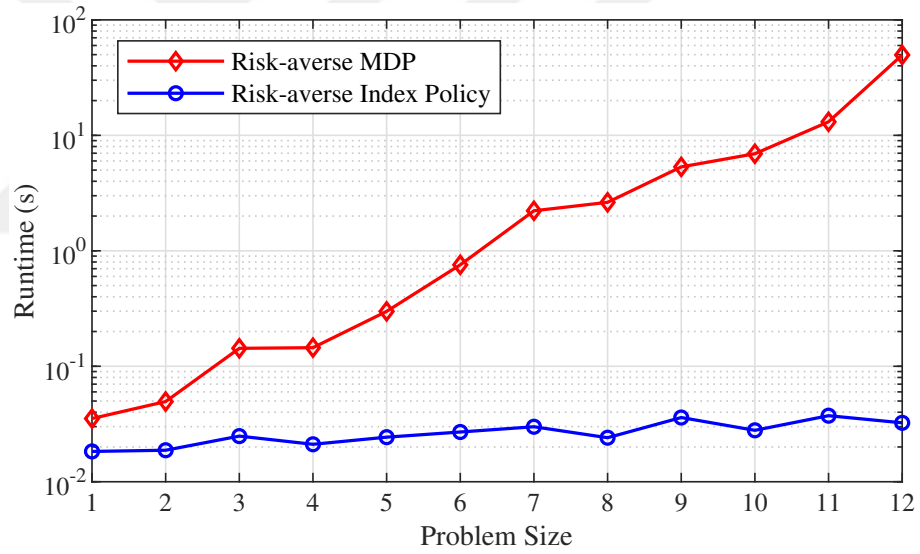


Figure 3.1: Run time of obtaining the optimal policy and index policy for first-order mean-semideviation

Chapter 4

Index Policy for Multi-armed Bandit Problem with Dynamic Risk Measures

In this chapter, we study risk-averse control of the multi-armed bandit problem in regard to the concept of dynamic risk measures with the goal of determining a policy with the best risk adjusted total discounted return. In respect of this specific setting, we present a theoretical analysis based on Whittle's retirement problem and propose a priority-index policy that reduces to Gittins index when the level of risk aversion converges to zero. We generalize the restart-in-state formulation of the Gittins index to effectively compute these risk-averse allocation indices. Numerical results exhibit excellent performance of this heuristic approach for two well known coherent risk measures of first-order mean-semideviation and mean-AVaR. Our experimental studies suggest that there is no guarantee an index-based optimal policy exists for our risk-averse problem. Nonetheless, our risk-averse allocation indices can achieve optimal or near-optimal policies which in some instances are easier to interpret compared to the exact optimal policy.

The aims of this chapter are the followings: (1) the characterization of an optimal index-based policy for single arm risk-averse MAB, (2) an effective algorithm

for estimating these indices, (3) an extension to the case of multi-arm problem, and (4) a numerical experiment to test the performance of our heuristic. To the best of authors' knowledge, this is the first study on the generalization of the work of [19] to find an index-based good feasible policy for risk-averse MAB.

4.1 Problem Description and Preliminaries

The classical risk-neutral multi-armed bandit problem is designed to maximize the expected total discounted reward. More explicitly, we can model MAB as a Markov decision process (MDP), in which a decision maker determines which arm to play, among K available arms, at each step of decision $t \in \mathbb{N}$, where $\mathbb{N} = \{1, 2, 3, \dots\}$. In this study, we consider a risk-averse MAB with dynamic risk measures. For mathematical convenience, we consider negative of rewards that could be translated as costs.

Our problem setting is very similar to that described in Chapter 3 with the following different elements:

- (i) A finite action space $\mathcal{U}$ and action sets $U(x) \subseteq \mathcal{U}$, $x \in \mathcal{X}$, where $U(x)$ is a nonempty set of possible actions at each state $x \in \mathcal{X}$. At each step $t \in \mathbb{N}$, we apply an *action* $u_t = \{u_t^1, u_t^2, \dots, u_t^K\}$, with $u_t^i \in \{0, 1\}$ as the action applied to arm $i \in \mathcal{K}$. Here, $u_t^i = 1$ indicates that arm i is activated at step t , while $u_t^i = 0$ reflects that arm i is not played. *Admissible actions* are those for which precisely one element of u_t equals to one at every step of $t \in \mathbb{N}$. That is, at each step, with any choice of admissible actions, the state of a single arm evolves independently of other arms in a Markovian manner, and there will be no changes to the states of the remaining arms.
- (ii) At each step, the state of the arm we activate evolves in a Markovian fashion according to the transition probabilities $Q^i(x, 1, y) = \mathbb{P}(x_{t+1}^i = y \mid x_t^i = x, u_t^i = 1)$, $i \in \mathcal{K}$, $x, y \in \mathcal{X}^i$, $t \in \mathbb{N}$. However, if we do not play an arm

its state remains unchanged, i.e., for all $i \in \mathcal{K}$, $x, y \in \mathcal{X}^i$, $t \in \mathbb{N}$ we have

$$Q^i(x, 0, y) = \mathbb{P}(x_{t+1}^i = y \mid x_t^i = x, u_t^i = 0) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise.} \end{cases}$$

- (iii) For each $i \in \mathcal{K}$, the *cost function* $g^i : \mathcal{X}^i \times \mathcal{X}^i \rightarrow \mathbb{R}_-$ is set to be finite and non-positive cost incurred by playing arm i at state x^i and its transition to state y^i . We characterize $c^i(x^i, u^i, y^i)$ as the cost of playing arm i transitioning from state $x^i \in \mathcal{X}^i$ to state $y^i \in \mathcal{X}^i$ under action $u^i \in \{0, 1\}$:

$$c^i(x^i, u^i, y^i) := \begin{cases} g^i(x^i, y^i), & \text{if } u^i = 1 \\ 0, & \text{if } u^i = 0, \end{cases} \quad (4.1)$$

and $c(x, u, y)$ denotes the cost incurred over all arms, that is $c(x, u, y) = \sum_{i \in \mathcal{K}} c^i(x^i, u^i, y^i)$, $x = (x^1, \dots, x^K) \in \mathcal{X}$, $y = (y^1, \dots, y^K) \in \mathcal{X}$ and $u = (u^1, \dots, u^K)$, $u^i \in \{0, 1\}$ for $i \in \mathcal{K}$. Please remind that we consider costs as negative of rewards.

In this study, we use dynamic risk measures to model risk-aversion and then we look for a stationary Markov policy to minimize the overall risk-averse discounted costs incurred over an infinite horizon. As far as the theory of [39] is concerned, the infinite horizon steady-state MDP with dynamic risk measures has a stationary optimal policy and we can obtain this policy either with *value iteration* or *policy iteration* algorithms. That being said, such algorithms could be computationally demanding, and sometimes even practically impossible for large-scale problems. Moreover, in certain cases, the structure of the optimal policy is not easy to interpret, which creates complications while implementing them to real life problems. Determined to resolve these difficulties, we are looking for reasonable index-based heuristic policies.

Prior to presenting our model, we initially provide description and properties of dynamic risk measures. Consider 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 $\mathcal{Z}_1 = \mathbb{R}$. In our case, considering that each $\mathcal{F}_t$ is finite, the spaces $\mathcal{Z}_t$ can be specified with finite dimensional vector spaces.

One-step conditional risk measure, $\rho_t : \mathcal{Z}_{t+1} \rightarrow \mathcal{Z}_t$, $t \in \{1, \dots, T\}$ is a measure satisfying the axioms provided below (see [41, 42]):

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

Conditions (A1)-(A4) are referred to as *convexity*, *monotonicity*, *translation invariance* and *positive homogeneity*, respectively. These are analogous to the axioms of coherent risk measures proposed by [38].

First-order mean-semideviation and *mean-average value-at-risk* (mean-AVaR) are two well-known coherent risk measures, which were extensively examined in the literature. Here, we offer the conditional variants of the two risk measures. Conditional first-order mean-semideviation risk measure is specified as

$$\rho_t(Z_{t+1}) = \mathbb{E}[Z_{t+1}|\mathcal{F}_t] + \kappa \mathbb{E}[(Z_{t+1} - \mathbb{E}[Z_{t+1}|\mathcal{F}_t])_+|\mathcal{F}_t], \quad (4.2)$$

where $\kappa \in [0, 1]$ and $(a)_+ := \max\{a, 0\}$ for $a \in \mathbb{R}$. With the parameters of $\alpha \in (0, 1)$ and $\lambda \in [0, 1]$, conditional mean-AVaR risk measure can be represented as

$$\rho_t(Z_{t+1}) = \lambda \mathbb{E}[Z_{t+1}|\mathcal{F}_t] + (1 - \lambda) \text{AVaR}_\alpha(Z_{t+1}|\mathcal{F}_t), \quad (4.3)$$

where conditional AVaR_α is

$$\text{AVaR}_\alpha(Z_{t+1}|\mathcal{F}_t) = \min_{\eta \in \mathcal{Z}_t} \left\{ \eta + \frac{1}{1 - \alpha} \mathbb{E}[(Z_{t+1} - \eta)_+|\mathcal{F}_t] \right\}.$$

A *dynamic risk measure* is identified as a sequence of one-step conditional risk measures [39, 49, 50, 51, 52]. The dynamic risk measure $\varrho_{1,T} : \mathcal{Z}_{1,T} \mapsto \mathcal{Z}_1$ on the

finite horizon with the length of T is given as:

$$\varrho_{1,T}(Z_2, Z_3, \dots, Z_{T+1}) := \rho_1(Z_2 + \rho_2(Z_3 + \rho_3(Z_4 + \dots + \rho_T(Z_{T+1}) \dots))), \quad (4.4)$$

and in the discounted case, with a discount factor of $\beta \in (0, 1)$, it can be presented as:

$$\varrho_{1,T}^\beta(Z_2, \dots, Z_{T+1}) := \rho_1(Z_2 + \rho_2(\beta Z_3 + \rho_3(\beta^2 Z_4 + \dots + \rho_T(\beta^{T-1} Z_{T+1}) \dots))). \quad (4.5)$$

Further, we describe the dynamic risk measure for infinite horizon as:

$$\varrho(Z_2, Z_3, Z_4, \dots) := \lim_{T \rightarrow \infty} \varrho_{1,T}(Z_2, Z_3, \dots, Z_{T+1}), \quad (4.6)$$

$$\varrho^\beta(Z_2, Z_3, Z_4, \dots) := \lim_{T \rightarrow \infty} \varrho_{1,T}^\beta(Z_2, Z_3, \dots, Z_{T+1}). \quad (4.7)$$

We employ dynamic risk measure (4.7) to evaluate the risk of cost sequences $c(x_t, u_t, x_{t+1}) \in \mathcal{Z}_{t+1}$, $t \in \mathbb{N}$, for policy π and initial state of $x_1 \in \mathcal{X}$:

$$\begin{aligned} R^\pi(x_1) &= \varrho^\beta(c(x_1, u_1, x_2), c(x_2, u_2, x_3), \dots) \\ &= \rho_1(c(x_1, u_1, x_2) + \rho_2(\beta c(x_2, u_2, x_3) + \rho_3(\beta^2 c(x_3, u_3, x_4) + \dots))). \end{aligned} \quad (4.8)$$

Considering $R(x_1)$ as the optimal risk-averse total discounted cost with the restriction of playing one arm at each step, we have

$$\begin{aligned} R(x_1) &= \min_{\pi \in \Pi} R^\pi(x_1) \\ &= \min_{\pi \in \Pi} \rho_1 \left(\sum_{i \in \mathcal{K}} c^i(x_1^i, u_1^i, x_2^i) + \rho_2 \left(\beta \sum_{i \in \mathcal{K}} c^i(x_2^i, u_2^i, x_3^i) + \dots \right) \right), \end{aligned} \quad (4.9)$$

where Π represents the class of stationary admissible policies for our problem, that is, we are only allowed to play one arm at each step. We can solve (4.9) with the help of the following dynamic programming equations of MDP with dynamic risk measures [39, 47, 53]:

$$R(x_1) = \min_{\pi \in \Pi} \sigma(c(x_1, u_1, x_2) + \beta R(x_2), x_1, Q(x_1, u_1, x_2)), \quad (4.10)$$

where σ is a *risk transition mapping*. Let $\mathcal{V}$ be the space of all real functions on $\mathcal{U} \times \mathcal{X}$. Then, $\sigma : \mathcal{V} \times \{(x, Q(x, u, y)) \mid u \in U(x), x \in \mathcal{X}, y \in \mathcal{X}\} \mapsto \mathbb{R}$ is a *risk transition mapping* if for every $x \in \mathcal{X}$ and $u \in U(x)$, the function $\varrho \mapsto \sigma(\varrho, x, m)$ is a coherent measure of risk on $\mathcal{V}$ (for details, see [39, 47, 53]). Considering that the *risk transition mapping* is coherent and finite valued with respect to the first argument, it is also continuous with respect to this argument. Therefore, there exists a closed convex set $\mathcal{A}(x, m)$ such that for all $v \in \mathcal{V}$ we have the following dual representation (see [54, Theorem 2.2] for details):

$$\sigma(\varrho, x, m) = \sup_{\mu \in \mathcal{A}(x, m)} \mathbb{E}_\mu[\nu].$$

As stated in [39], (4.10) has a stationary optimal policy and we can obtain this policy using risk-averse dynamic programming tools. We recall the following theorem which appeared as Theorem 4 of [39]:

Theorem 4.1.1. *Assume that the following conditions are satisfied:*

- (i) *For every $x, y \in \mathcal{X}$ the transition kernel $Q(x, \cdot, y)$ is continuous;*
- (ii) *The conditional risk measures ρ_t , $t = 1, \dots, T$, are Markov, stationary and such that for every $x \in \mathcal{X}$ the multifunction $\mathcal{A}(x, \cdot)$ is lower semicontinuous;*
- (iii) *The function $c(\cdot, \cdot, \cdot)$ is uniformly bounded;*
- (iv) *For every $x, y \in \mathcal{X}$ the function $c(x, \cdot, y)$ is lower semicontinuous;*
- (v) *For every $x \in \mathcal{X}$ the set $U(x)$ is compact.*

Then problem (4.9) has an optimal solution and its optimal value, as a function of the initial state x_1 , satisfies the dynamic programming equation in (4.10). Moreover, an optimal stationary Markov policy $\hat{\Pi} = \{\hat{\pi}, \hat{\pi}, \dots\}$ exists and satisfies the equation:

$$\hat{\pi}(x_1) = \arg \min_{u_1 \in U(x_1)} \sigma(c(x_1, u_1, x_2) + \beta R(x_2), x_1, Q(x_1, u_1, x_2)), \quad x_1 \in \mathcal{X}. \quad (4.11)$$

Conversely, every bounded solution of (4.10) is the optimal value of problem (4.9), and every measurable solution of (4.11) defines an optimal stationary Markov policy.

However, as discussed, these approaches are practically infeasible for medium-scale to large-scale problems. Moreover, as stated in [40], since the additivity of the objective function is one of the key assumptions in [18] to ensure the optimality of an index policy, index policies might not always provide an optimal solution to our risk-averse MAB, as coherent risk measures are sub-additive. Consequently, we can achieve an optimal policy by undertaking the much more complex procedure of solving risk-averse dynamic programming equations. At this stage, following the classic work of [19], and based on the theory and presentation of [55] and [56], we try to determine a heuristic solution to risk-averse MAB in the form of index policies. For that, we need to design a calibrating function for each arm. In order to come up with such functions, we introduce a generalized form of the problem, with which by focusing on a single arm, we can obtain proper calibrating functions.

4.2 Risk-averse Retirement Problem

In our analysis, we follow a similar approach as [19], and in order to design index policies for risk-averse MAB, we allow retiring from the process at each step with the cost of $M \in \mathbb{R}^-$, here referred to as *risk-averse retirement process*. With this new setting, the risk of the discounted cost sequences resulting from the policy π can be represented by:

$$R^\pi(x_1, M) = \varrho_{1,\tau}^\beta(c(x_1, u_1, x_2), c(x_2, u_2, x_3), \dots, c(x_{\tau-1}, u_{\tau-1}, x_\tau), M), \quad (4.12)$$

where $x_1 \in \mathcal{X}$ denotes the initial state, and τ is the time of retiring from the process. Here, we slightly abuse notation by considering $\varrho_{1,\tau}^\beta(Z_2, Z_3, \dots, Z_{\tau+1}) = \varrho^\beta(Z_2, Z_3, \dots, Z_{\tau+1}, 0, 0, \dots)$.

We introduce $R(x_1, M)$ as the optimal risk-averse total discounted cost under the restriction that at each step either one arm will be played or the play will be terminated with retirement cost of M . Considering Π_M as the class of stationary

admissible policies for the retirement problem, we have

$$R(x_1, M) = \min_{\pi \in \Pi_M} R^\pi(x_1, M).$$

Now, using the *risk transition mapping* σ , we can rewrite $R(x_1, M)$ as

$$R(x_1, M) = \min\{M, \min_{\pi \in \Pi} \sigma(c(x_1, u_1, x_2) + \beta R(x_2, M), x_1, Q(x_1, u_1, x_2))\}. \quad (4.13)$$

In the next theorem, we show that under some conditions the optimal value of risk-averse retirement problem would be the same as that of risk-averse MAB. Based on the result of this theorem, we can consider risk-averse MAB as a special case of risk-averse retirement problem which justifies the applicability of retirement problem solution strategies for risk-averse MAB.

Theorem 4.2.1. *Let $C_L \leq c(x, u, y) \leq C_U$ for $x \in \mathcal{X}$, $y \in \mathcal{X}$, $u = (u^1, u^2, \dots, u^K)$, $u^i \in \{0, 1\}$, $i \in \mathcal{K}$. For fixed $x_1 \in \mathcal{X}$, $R(x_1, M)$ is a non-decreasing function of M with*

$$R(x_1, M) = \begin{cases} R(x_1), & \text{for } M \geq C_U/(1 - \beta) \\ M, & \text{for } M \leq C_L/(1 - \beta), \end{cases} \quad (4.14)$$

where $R(x_1)$ represents the optimal risk-averse total discounted cost of risk-averse MAB (4.9).

Proof. The non-increasing property is obvious from (4.13) because an increase in M cannot decrease the minimal risk adjusted cost. For $M \leq C_L/(1 - \beta)$, any policy that does not require immediate retirement has a risk adjusted total cost bounded below by:

$$\inf_{\tau > 1} \varrho_{1,\tau}^\beta(C_L, C_L, \dots, M) = \inf_{\tau > 1} C_L \frac{1 - \beta^{\tau-1}}{1 - \beta} + \beta^{\tau-1} M \geq M.$$

Therefore, for $M \leq C_L/(1 - \beta)$, we observe optimality of immediate retirement. That is, $R(x_1, M) = M$, for $M \leq C_L/(1 - \beta)$.

Similarly, for $M \geq C_U/(1 - \beta)$, we have $\inf_{\tau > 1} \varrho_{1,\tau}^\beta(C_U, C_U, \dots, M) \leq M$, indicating immediate retirement is never optimal for $M > C_U/(1 - \beta)$. And with

retirement option off the game, the optimal policy of the risk-averse retirement problem and risk-averse MAB will be the same, providing the same minimal value. That is, $R(x_1, M) = R(x_1)$, for $M \geq C_U/(1 - \beta)$. $\square$

We are now seeking to heuristically find a solution of this problem belonging to the family of index policies. To obtain proper actions in each state, we need to develop a calibrating function for individual arms. To come up with such indices, we need to analyze each arm independently, and acquire reasonable calibrating functions using these separate problems.

4.3 Indexability and Indices

Consider a single arm $i \in \mathcal{K}$ problem. At any step, after acquiring information about the state of the arm, the decision maker has to decide whether to play the arm or to retire. In case the state is $x \in \mathcal{X}^i$ and it is decided to play (i.e., activate the arm), a bounded cost of $c^i(x^i, 1, y^i)$ is incurred with probability $Q^i(x^i, 1, y^i)$, $y \in \mathcal{X}^i$ and the process evolves in a Markovian manner. However, if it is decided to retire (that is, to terminate), then a terminal cost M is incurred. Let $R^i(x_1^i, M)$ denote the minimal risk adjusted total discounted payment with the initial state of $x_1^i \in \mathcal{X}^i$, then it satisfies the below optimality equation:

$$R^i(x_1^i, M) = \min\{M, \sigma(c^i(x_1^i, 1, x_2^i) + \beta R^i(x_2^i, M), x_1^i, Q^i(x_1^i, 1, x_2^i))\}. \quad (4.15)$$

Remark. Based on the structure of above optimality equation, we have $R^i(x_1^i, M) \leq M$ for all $x_1^i \in \mathcal{X}^i$ and $M \in \mathbb{R}^-$.

Our aim here is to propose a calibrating function for a risk-averse single arm bandit and to show that the risk-averse allocation indices generated with this calibrating function provide us with an optimal policy for the risk-averse single arm bandit. In other words, by showing that each risk-averse arm is indexable, we will show that risk-averse MAB is indexable. To achieve this goal, first we show in the first two lemmas that if it is optimal to retire in a given state with

termination cost of M , then retirement is also optimal once the termination cost is less than M . This feature is referred to as indexability. These topics are initially introduced by [31] for the case of average reward and afterwards are defined for the discounted case [57].

Lemma 4.3.1. *For fixed $x_1^i \in \mathcal{X}^i$, $R^i(x_1^i, M) - M$ is non-increasing in M .*

Proof. Let $R_n^i(x_1^i, M)$ represents the minimal risk adjusted cost when the decision maker is only allowed to play for at most n steps before retiring. So,

$$R_0^i(x_1^i, M) = M \quad (4.16)$$

$$R_n^i(x_1^i, M) = \min\{M, \sigma(c^i(x_1^i, 1, x_2^i) + \beta R_{n-1}^i(x_2^i, M), x_1^i, Q^i(x_1^i, 1, x_2^i))\}. \quad (4.17)$$

We will proceed the proof by induction on n . When $n = 0$, we have immediate retirement, therefore as stated in (4.16), the statement is satisfied. Now, assuming $R_{n-1}^i(x_1^i, M)$ is non-increasing in M and referring to (4.17), we have

$$\begin{aligned} R_n^i(x_1^i, M) - M &= \min\{M, \sigma(c^i(x_1^i, 1, x_2^i) + \beta R_{n-1}^i(x_2^i, M), x_1^i, Q^i(x_1^i, 1, x_2^i))\} - M \\ &= \min\{0, \sigma(c^i(x_1^i, 1, x_2^i) + \beta R_{n-1}^i(x_2^i, M), x_1^i, Q^i(x_1^i, 1, x_2^i)) - M\} \\ &= \min\{0, \sigma(c^i(x_1^i, 1, x_2^i) + \beta(R_{n-1}^i(x_2^i, M) - M), x_1^i, Q^i(x_1^i, 1, x_2^i)) \\ &\quad - (1 - \beta)M\}, \end{aligned}$$

where the last equality is due to (A3). Focusing on the last expression, as M increases, $R_{n-1}^i(x_2^i, M) - M$ is non-increasing due to the induction hypothesis. Therefore, based on monotonicity property of the conditional risk measures (A2), $\sigma(c^i(x_1^i, 1, x_2^i) + \beta(R_{n-1}^i(x_2^i, M) - M), x_1^i, Q^i(x_1^i, 1, x_2^i))$ is non-increasing. Since $-(1 - \beta)M$ is also non-increasing by increasing M , we conclude $R_n^i(x_1^i, M) - M$ is non-increasing in M for all $n \in \mathbb{N}$. Finally, since $R^i(x_1^i, M)$ can be obtained by the limit $\lim_{n \rightarrow \infty} R_n^i(x_1^i, M)$, it follows that $R^i(x_1^i, M) - M$ also satisfies the property, which completes the proof. $\square$

Lemma 4.3.2. *For $x^i \in \mathcal{X}^i$, if retirement is optimal with termination cost of M , it is also optimal with termination cost less than M .*

Proof. This holds because, for any $\bar{M} \leq M$, we have

$$0 \geq R^i(x^i, \bar{M}) - \bar{M} \geq R^i(x^i, M) - M = 0,$$

where the first and second inequalities follow from Remark 4.3 and Lemma 4.3.1, respectively, and the equality follows since retirement is optimal at M . So, $R^i(x^i, \bar{M}) = \bar{M}$, for $\bar{M} \leq M$. $\square$

Now, before introducing the calibrating function, we show in the following lemma that the optimal value for this problem is continuous with respect to M . This result will help us to provide a better representation of the indices.

Lemma 4.3.3. *For $x_1^i \in \mathcal{X}^i$, $R^i(x_1^i, M)$ is continuous in M .*

Proof. Let $R_n^i(x_1^i, M)$ represent the minimal risk adjusted cost when the decision maker is only allowed to play for at most n steps before retiring. We will proceed the proof by induction on n . When $n = 0$, we have $R_0^i(x_1^i, M) = M$, so the statement is satisfied. Now, assuming $R_{n-1}^i(x_1^i, M)$ is continuous in M , we have $R_n^i(x_1^i, M) = \min\{M, \sigma(c^i(x_1^i, 1, x_2^i) + \beta R_{n-1}^i(x_2^i, M), x_1^i, Q^i(x_1^i, 1, x_2^i))\}$. Note that the continuity of σ with respect to its first argument follows from the monotonicity and convexity axioms as discussed in [54, Proposition 3.1]. Therefore, as the composition of continuous functions is also continuous and with respect to the induction hypothesis, the second argument of the minimum is continuous in M . As minimum of two continuous functions is continuous, we have $R_n^i(x_1^i, M)$ is continuous in M .

Now, in order to complete the proof, we need to show that $R_n^i(x_1^i, M)$ converges to $R^i(x_1^i, M)$ uniformly with respect to M . For all $M \in \mathbb{R}^-$, we have

$$R^i(x_1^i, M) = \inf_{\tau \geq 1} \varrho_{1,\tau}^\beta(c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x_{\tau-1}^i, 1, x_\tau^i), M), \quad (4.18)$$

$$R_n^i(x_1^i, M) = \inf_{1 \leq \tau' \leq n+1} \varrho_{1,\tau'}^\beta(c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x_{\tau'-1}^i, 1, x_{\tau'}^i), M). \quad (4.19)$$

Let τ^* and τ'^* be the optimal solution of (4.18) and (4.19), respectively. In case $\tau^* \leq n + 1$ then $\tau'^* = \tau^*$, and $|R^i(x_1^i, M) - R_n^i(x_1^i, M)| = 0 \leq \epsilon$, $\forall \epsilon > 0$. Let

$C_L^i \leq c^i(x^i, 1, y^i) \leq C_U^i$ for all $x^i \in \mathcal{X}^i$ and $y^i \in \mathcal{Y}^i$, if $\tau^* > n + 1$ we have

$$\begin{aligned}
|R^i(x_1^i, M) - R_n^i(x_1^i, M)| &= R_n^i(x_1^i, M) - R^i(x_1^i, M) \\
&= \varrho_{1, \tau'^*}^\beta(c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x_{\tau'^*-1}^i, 1, x_{\tau'^*}^i), M) \\
&\quad - \varrho_{1, \tau^*}^\beta(c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x_{\tau^*-1}^i, 1, x_{\tau^*}^i), M) \\
&\leq \varrho_{1, n+1}^\beta(c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x_n^i, 1, x_{n+1}^i), M) \\
&\quad - \varrho_{1, \tau^*}^\beta(c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x_{\tau^*-1}^i, 1, x_{\tau^*}^i), M) \\
&\leq \varrho_{1, n+1}^\beta(c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \dots, c^i(x_n^i, 1, x_{n+1}^i), M) \\
&\quad - \varrho_{1, \tau^*}^\beta(c^i(x_1^i, 1, x_2^i), \dots, c^i(x_n^i, 1, x_{n+1}^i), C_L^i, \dots, C_L^i, M) \\
&\leq \varrho_{1, \tau^*}^\beta(\underbrace{0, \dots, 0}_n, M - C_L^i, -C_L^i, \dots, -C_L^i, -M) \\
&\quad + \varrho_{1, \tau^*}^\beta(c^i(x_1^i, 1, x_2^i), \dots, c^i(x_n^i, 1, x_{n+1}^i), C_L^i, C_L^i, \dots, C_L^i, M) \\
&\quad - \varrho_{1, \tau^*}^\beta(c^i(x_1^i, 1, x_2^i), \dots, c^i(x_n^i, 1, x_{n+1}^i), C_L^i, C_L^i, \dots, C_L^i, M) \\
&= \beta^n \varrho_{1, \tau^*-n}^\beta(M - C_L^i, -C_L^i, \dots, -C_L^i, -M) \\
&\leq \beta^n \varrho_{1, \tau^*-n}^\beta(-C_L^i, -C_L^i, \dots, -C_L^i, -M) \\
&\leq \beta^n \varrho_{1, \tau^*-n}^\beta(-C_L^i, -C_L^i, \dots, -C_L^i, -C_L^i/(1-\beta)) \\
&\leq \beta^n \varrho_{1, \tau^*-n}^\beta(-C_L^i/(1-\beta), \dots, -C_L^i/(1-\beta)) \\
&\leq \beta^n \varrho^\beta(-C_L^i/(1-\beta), -C_L^i/(1-\beta), \dots) \\
&= \beta^n \frac{-C_L^i}{(1-\beta)^2},
\end{aligned}$$

where the first inequality holds as τ'^* attains the minimum value, the second inequity is based on monotonicity property (A2), the third inequity relates to subadditivity property, the fourth inequality is again based on (A2), the fifth inequality is according to Theorem 4.2.1 where we know if $\tau^* > 1$ then $M > C_L^i/(1-\beta)$ along with the monotonicity property, and lastly the sixth and the seventh inequalities relate to the monotonicity axiom. Therefore, for a given $\epsilon > 0$, we can choose $N \geq \ln \frac{\epsilon(1-\beta)^2}{-C_L^i} / \ln \beta$, a choice of N that is independent of M so that $|R^i(x_1^i, M) - R_n^i(x_1^i, M)| < \epsilon$ for all $M \in \mathbb{R}^-$ whenever $n \geq N$. Thus $R_n^i(x_1^i, M)$ converges uniformly to $R^i(x_1^i, M)$ with respect to M on $\mathbb{R}^-$.

With respect to M , we showed that $\{R_n^i(x_1^i, M)\}_{n \geq 0}$ is a sequence of continuous functions and that $R_n^i(x_1^i, M) \rightarrow R^i(x_1^i, M)$ uniformly. Then $R^i(x_1^i, M)$ is

continuous. □

Let us define $M^i(x_1^i)$ as the largest value of M at which it is optimal to retire at x_1^i . That is,

$$M^i(x_1^i) := \sup \{M \mid R^i(x_1^i, M) = M\}.$$

Since $R^i(x_1^i, M)$ is continuous in M , as stated in Lemma 4.3.3, we can also write

$$M^i(x_1^i) := \max \{M \mid R^i(x_1^i, M) = M\}. \quad (4.20)$$

The risk-averse allocation index $M^i(x_1^i)$ can be interpreted as the highest retirement payment M for which the decision maker is indifferent between retiring and proceeding with arm i . Now, we provide the optimal policy for the single arm bandit in the following theorem.

Theorem 4.3.1. *Observing state x^i in a single arm i , retirement is optimal if $M < M^i(x^i)$ and it is optimal to play if $M > M^i(x^i)$. In other words, $M^i(x^i)$ specifies the value for termination cost that makes the decision maker indifferent between retiring or continuing.*

Proof. According to [31], a single arm bandit is indexable if the set of states where the termination action is optimal increases monotonically from the empty set to the full set of states as M decreases from 0 to $-\infty$. In other words, $M^i(x^i)$ requires to induce a consistent ordering of the states, such that if it is optimal for arm i being at state x_1^i to terminate under the termination cost of $M^i(x^i)$, it will also be optimal to terminate under the termination cost of $M < M^i(x^i)$. Therefore, the indexability of a risk-averse single arm retirement process with the calibrating function of $M^i(x^i)$ follows from Lemma 4.3.2. □

Suppose now that instead of a single arm, there are K arms available. Once the state of each arm has been observed, a decision must be taken to either play exactly one arm or to terminate. In case decision maker plays arm $i \in \mathcal{K}$ with state of x^i , a cost $c^i(x^i, 1, y^i)$ is incurred with probability $Q^i(x^i, 1, y^i)$, $y^i \in \mathcal{X}^i$, and that arm evolves in a Markovian manner whereas the states of the other arms remain frozen. If decision maker decides to retire, the process terminates and cost

of M will be incurred. As discussed, solving risk-averse MAB directly would be computationally expensive in that the number of states grows exponentially as K increases. Although index-based policy is an optimal policy for the classical MAB, this is highly unlikely to happen for a risk-averse model as coherent risk measures are subadditive. At the same time, provided that the risk-averse allocation index is optimal for a risk-averse single arm model, it is rational to consider a risk-averse variation of this policy, which we introduce as a suboptimal policy for the risk-averse MAB problem. Hence, we introduce an index-based heuristic by generalizing the indexability of each K constituent single arm bandit problems. In fact, if we let $M^i(x^i)$ denote the indifference value for state x^i when only one arm is available, our proposed index-based policy for the risk-averse retirement problem is as follows:

Definition 4.3.1. (Index Policy for Risk-averse Retirement Problem) For $t \in \mathbb{N}$ and the state of $x_t = (x_t^1, x_t^2, \dots, x_t^K)$, retire if $M^i(x_t^i) \geq M$ for all $i \in \mathcal{K}$, and play arm j if $\min_{i \in \mathcal{K}} M^i(x_t^i) = M^j(x_t^j) \leq M$.

A critical feature of the Gittins index is that it can be characterized by stopping times. This feature offers a probabilistic interpretation of the index and is often employed to develop other exact methods for computing the indices, see [58] for details. In the following proposition, we show that with respect to our setting, the index can be achieved by minimizing over stopping times.

Proposition 4.3.1. For $i \in \mathcal{K}$, the risk-averse allocation index for each initial state $x_1^i \in \mathcal{X}^i$ can be represented as:

$$M^i(x_1^i) = \inf_{\tau^i > 1} \sup_{\mu^i \in \mathcal{A}_{\tau^i}^i} \frac{\mathbb{E}_{\mu^i} \left[\sum_{t=1}^{\tau^i-1} \beta^{t-1} c^i(x_t^i, 1, x_{t+1}^i) \right]}{1 - \mathbb{E}_{\mu^i} [\beta^{\tau^i-1}]}, \quad (4.21)$$

where $\mathcal{A}_{\tau^i}^i := \mathcal{A}^i(x_{\tau^i-1}^i, Q^i(x_{\tau^i-1}^i, 1, x_{\tau^i}^i)) \circ \dots \circ \mathcal{A}^i(x_2^i, Q^i(x_2^i, 1, x_3^i)) \circ \mathcal{A}^i(x_1^i, Q^i(x_1^i, 1, x_2^i))$ is a convex and closed set of probability measures on Ω and $\mathcal{A}^i(x_t^i, Q^i(x_t^i, 1, x_{t+1}^i))$, $t \in \{1, \dots, \tau^i - 1\}$ is a convex and closed set associated with σ in (4.15) (for more details, see [59]).

Proof. Consider the single arm bandit problem with arm i in which we can also retire at any time and incur cost M . Hence, we rewrite the Bellman equation in

(4.15) as:

$$R^i(x_1^i, M) = \min \left\{ M, \sup_{\mu^i \in \mathcal{A}^i(x_1^i, Q^i(x_1^i, 1, x_2^i))} \mathbb{E}_{\mu^i} [c^i(x_1^i, 1, x_2^i) + \beta R^i(x_2^i, M)] \right\}.$$

First, consider the immediate retirement ($\tau^i = 1$) as the optimal policy. Let π be the policy that prescribes playing the arm up to time $\tau^i > 1$ and then stops to pay the cost M , and let $R^{i,\pi}(x_1^i, M)$ denote the risk of the cost sequences resulting from the policy π for the single arm i . Hence,

$$R^{i,\pi}(x_1^i, M) = \sup_{\mu^i \in \mathcal{A}_{\tau^i}^i} \mathbb{E}_{\mu^i} \left[\sum_{t=1}^{\tau^i-1} \beta^{t-1} c^i(x_t^i, 1, x_{t+1}^i) + \beta^{\tau^i-1} M \right].$$

As immediate retirement is strictly optimal, $M < R^{i,\pi}(x_1^i, M)$ and there exists $\mu^{i,*}$ such that

$$\begin{aligned} M &< \mathbb{E}_{\mu^{i,*}} \left[\sum_{t=1}^{\tau^i-1} \beta^{t-1} c^i(x_t^i, 1, x_{t+1}^i) + \beta^{\tau^i-1} M \right] \\ &= \mathbb{E}_{\mu^{i,*}} \left[\sum_{t=1}^{\tau^i-1} \beta^{t-1} c^i(x_t^i, 1, x_{t+1}^i) | X_1^i = x_1^i \right] + M \mathbb{E}_{\mu^{i,*}} [\beta^{\tau^i-1}]. \end{aligned}$$

Thus

$$M \left(1 - \mathbb{E}_{\mu^{i,*}} [\beta^{\tau^i-1}] \right) < \mathbb{E}_{\mu^{i,*}} \left[\sum_{t=1}^{\tau^i-1} \beta^{t-1} c^i(x_t^i, 1, x_{t+1}^i) \right].$$

So, we have

$$M < \sup_{\mu^i \in \mathcal{A}_{\tau^i}^i} \frac{\mathbb{E}_{\mu^i} \left[\sum_{t=1}^{\tau^i-1} \beta^{t-1} c^i(x_t^i, 1, x_{t+1}^i) \right]}{1 - \mathbb{E}_{\mu^i} [\beta^{\tau^i-1}]}.$$

Since this holds for an arbitrary π , we conclude $M < M^i(x_1^i)$.

Similarly, playing until $\tau^i > 1$ is strictly optimal for x_1^i if

$$M > \inf_{\tau^i > 1} \sup_{\mu^i \in \mathcal{A}_{\tau^i}^i} \frac{\mathbb{E}_{\mu^i} \left[\sum_{t=1}^{\tau^i-1} \beta^{t-1} c^i(x_t^i, 1, x_{t+1}^i) \right]}{1 - \mathbb{E}_{\mu^i} [\beta^{\tau^i-1}]}.$$

In conclusion, immediate retirement is optimal in state x_1^i if and only if $M^i(x_1^i) \geq M$. Also, retiring at time $\tau^i > 1$ is optimal in this state if and only if $M^i(x_1^i) \leq M$. From this argument, we can infer $M^i(x_1^i)$ can provide the risk-averse allocation indices with respect to (4.21). $\square$

4.4 Computation of the Risk-averse Indices

In this section, we consider a single arm bandit process i and assume that $\mathcal{X}^i = \{1, 2, \dots, N\}$. Similar to [23], in order to calculate the allocation indices, we can relate the retirement problem to a restart-in-state formulation, and develop a better representation of the indices. In this approach, when the player is indifferent between playing the bandit process starting in state $x^i \in \mathcal{X}^i$ or retiring, this option of retiring is replaced with the equivalent option of restarting in state x^i , and indices are computed based on the optimal value function of state x^i in this new formulation. To elaborate, first we provide formal description of *risk-averse restart-in-state problem* in Definition 4.4.1, then prove this equivalency in Theorem 4.4.1.

Definition 4.4.1. (Risk-averse Restart-in-state- x Problem) A variation of bandit process that evolves in a Markovian manner and a cost is incurred in each step until we decide to restart the process in state x and continue similarly afterwards. In other words, at each step, the player has an option to play the arm or instantaneously restart in state x . The aim is to choose the restart times so as to minimize the risk adjusted total discounted cost over an infinite time horizon.

Let $v^i(y_1^i, x^i)$ be the optimal risk-averse total discounted cost for the single arm i under the restriction that at each step either we play the arm or the play will be restarted at state x^i for the initial state of $y_1^i \in \mathcal{X}^i$. We have for all $y_1^i \in \mathcal{X}^i$

$$v^i(y_1^i, x^i) = \min\{\sigma(c(x^i, 1, y_2^i) + \beta v^i(y_2^i, x^i), x^i, Q(x^i, 1, y_2^i)), \sigma(c(y_1^i, 1, y_2^i) + \beta v^i(y_2^i, x^i), y_1^i, Q(y_1^i, 1, y_2^i))\}. \quad (4.22)$$

In the next theorem, we show that the optimal value function for state x^i in the restart-in-state- x^i problem described in Definition 4.4.1 is equal to the risk-averse allocation index of state x introduced in Definition 3.3.1 for risk-averse MAB with retirement option.

Theorem 4.4.1. For $i \in \mathcal{K}$, the risk-averse allocation index for each state $x^i \in \mathcal{X}^i$ can be represented as $M^i(x^i) = v^i(x^i, x^i)$.

Proof. Based on the optimality equation of retirement problem (4.15), for $x^i \in \mathcal{X}^i$, we have

$$\begin{aligned}
R^i(1, M) &= \min\{M, \sigma(c^i(1, 1, y_2^i) + \beta R^i(y_2^i, M), 1, Q^i(1, 1, y_2^i))\} \\
R^i(2, M) &= \min\{M, \sigma(c^i(2, 1, y_2^i) + \beta R^i(y_2^i, M), 2, Q^i(2, 1, y_2^i))\} \\
&\vdots \\
R^i(x^i, M) &= \min\{M, \sigma(c^i(x^i, 1, y_2^i) + \beta R^i(y_2^i, M), x^i, Q^i(x^i, 1, y_2^i))\} \\
&\vdots \\
R^i(N, M) &= \min\{M, \sigma(c^i(N, 1, y_2^i) + \beta R^i(y_2^i, M), N, Q^i(N, 1, y_2^i))\}.
\end{aligned}$$

On the other hand, the risk-averse allocation index of state x^i introduced in Definition 4.3.1 is the value of M for which starting from state x^i we are indifferent between actions of continue to play and retire. Therefore, for $M = \sigma(c^i(x^i, 1, y_2^i) + \beta R^i(y_2^i, M), x^i, Q^i(x^i, 1, y_2^i))$, M corresponds to the index of state x^i , i.e., $M = M^i(x^i)$. Replacing $R^i(y_2^i, M)$ where $M = \sigma(c^i(x^i, 1, y_2^i) + \beta R^i(y_2^i, M), x^i, Q^i(x^i, 1, y_2^i))$ with $v^i(y_2^i, x^i)$, we have

$$\begin{aligned}
v^i(1, x^i) &= \min\{\sigma(c^i(x^i, 1, y_2^i) + \beta v^i(y_2^i, x^i), x^i, Q^i(x^i, 1, y_2^i)), \\
&\quad \sigma(c^i(1, 1, y_2^i) + \beta v^i(y_2^i, x^i), 1, Q^i(1, 1, y_2^i))\} \\
v^i(2, M) &= \min\{\sigma(c^i(x^i, 1, y_2^i) + \beta v^i(y_2^i, x^i), x^i, Q^i(x^i, 1, y_2^i)), \\
&\quad \sigma(c^i(2, 1, y_2^i) + \beta v^i(y_2^i, x^i), 2, Q^i(2, 1, y_2^i))\} \\
&\vdots \\
v^i(x^i, x^i) &= \sigma(c^i(x^i, 1, y_2^i) + \beta v^i(y_2^i, x^i), x^i, Q^i(x^i, 1, y_2^i)) \\
&\vdots \\
v^i(N, x^i) &= \min\{\sigma(c^i(x^i, 1, y_2^i) + \beta v^i(y_2^i, x^i), x^i, Q^i(x^i, 1, y_2^i)), \\
&\quad \sigma(c^i(N, 1, y_2^i) + \beta v^i(y_2^i, x^i), N, Q^i(N, 1, y_2^i))\}.
\end{aligned}$$

According to (4.22), the above equations provide optimal value functions for restart-in-state- x^i problem, where we are finding element-wise minimum of the value we achieve by restarting in state x^i and continuing the play. Therefore, as $v^i(x^i, x^i) = M$, the assertion of the theorem follows. $\square$

Algorithm 4 calculates the *risk-averse indices* for the states of a single arm,

with the inputs of the costs, the discount factor, as well as the risk measure parameters. Thereby, in multi-arm bandits, we compute the indices for each arm separately and our index policy for risk-averse MAB will be playing the arm with the highest index at each step. In the process of computing the index for each state, we need to solve risk-averse dynamic programs, presented in (19), using standard tools discussed in [39] such as risk-averse value iteration or policy iteration. Unlike offline algorithms in which we compute the indices in decreasing order of their index values, restart-in-state formulation is an online algorithm and calculates each index by solving separate dynamic programs with no specific order. That is why, it can be generalized to countable state spaces relying on standard dynamic programming extensions to countable state spaces. Moreover, using an online algorithm enables us to only compute indices for a certain portion of states in case knowing the values for all indices is not required.

Algorithm 4 Computing *Risk-averse Indices* for Arm i

```

1: for  $k = 1$  to  $N$  do
2:   Restart Option:
3:   for  $m = 1$  to  $N$  do
4:      $x_m^{(0)} = k$  ▷ initial state vector for restart option
5:     for  $n = 1$  to  $N$  do
6:        $Q_{mn}^{(0)} = Q^i(k, 1, n)$  ▷ transition probability matrix for restart option
7:        $c_{mn}^{(0)} = c^i(k, 1, n)$  ▷ cost matrix for restart option
8:     end for
9:   end for
10:  Continuation Option:
11:  for  $m = 1$  to  $N$  do
12:     $x_m^{(1)} = m$  ▷ initial state vector for continuation option
13:    for  $n = 1$  to  $N$  do
14:       $Q_{mn}^{(1)} = Q^i(m, 1, n)$  ▷ transition probability matrix for continuation option
15:       $c_{mn}^{(1)} = c^i(m, 1, n)$  ▷ cost matrix for continuation option
16:    end for
17:  end for
18:  Dynamic Program:
19:     $v = \min \{ \sigma(c^{(0)} + \beta v, x^{(0)}, Q^{(0)}), \sigma(c^{(1)} + \beta v, x^{(1)}, Q^{(1)}) \}$  (4.23)
20:     $M^i(k) = v(k)$ 
21:  end for
22: return  $M^i(k), k \in \{1, \dots, N\}$ 

```

4.5 Computational Experiments

In an effort to evaluate the performance of the proposed index-based policy, we perform a series of computational experiments. The algorithms are coded in MATLAB R2019a and the experiments are conducted on a PC (Intel 3.60GHz, 64 GB memory) running Windows 10. We seek the following policies for each test instance:

- (1) An optimal policy for our problem, achieved by employing the *risk-averse policy iteration algorithm with convex optimization method* (see, [47]).
- (2) The risk-averse index policy (*RA*) achieved by *risk-averse allocation indices*.
- (3) The risk-neutral policy (*RN*) achieved by Gittins indices.

We consider a MAB consisting of three arms with four states in each arm, leading to a state space of $4^3 = 64$ elements. For each test instance, the transition probabilities of each arm under the action of play are taken from the $U(0, 1)$ distribution and the consequent transition matrix is normalized across rows, that is the entries of each row are divided by their sum. Trying to follow [48], the cost of transitioning from a state to another is sampled from a truncated normal distribution with a mean value of $U(-6, -5)$ and a standard deviation of σ taking values of $\{0.01, 0.5, 1\}$. Experiments are performed with a discount factor of $\beta \in \{0.90, 0.95\}$.

We use first-order mean-semideviation given in (4.2), with $\kappa \in \{0.25, 0.50, 0.75, 1\}$, along with mean-AVaR provided in (4.3), with parameters of $\alpha \in \{0.80, 0.90, 0.95\}$ and $\lambda \in \{0, 0.5\}$, where $\lambda = 0$ represents AVaR risk measure.

Regarding each set of parameters, we generate 1000 test instances at random, and for each generated instance, we are comparing *RA* and *RN* policies on the basis of two separate performance measures:

- (1) the suboptimality percentages of policy $\pi \in \{RA, RN\}$ for each initial state $x \in \mathcal{X}$:

$$100 \times \frac{R(x) - R^\pi(x)}{R(x)},$$

- (2) similarity percentage of policy $\pi \in \{RA, RN\}$ compared to optimal policy π^* :

$$100 \times \frac{\sum_{x \in \mathcal{X}} \mathbb{1}_{(\pi, \pi^*)}(x)}{|\mathcal{X}|},$$

where

$$\mathbb{1}_{(\pi, \pi^*)}(x) := \begin{cases} 1, & \text{if } \pi^*(x) = \pi(x) \\ 0, & \text{otherwise.} \end{cases}$$

For each generated instance, we calculate median and maximum percentages of suboptimality across all states. Table 4.1 summarizes the statistics of maximum suboptimality percentages for the risk measure of first-order mean-semideviation. Distinguished by the third column, we report mean and maximum of maximum suboptimality percentages over all the set of 1000 test instances. With respect to this performance measure, superiority of *RA* policy to *RN* policy is evident. For example, for risk-averse MAB with parameters of $\sigma = 1$, $\beta = 0.90$, and $\kappa = 1$, the maximum value for maximum suboptimality percentage for *RN* and *RA* are 10.23% and 0.114%, respectively. Table 4.1 indicates that by increasing the cost value variability, i.e. σ , a considerably higher suboptimality percentage is observed for both *RN* and *RA* policies. This increase is, however, much more significant for *RN*. In addition, a rise in the degree of risk-aversion (in other words, an increase in the value of κ) leads to higher maximum suboptimality percentage in both policies, but higher increase is for *RN*. Regarding the discount factor β , even though there is no consistent behavior in *RN* policy, reduction in maximum suboptimality percentages of *RA* policy is observed as the discount factor increases.

Table 4.2 reports the median of suboptimality percentages concerning first-order mean-semideviation risk measure. This table, with few non-zero values for *RA* policy, reveals similar pattern as in Table 4.1. Table 4.3 states the average similarity percentage values for this risk measure. As far as these results are

Table 4.1: Statistics of maximum suboptimality percentage of policies RN and RA for first-order mean-semideviation

σ	β		$\kappa = 0.25$		$\kappa = 0.50$		$\kappa = 0.75$		$\kappa = 1$	
			RN	RA	RN	RA	RN	RA	RN	RA
0.01	0.90	mean	0	0	0	0	0	0	0	0
		max	0.006	0	0.024	0	0.039	0	0.053	0.001
	0.95	mean	0	0	0	0	0	0	0	0
		max	0.007	0	0.025	0	0.04	0	0.054	0
0.5	0.90	mean	0.014	0	0.049	0	0.102	0	0.176	0.001
		max	0.692	0.015	1.466	0.028	2.228	0.04	3.005	0.128
	0.95	mean	0.012	0	0.038	0	0.077	0	0.133	0
		max	1.036	0.005	2.113	0.008	3.167	0.022	4.215	0.031
1	0.90	mean	0.043	0	0.158	0.001	0.329	0.001	0.559	0.003
		max	1.806	0.021	4.379	0.055	7.096	0.081	10.232	0.114
	0.95	mean	0.029	0	0.114	0	0.248	0	0.433	0.001
		max	1.665	0.007	3.718	0.014	6.543	0.035	9.883	0.05

concerned, the performance of our risk-averse index policy RA is remarkably good. For example, for $\sigma = 1$, $\beta = 0.95$, $\kappa = 1$, the difference between average similarity percentages of RN and RA is close to 11%.

In addition, we concentrate on for what percentage of instances RN and RA policies offer optimal policy. In this regard, RA is optimal in more than 95% of instances and in no occasion it is more than 0.13% suboptimal, whereas RN is optimal only for 62% of cases and the suboptimality percentage can be more than 10%.

Regarding mean-AVaR risk measure, the results of maximum suboptimality percentage are presented in Table 4.4. Much like in the case of first-order mean-semideviation, increasing variability of cost values, σ , leads to an increase in maximum suboptimality percentage of RN and RA , with more noticeable changes in RN . And, with regard to the increase in the discount factor β , the reduction in the maximum suboptimality percentage is generally observed. On the other hand, increasing the level of risk-aversion by reducing λ or increasing α results in a reduction in the performance of RN and RA policies. The success of our RA policy is much higher than that of RN policy. One can see that, when $\sigma = 1$, $\beta = 0.90$, $\lambda = 0$ and $\alpha = 0.95$, RN experiences the worst suboptimality percentage of 48.297%, whereas this value is 1.101% for RA .

Table 4.2: Statistics of median suboptimality percentage of policies RN and RA for first-order mean-semideviation

σ	β		$\kappa = 0.25$		$\kappa = 0.50$		$\kappa = 0.75$		$\kappa = 1$	
			RN	RA	RN	RA	RN	RA	RN	RA
0.01	0.90	mean	0	0	0	0	0	0	0	0
		max	0.003	0	0.013	0	0.022	0	0.031	0
	0.95	mean	0	0	0	0	0	0	0	0
		max	0.004	0	0.019	0	0.033	0	0.047	0
0.5	0.90	mean	0.004	0	0.015	0	0.035	0	0.067	0
		max	0.307	0	0.959	0	1.709	0.003	2.494	0.011
	0.95	mean	0.006	0	0.02	0	0.043	0	0.077	0
		max	0.721	0	1.58	0	2.435	0.002	3.292	0.002
1	0.90	mean	0.012	0	0.053	0	0.123	0	0.229	0
		max	1.21	0	3.416	0.001	5.642	0.002	8.166	0.005
	0.95	mean	0.014	0	0.062	0	0.144	0	0.263	0
		max	1.124	0	3.186	0	5.872	0.002	8.864	0.003

Table 4.3: Average similarity percentages of policies RN and RA for first-order mean-semideviation

σ	β	$\kappa = 0.25$		$\kappa = 0.50$		$\kappa = 0.75$		$\kappa = 1$	
		RN	RA	RN	RA	RN	RA	RN	RA
0.01	0.90	99.913	100	99.813	99.981	99.75	100	99.675	99.994
	0.95	99.875	99.994	99.788	99.994	99.719	100	99.675	99.988
0.5	0.90	97.844	99.883	95.53	99.82	93.377	99.631	91.072	99.434
	0.95	97.75	99.877	95.509	99.752	93.186	99.508	90.906	99.308
1	0.90	96.939	99.786	93.923	99.598	91.238	99.5	88.183	99.148
	0.95	96.905	99.723	93.88	99.603	91.144	99.345	87.984	98.955

Table 4.4: Statistics of maximum suboptimality percentage of policies RN and RA for mean-AVaR

σ	β		$\lambda = 0$						$\lambda = 0.5$					
			$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$		$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$	
			RN	RA	RN	RA	RN	RA	RN	RA	RN	RA	RN	RA
0.01	0.90	mean	0.002	0	0.003	0	0.003	0	0	0	0.001	0	0.001	0
		max	0.161	0.001	0.24	0.001	0.237	0.002	0.04	0.001	0.068	0.001	0.079	0.001
	0.95	mean	0.002	0	0.002	0	0.002	0	0	0	0	0	0	0
		max	0.156	0	0.226	0	0.225	0	0.036	0	0.064	0	0.077	0
0.5	0.90	mean	1.801	0.023	2.167	0.025	2.357	0.026	0.321	0.003	0.394	0.003	0.427	0.004
		max	21.889	0.248	21.851	0.328	21.628	0.328	4.62	0.059	5.69	0.093	6.017	0.096
	0.95	mean	1.516	0.007	1.811	0.009	2	0.009	0.238	0.001	0.293	0.001	0.318	0.001
		max	21.004	0.065	21.06	0.103	20.898	0.105	5.201	0.022	5.324	0.02	5.598	0.025
1	0.90	mean	5.485	0.07	6.639	0.079	7.245	0.083	1.039	0.007	1.219	0.009	1.323	0.011
		max	42.232	1.033	48.297	1.103	48.297	1.101	13.341	0.141	13.003	0.172	12.69	0.357
	0.95	mean	4.845	0.026	5.898	0.026	6.516	0.029	0.802	0.003	0.949	0.004	1.026	0.004
		max	40.538	0.545	48.451	0.338	48.451	0.303	12.879	0.045	13.227	0.087	12.597	0.104

Similarly, median suboptimality percentage results for mean-AVaR are presented in Table 4.5. This table reveals similar trends as in Table 4.4. Also, regarding this risk measure, the average similarity percentages are reported in Table 4.6. With respect to all the performance measures, we observe superiority of RA compared to RN . Although the average similarity percentage for RA under any parameter setting we take into account is more than 92%, it can decline to around 64% while dealing with RN .

Overall, the performance of our index policy RA is higher than that of RN when using mean-AVaR as risk measure. This superiority is easier to understand if we focus on the maximum values in Table 4.4 for $\sigma = 1$, where the percentage of RN suboptimality is more than 70 times that of RA in general. Specifically, for $\lambda = 0$, $\alpha = 0.95$, and $\beta = 0.95$, the maximum values for maximum suboptimality percentage of RN and RA are 48.451% and 0.303%, respectively.

Furthermore, our index policy is optimal in more than 91% of cases and is in no occasion worse than 1.2% suboptimal, whereas RN policy is optimal in only 41.5% of instances and could have a suboptimal percentage of about 48.5%.

In addition, we conduct another experiment to visualize the time efficiency of

Table 4.5: Statistics of median suboptimality percentage of policies RN and RA for mean-AVaR

σ	β		$\lambda = 0$						$\lambda = 0.5$					
			$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$		$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$	
			RN	RA	RN	RA	RN	RA	RN	RA	RN	RA	RN	RA
0.01	0.90	mean	0.001	0	0.001	0	0.001	0	0	0	0	0	0	0
		max	0.139	0	0.139	0	0.139	0	0.029	0	0.05	0	0.06	0
	0.95	mean	0.001	0	0.001	0	0.001	0	0	0	0	0	0	0
		max	0.145	0	0.146	0	0.146	0	0.031	0	0.055	0	0.067	0
0.5	0.90	mean	0.971	0	1.15	0	1.26	0	0.127	0	0.165	0	0.183	0
		max	16.209	0.02	17.034	0.031	16.215	0.028	3.149	0.002	3.989	0.002	4.289	0.004
	0.95	mean	1.029	0	1.21	0	1.329	0	0.143	0	0.183	0	0.199	0
		max	18.083	0.02	19.028	0.008	18.116	0.01	4.223	0.005	4.357	0.002	4.716	0.002
1	0.90	mean	3.069	0.001	3.627	0	4.07	0	0.454	0	0.538	0	0.587	0
		max	33.744	0.1	37.009	0.08	36.193	0.076	8.761	0.004	8.268	0.005	8.588	0.007
	0.95	mean	3.391	0	4.005	0	4.473	0	0.498	0	0.599	0	0.648	0
		max	36.129	0.035	42.359	0.037	41.165	0.047	10.473	0.002	11.007	0.005	10.416	0.007

Table 4.6: Average similarity percentages of policies RN and RA for mean-AVaR

σ	β	$\lambda = 0$						$\lambda = 0.5$					
		$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$		$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$	
		RN	RA	RN	RA	RN	RA	RN	RA	RN	RA	RN	RA
0.01	0.90	98.615	99.937	98.536	99.921	98.552	99.937	99.307	99.937	99.213	99.906	99.15	99.89
	0.95	98.615	99.921	98.567	99.89	98.552	99.937	99.276	99.921	99.213	99.89	99.134	99.874
0.5	0.90	75.445	96.174	73.878	95.942	73.04	95.907	87.232	98.992	85.827	98.89	85.056	98.717
	0.95	74.921	95.497	73.465	95.061	72.717	94.923	87.126	98.744	85.902	98.776	85.052	98.497
1	0.90	68.002	94.561	66.148	94.163	64.948	93.789	82.958	98.642	81.065	98.351	79.987	98.166
	0.95	67.522	93.474	65.432	92.947	64.405	92.302	82.517	98.272	80.801	97.957	79.66	97.835

Table 4.7: Description of problem sizes

Problem size	1	2	3	4	5	6	7	8	9	10	11	12
No. of arms	2	2	2	3	2	2	2	3	2	4	2	3
No. of states in each arm	3	4	5	3	6	7	8	4	9	3	10	5
State space (MDP)	9	16	25	27	36	49	64	64	81	81	100	125
State space (<i>RAI</i> policy)	6	8	10	9	12	14	16	12	18	12	20	15

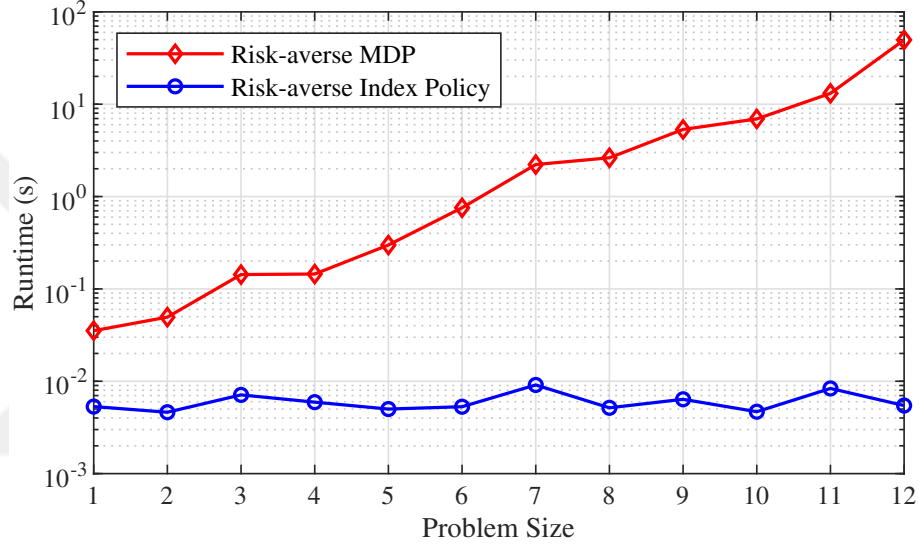


Figure 4.1: Runtime of obtaining the optimal policy via solving risk-averse MDP along with executing risk-averse index heuristic versus the size of the risk-averse MAB for first-order mean-semideviation risk measure.

our index-based policy against optimally solving the corresponding MDP. Figures 4.1 and 4.2 show the CPU running time of these two algorithms for 12 problem sizes with state space of 9 to 125 elements for the two risk measures of first-order semi-deviation and mean-AVaR, respectively. The details of these problem sizes are described in Table 4.7. These figures are on a logarithmic scale in order to demonstrate better comparison. They indicate that the computation of our risk-averse index-based heuristic takes considerably less time than finding an optimal policy. By increasing the size of the problem, the computation time of our risk-averse index-based policy seems to increase linearly, whereas the computation time of the corresponding risk-averse MDP displays exponential growth.

We also realize that an optimal solution of risk-averse MAB may not be index-based. The optimal solution for first-order mean-semideviation and mean-AVaR

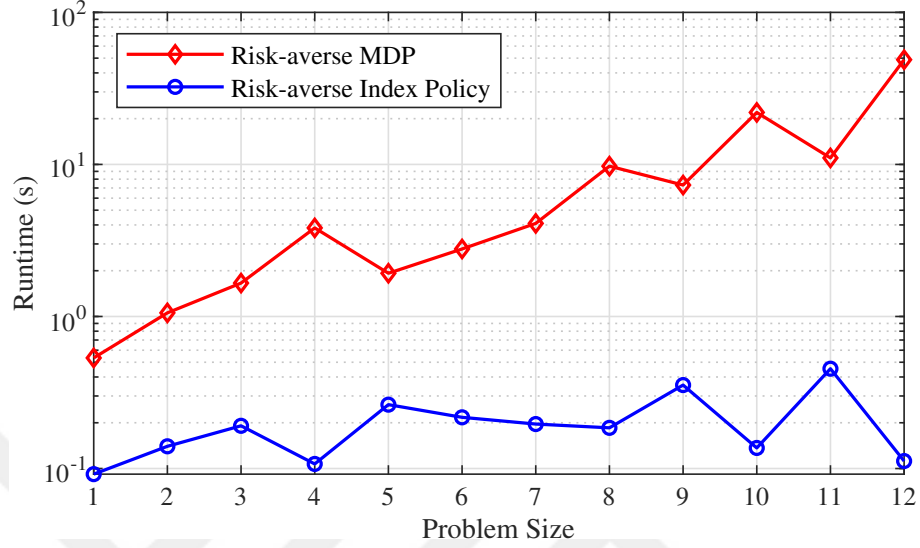


Figure 4.2: Runtime of obtaining the optimal policy via solving risk-averse MDP along with executing risk-averse index heuristic versus the size of the risk-averse MAB for Mean-AVaR risk measure.

risk measures was not achieved in the form of index policy in 0.454% and 0.202% of cases, respectively.

Finally, we compare the performance of the heuristic presented in Chapter 3, here referred to as *RA1*, with the heuristic presented in this chapter, here referred to as *RA2*, using the same setting employed in Section 3.6. Tables 4.8 and 4.9 summarize the results of maximum suboptimality percentage and similarity percentages for the risk measure of first-order mean-semideviation, respectively. Moreover, Tables 4.10 and 4.11 summarize the results of maximum suboptimality percentage and similarity percentages for the risk measure of mean-AVaR, respectively. We observe that in both risk measures *RA2* is outperforming *RA1* in terms of maximum suboptimality percentage and similarity percentages.

4.6 Conclusion

In this chapter, we analyze a risk-averse formulation of the multi-armed bandit problem. This model helps decision makers with different extent of risk-aversion

Table 4.8: Comparison of maximum suboptimality percentage of policies RA1 and RA2 for first-order mean-semideviation

σ	β		$\kappa = 0.25$		$\kappa = 0.50$		$\kappa = 0.75$		$\kappa = 1$	
			RA1	RA2	RA1	RA2	RA1	RA2	RA1	RA2
0.01	0.90	mean	0	0	0	0	0	0	0	0
		max	0	0	0.002	0	0.001	0	0.006	0
	0.95	mean	0	0	0	0	0	0	0	0
		max	0.001	0	0.002	0	0.003	0	0.002	0
0.5	0.90	mean	0.003	0	0.011	0	0.019	0	0.032	0
		max	0.356	0.01	0.953	0.011	1.136	0.03	1.8	0.025
	0.95	mean	0.003	0	0.01	0	0.02	0	0.03	0
		max	0.487	0.003	1.143	0.005	1.769	0.009	2.362	0.007
1	0.90	mean	0.01	0	0.031	0	0.059	0.001	0.087	0.001
		max	1.103	0.009	2.1	0.052	3.209	0.105	3.813	0.193
	0.95	mean	0.008	0	0.023	0	0.048	0	0.067	0
		max	1.167	0.005	1.87	0.02	3.134	0.035	4.535	0.064

Table 4.9: Average similarity percentages of policies RA1 and RA2 for first-order mean-semideviation

σ	β		$\kappa = 0.25$		$\kappa = 0.50$		$\kappa = 0.75$		$\kappa = 1$	
			RA1	RA2	RA1	RA2	RA1	RA2	RA1	RA2
0.01	0.90	100	100	99.981	100	99.977	100	99.953	99.988	
	0.95	99.988	100	99.981	99.994	99.969	100	99.947	99.994	
0.5	0.90	99.603	99.905	99.217	99.819	98.844	99.739	98.459	99.603	
	0.95	99.58	99.883	99.078	99.778	98.628	99.684	98.291	99.572	
1	0.90	99.514	99.911	98.969	99.798	98.502	99.655	98.291	99.603	
	0.95	99.459	99.875	98.953	99.744	98.358	99.584	98.028	99.522	

to play the arm with the best risk adjusted total discounted return. Our study shows that, due to subadditivity of our risk measure, MAB with dynamic risk measures may not admit an optimal index-based policy. However, we propose a novel priority-index heuristic, which converges to Gittins index when the extent of risk-aversion goes to zero. In our numerical experiments, we use two well known coherent risk measures of *first order mean-semideviation* and *mean-AVaR* to evaluate the performance of our index policy. The experiments reveal that our index-based policy has much lower percentages of suboptimality and is more comparable to the optimal policy relative to the Gittins index policy.

Table 4.10: Comparison of maximum suboptimality percentage of policies RA1 and RA2 for mean-AVaR

σ	β		$\lambda = 0$						$\lambda = 0.5$					
			$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$		$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$	
			RA1	RA2	RA1	RA2	RA1	RA2	RA1	RA2	RA1	RA2	RA1	RA2
0.01	0.90	mean	0	0	0	0	0	0	0	0	0	0	0	0
		max	0	0	0	0	0	0	0	0	0	0	0	0
	0.95	mean	0	0	0	0	0	0	0	0	0	0	0	0
		max	0	0	0	0	0	0	0	0	0	0	0	0
0.5	0.90	mean	0.028	0.002	0.006	0.001	0	0	0.018	0	0.002	0	0.002	0
		max	4.677	0.087	1.323	0.066	0.044	0.044	1.979	0.113	0.232	0.064	0.237	0
	0.95	mean	0.021	0.001	0.003	0	0	0	0.015	0	0.006	0	0.001	0
		max	4.498	0.059	0.749	0.029	0.012	0.012	1.641	0.034	2.018	0.025	0.157	0.005
1	0.90	mean	0.027	0.002	0.012	0.001	0.001	0	0.024	0	0.005	0	0.005	0
		max	7.538	0.127	2.761	0.075	0.285	0.041	1.817	0.064	0.828	0.053	0.999	0.003
	0.95	mean	0.022	0.001	0.006	0.001	0.001	0	0.02	0	0.007	0	0.005	0
		max	7.05	0.108	1.44	0.056	0.133	0.011	1.251	0.02	1.022	0.017	1.173	0.01

Table 4.11: Average similarity percentages of policies RA1 and RA2 for mean-AVaR

σ	β	$\lambda = 0$						$\lambda = 0.5$					
		$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$		$\alpha = 0.80$		$\alpha = 0.90$		$\alpha = 0.95$	
		RA1	RA2	RA1	RA2	RA1	RA2	RA1	RA2	RA1	RA2	RA1	RA2
0.01	0.90	99.984	99.984	99.984	99.984	100	100	99.968	99.984	99.984	99.984	100	100
	0.95	99.984	99.984	99.984	99.984	100	100	99.984	99.984	99.984	99.984	100	100
0.5	0.90	99.543	99.722	99.881	99.897	99.952	99.952	99.579	99.948	99.777	99.96	99.932	100
	0.95	99.36	99.535	99.773	99.789	99.948	99.948	99.535	99.928	99.781	99.968	99.889	99.98
1	0.90	99.622	99.726	99.805	99.861	99.972	99.988	99.63	99.94	99.809	99.964	99.897	99.976
	0.95	99.443	99.575	99.73	99.785	99.928	99.956	99.594	99.936	99.845	99.956	99.881	99.976

Chapter 5

Multi-armed Bandit Designs with Dynamic Risk Measures for Clinical Trials

The objective of risk-neutral MAB for clinical trial designs is to maximize the expected total number of effective treatments. However, this may not be the most desirable objective as the treatment that works best on average may also have significant variability, resulting in negative side effects for some patients. That's why, in this chapter, we apply the proposed approach in Chapter 4 to perform risk-averse treatment allocation based on a Bayesian Bernoulli model. We evaluate the performance of this approach against other allocation rules, including fixed randomization.

5.1 Problem Description and Preliminaries

In classical risk-neutral multi-armed bandit problem (MAB), a decision maker faces a set of available arms $\mathcal{K} := \{1, 2, \dots, K\}$ and determines which one to play at each step $t \in \mathbb{N}$, where $\mathbb{N} = \{1, 2, 3, \dots\}$. After playing each arm, a

reward will be obtained and the objective would be to maximize the expected total discounted reward.

In this chapter, we consider a risk-averse Bayesian Bernoulli MAB (BB-MAB) with dynamic risk measures, where each arm represents a treatment with an unknown success rate, each decision step is an incoming patient who will receive a treatment, and consequently each play will be assigning a treatment to a patient. At each time t , playing arm $i \in \mathcal{K}$ can be considered as drawing a sample observation from a Bernoulli population with unknown success probability p^i , with which we earn a reward that can be either 0 or 1. For mathematical convenience, we consider negative of rewards that could be translated as costs, therefore we consider that costs of 0 and -1 are incurred. The Bayesian feature is incorporated by allowing each parameter p^i to have a Beta prior with parameters $s_1^i \in \mathbb{N}$ and $f_1^i \in \mathbb{N}$ at the initial time (that is, at $t = 1$). For any step $t > 1$, after observing $s_t^i \in \mathbb{N}$ successes and $f_t^i \in \mathbb{N}$ failures, the posterior density is a Beta distribution with parameters $(s_1^i + s_t^i, f_1^i + f_t^i)$:

$$f(p^i | s_1^i + s_t^i, f_1^i + f_t^i) = \frac{\Gamma(s_1^i + s_t^i + f_1^i + f_t^i)}{\Gamma(s_1^i + s_t^i)\Gamma(f_1^i + f_t^i)} (p^i)^{s_1^i + s_t^i - 1} (1 - p^i)^{f_1^i + f_t^i - 1}, p^i \in [0, 1], \quad (5.1)$$

and the distribution of each arm is independent of others. In general, the probability of success at each trial for arm i in state $(s_1^i + s_t^i, f_1^i + f_t^i)$ is $\frac{s_1^i + s_t^i}{s_1^i + s_t^i + f_1^i + f_t^i}$, that is the mean of a $\text{Beta}(s_1^i + s_t^i, f_1^i + f_t^i)$ distribution. Similarly, one can compute the probability of failure, which is $\frac{f_1^i + f_t^i}{s_1^i + s_t^i + f_1^i + f_t^i}$. Figure 5.1 illustrates the resulting Markov chain process in each arm i with the priors of (s_1^i, f_1^i) .

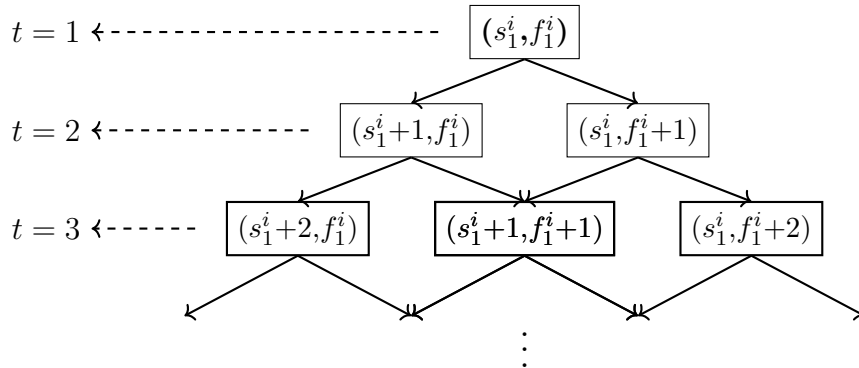


Figure 5.1: Markov chain process of arm i .

We consider $\beta \in (0, 1)$ as the discount factor, therefore $1 - \beta$ can be interpreted as the probability that the entire experiment will be terminated at any given time. The problem setting for our BB-MAB is defined with respect to the following elements:

(i) With the assumption that there exists no common state in distinct arms, let $\mathcal{X} = \times_{i=1}^K \mathcal{X}^i$ be the *state space* of the resulting MDP, where each arm $i \in \mathcal{K}$ is a Markov chain with an infinite state space $\mathcal{X}^i = \{(s_1^i + s_t^i, f_1^i + f_t^i) \in \mathbb{N}^2 \mid s_t^i, f_t^i \geq 0\}$. At each time $t \in \mathbb{N}$, the state space of arm i is all the possible number of successes and failures with respect to the initial prior. Thus, $\mathcal{X}_t^i = \{(s_1^i + s_t^i, f_1^i + f_t^i) \in \mathbb{N}^2 \mid s_t^i, f_t^i \geq 0, s_t^i + f_t^i < t\}$. We also denote the state arm of i at decision step t as $x_t^i = (s_1^i + s_t^i, f_1^i + f_t^i)$ with the initial prior as $x_1^i = (s_1^i, f_1^i)$.

(ii) A finite action space $\mathcal{U}$ and action sets $U(x) \subseteq \mathcal{U}$, $x \in \mathcal{X}$, where $U(x)$ is a nonempty set of possible actions at each state $x \in \mathcal{X}$. At step $t \in \mathbb{N}$, an *action* $u_t = \{u_t^1, u_t^2, \dots, u_t^K\}$ is applied, where u_t^i is the action of assigning ($u_t^i = 1$) or not assigning ($u_t^i = 0$) patient (step) t to treatment (arm) $i \in \mathcal{K}$. The *admissible actions* are the ones for which exactly one element of u_t is one at each step $t \in \mathbb{N}$. That is, for each patient, under any choice of admissible actions, only one treatment can be assigned.

(iii) The state of arm i that we play evolves in a Markovian fashion from step t to $t + 1$ with respect to the transition probabilities $P^i(x_{t+1}^i \mid x_t^i)$, and it is given by

$$x_{t+1}^i = \begin{cases} (s_1^i + s_t^i + 1, f_1^i + f_t^i), & \text{w.p. } \frac{s_1^i + s_t^i}{s_1^i + s_t^i + f_1^i + f_t^i} \\ (s_1^i + s_t^i, f_1^i + f_t^i + 1), & \text{w.p. } \frac{f_1^i + f_t^i}{s_1^i + s_t^i + f_1^i + f_t^i}. \end{cases}$$

However, the state of an arm remains unchanged if it is not played, that is, we have $x_{t+1}^i = x_t^i$ if $u_t^i = 0$ for $i \in \mathcal{K}$, $t \in \mathbb{N}$. Also, we denote controlled transition kernel $Q(x, u, y)$ as the transition probability from state x to y , when action u is applied. Therefore, we have $Q(x_t, u_t, x_{t+1}) = P^i(x_{t+1}^i \mid x_t^i)$, where $u_t = \{0, \dots, 0, 1, 0, \dots, 0\}$ with the nonzero component at position i , that is,

$i \in \mathcal{K}$ is the only treatment we assign at step $t \in \mathbb{N}$. Moreover, we denote $Q^i(x_t^i, 1, x_{t+1}^i) = P^i(x_{t+1}^i \mid x_t^i)$.

(iv) For each $i \in \mathcal{K}$, the *cost function* $g_t^i : \mathcal{X}_t^i \times \mathcal{X}_{t+1}^i \rightarrow \{0, -1\}$ is the cost of transition from step t to step $t + 1$ by playing arm i . So, we have:

$$g_t^i(x_t^i, x_{t+1}^i) = \begin{cases} -1, & \text{if } x_{t+1}^i = (s_1^i + s_t^i + 1, f_1^i + f_t^i) \\ 0, & \text{if } x_{t+1}^i = (s_1^i + s_t^i, f_1^i + f_t^i + 1). \end{cases} \quad (5.2)$$

We define $c^i(x_t^i, u_t^i, x_{t+1}^i)$ as the cost obtained by arm i if it transitions from state $x_t^i \in \mathcal{X}_t^i$ to state $x_{t+1}^i \in \mathcal{X}_{t+1}^i$ under action $u_t^i \in \{0, 1\}$, $i \in \mathcal{K}$:

$$c^i(x_t^i, u_t^i, x_{t+1}^i) := \begin{cases} g_t^i(x_t^i, x_{t+1}^i), & \text{if } u_t^i = 1 \\ 0, & \text{if } u_t^i = 0. \end{cases} \quad (5.3)$$

We also define $c(x, u, y)$ representing the cost over all arms, that is $c(x, u, y) = \sum_{i \in \mathcal{K}} c^i(x^i, u^i, y^i)$, $x = (x^1, \dots, x^K) \in \mathcal{X}$, and $u = (u^1, \dots, u^K)$ where $u^i \in \{0, 1\}$, $i \in \mathcal{K}$.

(v) We denote a stationary (Markov) *policy* as $\pi : \mathcal{X} \rightarrow \{0, 1\}^K$. Note that, with respect to our notation, $u_t = \pi(x_t)$, that is, both denote the decision to be taken at state $x_t \in \mathcal{X}$ for $t \in \mathbb{N}$. We also define $\pi^i : \mathcal{X}^i \rightarrow \{0, 1\}$ which prescribes the action to be taken for arm $i \in \mathcal{K}$ at a particular state.

Here, we search for a stationary Markov policy, π , that minimizes the total risk-averse discounted cost incurred over an infinite horizon. With respect to the theory of [39], an infinite horizon stationary MDP with dynamic risk measures has a stationary optimal policy and it is possible to find this policy using *value iteration* or *policy iteration* algorithms. However, these algorithms may be computationally expensive or even practically infeasible for large-scale real-life problems. Furthermore, in some cases, the structure of optimal policy makes it difficult to interpret, which creates complication while applying them. That's why, we take advantage of an index-based heuristic introduced in Chapter 4 for finding optimal or near-optimal policies for the total risk-averse discounted outcome with dynamic risk measures over an infinite horizon.

Before providing the model, we first need to introduce dynamic risk measures. Similar to Chapter 4, we consider 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 $\mathcal{Z}_1 = \mathbb{R}$. Due to the finiteness of each $\mathcal{F}_t$, in our case the spaces $\mathcal{Z}_t$ is identified with finite dimensional vector spaces.

In this context, $\rho_t : \mathcal{Z}_{t+1} \rightarrow \mathcal{Z}_t$, $t \in \{1, \dots, T\}$ satisfying the following axioms is defined as *one-step conditional risk measure* (see [41, 42]):

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

Conditions (A1)-(A4) are called *convexity*, *monotonicity*, *translation invariance* and *positive homogeneity*, respectively. They are analogous to the axioms introduced by [38] for coherent measures of risk.

Two important coherent risk measures, which have been extensively studied in the literature, are *first-order mean-semideviation* [43] and *mean-average value-at-risk* (mean-AVaR) [44]. We now provide the conditional versions of these two risk measures. Conditional first-order mean-semideviation risk measure is defined as

$$\rho_t(Z_{t+1}) = \mathbb{E}[Z_{t+1} | \mathcal{F}_t] + \kappa \mathbb{E}[(Z_{t+1} - \mathbb{E}[Z_{t+1} | \mathcal{F}_t])_+ | \mathcal{F}_t], \quad (5.4)$$

where $\kappa \in [0, 1]$ and $(a)_+ := \max\{a, 0\}$ for $a \in \mathbb{R}$. Given two parameters of $\alpha \in (0, 1)$ and $\lambda \in [0, 1]$, conditional mean-AVaR risk measure is defined as

$$\rho_t(Z_{t+1}) = \lambda \mathbb{E}[Z_{t+1} | \mathcal{F}_t] + (1 - \lambda) \text{AVaR}_\alpha(Z_{t+1} | \mathcal{F}_t), \quad (5.5)$$

where conditional AVaR_α can be represented as

$$\text{AVaR}_\alpha(Z_{t+1}|\mathcal{F}_t) = \min_{\eta \in \mathcal{Z}_t} \left\{ \eta + \frac{1}{1-\alpha} \mathbb{E}[(Z_{t+1} - \eta)_+ | \mathcal{F}_t] \right\}.$$

A *dynamic risk measure* is a sequence of one-step conditional risk measures [39, and references therein]. Thereby, dynamic risk measure $\varrho_{1,T}^\beta : \mathcal{Z}_{1,T} \mapsto \mathcal{Z}_1$ on a finite horizon of length T with discount factor $\beta \in (0, 1)$ is defined as follows:

$$\varrho_{1,T}^\beta(Z_2, Z_3, \dots, Z_{T+1}) := \rho_1(Z_2 + \rho_2(\beta Z_3 + \dots + \rho_T(\beta^{T-1} Z_{T+1}) \dots)). \quad (5.6)$$

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

$$\varrho^\beta(Z_2, Z_3, Z_4, \dots) = \lim_{T \rightarrow \infty} \varrho_{1,T}^\beta(Z_2, Z_3, \dots, Z_{T+1}). \quad (5.7)$$

Applying the dynamic risk measure (5.7), we are able to evaluate the risk of the cost sequences $c(x_t, u_t, x_{t+1}) \in \mathcal{Z}_{t+1}$, $t \in \mathbb{N}$ resulted by policy π . The performance objective in the risk-averse BB-MAB is to minimize the risk-adjusted total discounted negative number of successes. Considering $R(x_1)$ as this optimal value, we have

$$\begin{aligned} R(x_1) &= \min_{\pi \in \Pi} \varrho^\beta(c(x_1, u_1, x_2), c(x_2, u_2, x_3), c(x_3, u_3, x_4), \dots) \\ &= \min_{\pi \in \Pi} \rho_1 \left(\sum_{i \in \mathcal{K}} c^i(x_1^i, u_1^i, x_2^i) + \rho_2 \left(\beta \sum_{i \in \mathcal{K}} c^i(x_2^i, u_2^i, x_3^i) + \dots \right) \right), \end{aligned} \quad (5.8)$$

where x_1 is the initial state and Π is the set of all the feasible policies for our problem, that is, the class of policies that only allow allocation of one treatment for each patient. We can solve (5.8) with the help of the following dynamic programming equations of MDP with dynamic risk measures [39, 47, 53]:

$$R(x_1) = \min_{\pi \in \Pi} \sigma(c(x_1, u_1, x_2) + \beta R(x_2), x_1, Q(x_1, u_1, x_2)), \quad (5.9)$$

where σ is a *risk transition mapping*. Let $\mathcal{V}$ be the space of all real functions on $\mathcal{U} \times \mathcal{X}$. Then, $\sigma : \mathcal{V} \times \{(x, Q(x, u, y)) \mid u \in U(x), x \in \mathcal{X}, y \in \mathcal{X}\} \mapsto \mathbb{R}$ is a *risk transition mapping* if for every $x \in \mathcal{X}$ and $u \in U(x)$, the function $\varrho \mapsto \sigma(\varrho, x, m)$ is a coherent measure of risk on $\mathcal{V}$ (for details, see [39, 47, 53]). Considering that

the *risk transition mapping* is coherent and finite valued with respect to the first argument, it is also continuous with respect to this argument. Therefore, there exists a closed convex set $\mathcal{A}(x, m)$ such that for all $v \in \mathcal{V}$ we have the following dual representation (see [54, Theorem 2.2] for details):

$$\sigma(\varrho, x, m) = \sup_{\mu \in \mathcal{A}(x, m)} \mathbb{E}_{\mu} [\nu].$$

5.2 An Index Policy for Risk-averse MAB

In Chapter 4, by modifying the approach introduced by [19], we consider retirement problem in order to design good index policies for risk-averse MAB. *Risk-averse retirement process* is the same as risk-averse MAB, only with additional option of retiring from the process at any time by paying cost of $M \in \mathbb{R}^-$. Based on this setting, the risk of the cost sequences of policy π with retirement payment of M is represented by:

$$R^{\pi}(x_1, M) = \varrho_{1,\tau}^{\beta}(c(x_1, u_1, x_2), c(x_2, u_2, x_3), \dots, c(x_{\tau-1}, u_{\tau-1}, x_{\tau}), M), \quad (5.10)$$

where $x_1 \in \mathcal{X}$ denotes the initial state, and τ is the time of retiring from the process. We denote $R(x_1, M)$ as the minimum value for the risk-adjusted total discounted cost over all feasible policies for the retirement problem with the retirement cost of $M \in \mathbb{R}^-$. That is, for each patient only one treatment will be used until the retirement. Considering Π_M as the class of stationary admissible policies for the retirement problem with M as the retirement cost, we have

$$\begin{aligned} R(x_1, M) &= \min_{\pi \in \Pi_M} R^{\pi}(x_1, M) \\ &= \min\{M, \min_{\pi \in \Pi} (\sigma(c(x_1, u_1, x_2) + \beta R(x_2, M), x_1, Q(x_1, u_1, x_2)))\}. \end{aligned} \quad (5.11)$$

Focusing on a single arm $i \in \mathcal{K}$, let $R^i(x_1^i, M)$ denote the minimal risk-adjusted total discounted cost with the initial state of $x_1^i \in \mathcal{X}^i$. Then, the following optimality equation holds:

$$R^i(x_1^i, M) = \min\{M, \sigma(c^i(x_1^i, 1, x_2^i) + \beta R^i(x_2^i, M), x_1^i, Q^i(x_1^i, 1, x_2^i))\}. \quad (5.12)$$

Finally, in Chapter 4, we define the index value for each state x_1^i as $M^i(x_1^i)$ which is the largest value of M at which it is optimal to retire at x_1^i . That is,

$$M^i(x_1^i) := \max \{ M \in \mathbb{R}^- \mid R^i(x_1^i, M) = M \}.$$

Moreover, in Chapter 4, we introduce the index-based heuristic by generalizing the indexability of each single arm bandit as follows:

Definition 5.2.1. For $t \in \mathbb{N}$ and the state of $x_t = (x_t^1, x_t^2, \dots, x_t^K)$, retire if $M^i(x_t^i) \geq M$ for all $i \in \mathcal{K}$, and play arm j if $\min_{i \in \mathcal{K}} M^i(x_t^i) = M^j(x_t^j) \leq M$.

In order to compute these risk-averse indices, one needs to separately consider each arm i as a single arm bandit process. In Chapter 4, we show that in order to compute these indices, one can relate the retirement problem with initial state x to a restart-in-state- x formulation, in which the option of retiring is replaced with the equivalent option of restarting in state x , and indices are computed based on the optimal value function of state x in this new formulation. This approach is restated in the following theorem:

Theorem 5.2.1. *Optimal value function for state x in the restart-in-state- x problem is equal to the risk-averse allocation index of state x for risk-averse MAB with retirement option.*

5.3 Computation of Risk-averse Sequential Allocation Rules in Clinical Trials

So far, we discussed a specific characterization of the risk-averse indices. The purpose of this section is to illustrate the computation of these indices in the context of the clinical trials problem based on this characterization. This characterization casts the index computation into a specific form. Here, we follow the work of [60] and are only concerned with a single arm i . Let Π_R^i be the class of policies in which we either continue to play arm i or to instantaneously restart

the problem at its initial state x_1^i . Then, according to Theorem 5.2.1,

$$M^i(x_1^i) = \min_{\pi^i \in \Pi_R^i} \varrho^\beta(c^i(x_1^i, u_1^i, x_2^i), c^i(x_2^i, u_2^i, x_3^i), c^i(x_3^i, u_3^i, x_4^i), \dots). \quad (5.13)$$

Next, we will elaborate on how to use Eq. (5.13) in order to estimate $M^i(x_1^i)$. Since computation of $M^i(x_t^i)$ when $x_t^i = (s_1^i + s_t^i, f_1^i + f_t^i)$ is essentially the same as computing $M^i(x_1^i)$ with $x_1^i = (s_1^i, f_1^i)$ by changing the prior numbers of success and failure from (s_1^i, f_1^i) to $(s_1^i + s_t^i, f_1^i + f_t^i)$, without loss of generality we would only discuss the calculation of $M^i(x_1^i)$. In order to solve Eq. (5.13) for the initial state $x_1^i = (s_1^i, f_1^i)$, we first need to solve the following dynamic programming equations:

$$\begin{aligned} v^i(x_t^i) = \min \{ & \sigma(c^i(x_t^i, 1, x_{t+1}^i) + \beta v^i(x_{t+1}^i), x_t^i, Q^i(x_t^i, 1, x_{t+1}^i)), \\ & \sigma(c^i(x_1^i, 1, x_2^i) + \beta v^i(x_2^i), x_1^i, Q^i(x_1^i, 1, x_2^i)) \}, \text{ for all } x_t^i \in \mathcal{X}^i. \end{aligned} \quad (5.14)$$

Before going any further, similar to [39], we define the operator $\mathfrak{D} : \mathcal{V} \mapsto \mathcal{V}$ as follows:

$$\begin{aligned} \mathfrak{D}v^i(x_t^i) = \min \{ & \sigma(c^i(x_t^i, 1, x_{t+1}^i) + \beta v^i(x_{t+1}^i), x_t^i, Q^i(x_t^i, 1, x_{t+1}^i)), \\ & \sigma(c^i(x_1^i, 1, x_2^i) + \beta v^i(x_2^i), x_1^i, Q^i(x_1^i, 1, x_2^i)) \}, \text{ for all } x_t^i \in \mathcal{X}^i. \end{aligned} \quad (5.15)$$

With this notation, Eq. (5.14) can be compactly written as

$$v^i = \mathfrak{D}v^i. \quad (5.16)$$

Given the optimal values of $v^i(x_t^i)$ for all $x_t^i \in \mathcal{X}^i$ in Eq. (5.14), then $M^i(x_1^i) = v^i(x_1^i)$. However, solving Eq. (5.14) involves an infinite number of variables. In this regard, we introduce two systems of equations by truncating the length of the time horizon and considering $\mathcal{X}^i = \mathcal{X}_T^i$, where $\mathcal{X}_T^i = \{(s_1^i + s_t^i, f_1^i + f_t^i) \in \mathcal{X}^i \mid 0 \leq s_t^i + f_t^i < T\}$ is a finite subset of $\mathcal{X}^i$:

$$\begin{cases} u_T^i(x_t^i) = \min \{ \sigma(c^i(x_t^i, 1, x_{t+1}^i) + \beta u_T^i(x_{t+1}^i), x_t^i, Q^i(x_t^i, 1, x_{t+1}^i)), \\ \quad \sigma(c^i(x_1^i, 1, x_2^i) + \beta u_T^i(x_2^i), x_1^i, Q^i(x_1^i, 1, x_2^i)) \}, & s_t^i + f_t^i < T - 1, \\ u_T^i(x_t^i) = -\frac{1}{1 - \beta}, & s_t^i + f_t^i = T - 1. \end{cases} \quad (5.17)$$

$$\begin{cases} U_T^i(x_t^i) = \min \left\{ \sigma \left(c^i(x_t^i, 1, x_{t+1}^i) + \beta U_T^i(x_{t+1}^i), x_t^i, Q^i(x_t^i, 1, x_{t+1}^i) \right), \right. \\ \quad \left. \sigma \left(c^i(x_1^i, 1, x_2^i) + \beta U_T^i(x_2^i), x_1^i, Q^i(x_1^i, 1, x_2^i) \right) \right\}, & s_t^i + f_t^i < T - 1, \\ U_T^i(x_t^i) = w^i(x_t^i), & s_t^i + f_t^i = T - 1. \end{cases} \quad (5.18)$$

We will employ the following compact notations for Eq. (5.17) and Eq. (5.18):

$$u_T^i = \mathfrak{D}_1 u_T^i, \quad (5.19)$$

$$U_T^i = \mathfrak{D}_2 U_T^i. \quad (5.20)$$

Note that $w^i(x_t^i) \in [-1/1-\beta, 0]$ and depends on the dynamic risk measure employed. $w^i(x_t^i)$ can be defined as the risk-adjusted discounted cost achieved by the suboptimal policy of never restarting the process in state x_1^i when the initial state is x_t^i . With the aim of finding proper expressions for $w^i(x_t^i)$ for certain risk measures, first we provide a structural property of this value function in Lemma 5.3.1. Then, in Propositions 5.3.1 and 5.3.2, we provide $w^i(x_t^i)$ for risk measures of first-order mean-semideviation and mean-AVaR, respectively.

Lemma 5.3.1. *For $t \in \mathbb{N}$, we have $w^i(s_1^i + s_t^i + 1, f_1^i + f_t^i) - w^i(s_1^i + s_t^i, f_1^i + f_t^i + 1) \leq 1/\beta$.*

Proof. Let $y_{t+1}^i := (s_1^i + s_t^i + 1, f_1^i + f_t^i)$ and $z_{t+1}^i := (s_1^i + s_t^i, f_1^i + f_t^i + 1)$. Also, let $x_{t+2}^{i,(y)}$ and $x_{t+2}^{i,(z)}$ as the states to which the process transits at $t+2$ from the states of y_{t+1}^i and z_{t+1}^i , respectively. Thus, we need to show $w^i(y_{t+1}^i) - w^i(z_{t+1}^i) \leq 1/\beta$ and we show this by backward induction. The base case holds as $w^i(y_{T+1}^i) - w^i(z_{T+1}^i) = 0 \leq 1/\beta$. Assuming $w^i(y_{t+2}^i) - w^i(z_{t+2}^i) \leq 1/\beta$, we show $w^i(y_{t+1}^i) - w^i(z_{t+1}^i) \leq 1/\beta$:

$$\begin{aligned} -1/\beta + w^i(y_{t+1}^i) &= -1/\beta + \sigma \left(c^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}) + \beta w^i(x_{t+2}^{i,(y)}), y_{t+1}^i, Q^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}) \right) \\ &= \sigma \left(-1/\beta + c^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}) + \beta w^i(x_{t+2}^{i,(y)}), y_{t+1}^i, Q^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}) \right) \\ &= \sup_{\mu^i \in \mathcal{A}^i(y_{t+1}^i, Q^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}))} \mathbb{E}_{\mu^i} \left[-1/\beta + c^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}) + \beta w^i(x_{t+2}^{i,(y)}) \right] \\ &\leq \sup_{\mu^i \in \mathcal{A}^i(z_{t+1}^i, Q^i(z_{t+1}^i, 1, x_{t+2}^{i,(z)}))} \mathbb{E}_{\mu^i} \left[c^i(z_{t+1}^i, 1, x_{t+2}^{i,(z)}) + \beta w^i(x_{t+2}^{i,(z)}) \right] \\ &= \sigma \left(c^i(z_{t+1}^i, 1, x_{t+2}^{i,(z)}) + \beta w^i(x_{t+2}^{i,(z)}), z_{t+1}^i, Q^i(z_{t+1}^i, 1, x_{t+2}^{i,(z)}) \right) \\ &= w^i(z_{t+1}^i), \end{aligned}$$

where the second equality holds due to (A3). To show the inequality holds, we need to show for any probability measures μ and μ' , we have

$$\mathbb{E}_\mu \left[-1/\beta + c^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}) + \beta w^i(x_{t+2}^{i,(y)}) \right] \leq \mathbb{E}_{\mu'} \left[c^i(z_{t+1}^i, 1, x_{t+2}^{i,(z)}) + \beta w^i(x_{t+2}^{i,(z)}) \right].$$

For that, we show that the inequality $-1/\beta + c^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}) + \beta w^i(x_{t+2}^{i,(y)}) \leq c^i(z_{t+1}^i, 1, x_{t+2}^{i,(z)}) + \beta w^i(x_{t+2}^{i,(z)})$ holds for any realization of the process, so it will also hold for their convex combinations. We show this by considering the four following scenarios for the transition at $t+2$:

- (1) Consider $x_{t+2}^{i,(y)} = (s_1^i + s_t^i + 2, f_1^i + f_t^i)$ and $x_{t+2}^{i,(z)} = (s_1^i + s_t^i, f_1^i + f_t^i + 2)$, that is, we have success in the transition to $t+2$ starting from y_{t+1}^i and we have failure in the transition to $t+2$ starting from z_{t+1}^i , thus

$$\begin{aligned} -1/\beta + c^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}) + \beta w^i(x_{t+2}^{i,(y)}) &= -1/\beta - 1 + \beta w^i(x_{t+2}^{i,(y)}) \\ &= -1/\beta - 1 + \beta w^i(s_1^i + s_t^i + 2, f_1^i + f_t^i) \\ &\leq -1/\beta + \beta w^i(s_1^i + s_t^i + 1, f_1^i + f_t^i + 1) \\ &\leq -1/\beta + 1 + \beta w^i(s_1^i + s_t^i, f_1^i + f_t^i + 2) \\ &= -1/\beta + 1 + \beta w^i(x_{t+2}^{i,(z)}) \\ &\leq \beta w^i(x_{t+2}^{i,(z)}) \\ &= c^i(z_{t+1}^i, 1, x_{t+2}^{i,(z)}) + \beta w^i(x_{t+2}^{i,(z)}), \end{aligned}$$

where the first and second inequalities follow from the induction hypothesis and the third inequality is due to the fact that $0 \leq \beta \leq 1$.

- (2) Consider $x_{t+2}^{i,(y)} = (s_1^i + s_t^i + 2, f_1^i + f_t^i)$ and $x_{t+2}^{i,(z)} = (s_1^i + s_t^i + 1, f_1^i + f_t^i + 1)$, that is, we have success in the transition to $t+2$ both starting from y_{t+1}^i and z_{t+1}^i , thus

$$\begin{aligned} -1/\beta + c^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}) + \beta w^i(x_{t+2}^{i,(y)}) &= -1/\beta - 1 + \beta w^i(x_{t+2}^{i,(y)}) \\ &= -1/\beta - 1 + \beta w^i(s_1^i + s_t^i + 2, f_1^i + f_t^i) \\ &\leq -1/\beta + \beta w^i(s_1^i + s_t^i + 1, f_1^i + f_t^i + 1) \\ &= -1/\beta + \beta w^i(x_{t+2}^{i,(z)}) \\ &\leq -1 + \beta w^i(x_{t+2}^{i,(z)}) \\ &= c^i(z_{t+1}^i, 1, x_{t+2}^{i,(z)}) + \beta w^i(x_{t+2}^{i,(z)}), \end{aligned}$$

where the first inequality follows from the induction hypothesis and the second inequality is due to $0 \leq \beta \leq 1$.

- (3) Consider $x_{t+2}^{i,(y)} = x_{t+2}^{i,(z)} = (s_1^i + s_t^i + 1, f_1^i + f_t^i + 1)$, that is, we have failure in the transition to $t + 2$ starting from y_{t+1}^i and we have success in the transition to $t + 2$ starting from z_{t+1}^i , thus

$$\begin{aligned} -1/\beta + c^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}) + \beta w^i(x_{t+2}^{i,(y)}) &= -1/\beta + \beta w^i(x_{t+2}^{i,(y)}) \\ &= -1/\beta + \beta w^i(x_{t+2}^{i,(z)}) \\ &\leq -1 + \beta w^i(x_{t+2}^{i,(z)}) \\ &= c^i(z_{t+1}^i, 1, x_{t+2}^{i,(z)}) + \beta w^i(x_{t+2}^{i,(z)}), \end{aligned}$$

where the inequality is due to $0 \leq \beta \leq 1$.

- (4) Consider $x_{t+2}^{i,(y)} = (s_1^i + s_t^i + 1, f_1^i + f_t^i + 1)$ and $x_{t+2}^{i,(z)} = (s_1^i + s_t^i, f_1^i + f_t^i + 2)$, that is, we have failure in the transition to $t + 2$ both starting from y_{t+1}^i and z_{t+1}^i , thus

$$\begin{aligned} -1/\beta + c^i(y_{t+1}^i, 1, x_{t+2}^{i,(y)}) + \beta w^i(x_{t+2}^{i,(y)}) &= -1/\beta + \beta w^i(x_{t+2}^{i,(y)}) \\ &= -1/\beta + \beta w^i(s_1^i + s_t^i + 1, f_1^i + f_t^i + 1) \\ &\leq -1/\beta + 1 + \beta w^i(s_1^i + s_t^i, f_1^i + f_t^i + 2) \\ &= -1/\beta + 1 + \beta w^i(x_{t+2}^{i,(z)}) \\ &\leq \beta w^i(x_{t+2}^{i,(z)}) \\ &= c^i(z_{t+1}^i, 1, x_{t+2}^{i,(z)}) + \beta w^i(x_{t+2}^{i,(z)}), \end{aligned}$$

where the first inequality follows from the induction hypothesis and the second inequality is due to $0 \leq \beta \leq 1$.

Thus, the assertion of the lemma follows. $\square$

Proposition 5.3.1. *The risk-adjusted discounted cost for first-order mean-semideviation risk measure achieved by the suboptimal policy of never restarting the process in state x_1^i is equal to*

$$w^i(x_t^i) = -\frac{1}{1-\beta} \left(\frac{s_1^i + s_t^i}{s_1^i + s_t^i + f_1^i + f_t^i} \right) \left(1 - \kappa \frac{f_1^i + f_t^i}{s_1^i + s_t^i + f_1^i + f_t^i} \right), \text{ for all } x_t^i \in \mathcal{X}^i. \quad (5.21)$$

Proof. With reference to [39], the risk-adjusted discounted cost achieved by the suboptimal policy of never restarting the process in state x_1^i can be found as:

$$w^i(x_t^i) = \sigma \left(c^i(x_t^i, 1, x_{t+1}^i) + \beta w^i(x_{t+1}^i), x_t^i, Q^i(x_t^i, 1, x_{t+1}^i) \right), \text{ for all } x_t^i \in \mathcal{X}^i.$$

With respect to the definition of conditional first-order mean-semideviation risk measure defined in (5.4), we denote $p_t^i := \frac{s_1^i + s_t^i}{s_1^i + s_t^i + f_1^i + f_t^i}$, $\hat{s}_t^i := s_1^i + s_t^i$, and $\hat{f}_t^i := f_1^i + f_t^i$. It is worth noting that with this notation, we have $x_t^i = (\hat{s}_t^i, \hat{f}_t^i)$. Thus

$$\begin{aligned} w^i(x_t^i) &= \mathbb{E} \left[c^i(x_t^i, 1, x_{t+1}^i) + \beta w^i(x_{t+1}^i) \mid x_t^i \right] \\ &\quad + \kappa \mathbb{E} \left[\left(c^i(x_t^i, 1, x_{t+1}^i) + \beta w^i(x_{t+1}^i) - \mathbb{E} \left[c^i(x_t^i, 1, x_{t+1}^i) + \beta w^i(x_{t+1}^i) \mid x_t^i \right] \right)_+ \mid x_t^i \right] \\ &= -p_t^i + \beta \left(p_t^i w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + (1 - p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right) + \kappa \left[p_t^i \left(-1 + \beta w^i(\hat{s}_t^i + 1, \hat{f}_t^i) \right. \right. \\ &\quad \left. \left. + p_t^i - \beta \left(p_t^i w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + (1 - p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right) \right)_+ \right. \\ &\quad \left. + (1 - p_t^i) \left(\beta w^i(\hat{s}_t^i, \hat{f}_t^i + 1) + p_t^i - \beta \left(p_t^i w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + (1 - p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right) \right)_+ \right] \\ &= -p_t^i + \beta \left(p_t^i w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + (1 - p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right) \\ &\quad + \kappa \left[p_t^i \left((1 - p_t^i) \left(-1 + \beta (w^i(\hat{s}_t^i + 1, \hat{f}_t^i) - w^i(\hat{s}_t^i, \hat{f}_t^i + 1)) \right) \right)_+ \right. \\ &\quad \left. + (1 - p_t^i) \left(-p_t^i \left(-1 + \beta (w^i(\hat{s}_t^i + 1, \hat{f}_t^i) - w^i(\hat{s}_t^i, \hat{f}_t^i + 1)) \right) \right)_+ \right] \\ &= -p_t^i + \beta \left(p_t^i w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + (1 - p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right) \\ &\quad - \kappa \left[p_t^i \left((1 - p_t^i) \left(-1 + \beta (w^i(\hat{s}_t^i + 1, \hat{f}_t^i) - w^i(\hat{s}_t^i, \hat{f}_t^i + 1)) \right) \right) \right. \\ &= -(p_t^i - \kappa p_t^i (1 - p_t^i)) + \beta ((p_t^i - \kappa p_t^i (1 - p_t^i)) w^i(\hat{s}_t^i + 1, \hat{f}_t^i) \\ &\quad \left. + (1 - p_t^i)(1 + \kappa p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1)), \right] \end{aligned}$$

where the fourth equality is based on Lemma 5.3.1. Due to the structure of the final result, it is clear that this risk-averse value function is equivalent to the risk-neutral value function where the probability of success is equal to $p_t^i(1 - \kappa(1 - p_t^i))$. Since the value function related to the risk-neutral discounted cost achieved by the suboptimal policy of never restarting the process in state x_1^i is equal to the probability of success multiplied by $-1/(1 - \beta)$ (see [60]), we can write $w^i(x_t^i) = -p_t^i(1 - \kappa(1 - p_t^i))/(1 - \beta)$. $\square$

Proposition 5.3.2. *The risk-adjusted discounted cost for mean-AVaR risk measure achieved by the suboptimal policy of never restarting the process in state x_1^i for all $x_t^i \in \mathcal{X}^i$ is equal to*

$$w^i(x_t^i) = -\frac{1}{1 - \beta} \max \left\{ \lambda \left(\frac{s_1^i + s_t^i}{s_1^i + s_t^i + f_1^i + f_t^i} \right), 1 - \lambda' \left(\frac{f_1^i + f_t^i}{s_1^i + s_t^i + f_1^i + f_t^i} \right) \right\}, \quad (5.22)$$

where $\lambda' = 1 + \frac{\alpha}{1-\alpha}(1 - \lambda)$.

Proof. With reference to [39], the risk-adjusted discounted cost achieved by the suboptimal policy of never restarting the process in state x_1^i can be found as:

$$w^i(x_t^i) = \sigma \left(c^i(x_t^i, 1, x_{t+1}^i) + \beta w^i(x_{t+1}^i), x_t^i, Q^i(x_t^i, 1, x_{t+1}^i) \right), \text{ for all } x_t^i \in \mathcal{X}^i.$$

We denote $p_t^i := \frac{s_1^i + s_t^i}{s_1^i + s_t^i + f_1^i + f_t^i}$, $q_t^i := \frac{f_1^i + f_t^i}{s_1^i + s_t^i + f_1^i + f_t^i}$, $\hat{s}_t^i := s_1^i + s_t^i$, and $\hat{f}_t^i := f_1^i + f_t^i$. Consequently, we have $x_t^i = (\hat{s}_t^i, \hat{f}_t^i)$. With respect to the definition of conditional mean-AVaR risk measure defined in (5.5), we can write

$$\begin{aligned} w^i(x_t^i) &= \lambda \mathbb{E} \left[c^i(x_t^i, 1, x_{t+1}^i) + \beta w^i(x_{t+1}^i) \mid x_t^i \right] \\ &\quad + (1 - \lambda) \min_{\eta \in \mathcal{Z}_t} \left\{ \eta + \frac{1}{1 - \alpha} \mathbb{E} \left[(c^i(x_t^i, 1, x_{t+1}^i) + \beta w^i(x_{t+1}^i) - \eta)_+ \mid x_t^i \right] \right\} \\ &= -\lambda p_t^i + \beta \left(\lambda p_t^i w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + \lambda(1 - p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right) \\ &\quad + (1 - \lambda) \min_{\eta \in \mathcal{Z}_t} \left\{ \eta + \frac{1}{1 - \alpha} \left[p_t^i (-1 + \beta w^i(\hat{s}_t^i + 1, \hat{f}_t^i) - \eta)_+ \right. \right. \\ &\quad \left. \left. + (1 - p_t^i) (\beta w^i(\hat{s}_t^i, \hat{f}_t^i + 1) - \eta)_+ \right] \right\}. \end{aligned}$$

According to the definition of *value at risk*, that is, $\text{VaR}_\alpha(Z) = \min_t \{t : P(Z \leq t) \geq \alpha\}$, and considering Lemma 5.3.1 which states $-1 + \beta w^i(\hat{s}_t^i + 1, \hat{f}_t^i) \leq \beta w^i(\hat{s}_t^i, \hat{f}_t^i + 1)$, now we can consider the following two cases to find $w^i(x_t^i)$ by assessing two possible ranges for α :

- (1) Consider $\alpha \leq p_t^i$, we have $\eta^* = -1 + \beta w^i(\hat{s}_t^i + 1, \hat{f}_t^i)$, so $\text{AVaR}_\alpha(c^i(x_t^i, 1, x_{t+1}^i) + \beta w^i(x_{t+1}^i) \mid x_t^i) = -\frac{p_t^i - \alpha}{1 - \alpha} + \beta \left(\frac{p_t^i - \alpha}{1 - \alpha} w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + (1 - \frac{p_t^i - \alpha}{1 - \alpha}) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right)$

$\frac{p_t^i - \alpha}{1 - \alpha} w^i(\hat{s}_t^i, \hat{f}_t^i + 1)$). Then, we have

$$\begin{aligned}
w^i(x_t^i) &= -\lambda p_t^i + \beta \left(\lambda p_t^i w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + \lambda(1 - p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right) \\
&\quad + (1 - \lambda) \min_{\eta} \left\{ \eta + \frac{1}{1 - \alpha} \left[p_t^i (-1 + \beta w^i(\hat{s}_t^i + 1, \hat{f}_t^i) - \eta) \right. \right. \\
&\quad \left. \left. + (1 - p_t^i) (\beta w^i(\hat{s}_t^i, \hat{f}_t^i + 1) - \eta) \right] \right\} \\
&= -\lambda p_t^i + \beta \left(\lambda p_t^i w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + \lambda(1 - p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right) \\
&\quad + (1 - \lambda) \left(\frac{-p_t^i}{1 - \alpha} + \beta \left(\frac{p_t^i}{1 - \alpha} w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + \left(1 - \frac{p_t^i}{1 - \alpha} \right) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right) \right. \\
&\quad \left. + \min_{\eta} \left\{ \frac{-\eta \alpha}{1 - \alpha} \right\} \right) \\
&= -1 + \lambda' q_t^i + \beta \left((1 - \lambda' q_t^i) w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + \lambda' q_t^i w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right),
\end{aligned}$$

where $\lambda' = 1 + \frac{\alpha}{1 - \alpha} (1 - \lambda)$.

(2) Consider $\alpha > p_t^i$, we will have $\eta^* = \beta w^i(\hat{s}_t^i, \hat{f}_t^i + 1)$, so $\text{AVaR}_{\alpha}(c^i(x_t^i, 1, x_{t+1}^i) + \beta w^i(x_{t+1}^i) \mid x_t^i) = \beta w^i(\hat{s}_t^i, \hat{f}_t^i + 1)$. Thus

$$\begin{aligned}
w^i(x_t^i) &= -\lambda p_t^i + \beta \left(\lambda p_t^i w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + \lambda(1 - p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right) \\
&\quad + (1 - \lambda) \beta w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \\
&= -\lambda p_t^i + \beta \left(\lambda p_t^i w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + (1 - \lambda p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right).
\end{aligned}$$

Based on the above results, we have

$$w^i(x_t^i) = \begin{cases} -1 + \lambda' q_t^i + \beta \left((1 - \lambda' q_t^i) w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + \lambda' q_t^i w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right), & \text{if } \alpha \leq p_t^i \\ -\lambda p_t^i + \beta \left(\lambda p_t^i w^i(\hat{s}_t^i + 1, \hat{f}_t^i) + (1 - \lambda p_t^i) w^i(\hat{s}_t^i, \hat{f}_t^i + 1) \right), & \text{if } \alpha > p_t^i. \end{cases}$$

The final result in each part can be interpreted as the risk-neutral value function where the probability of success is equal to $1 - \lambda' q_t^i$ and λp_t^i , respectively. Based on the results for risk-neutral value function [60], we have

$$w^i(x_t^i) = \begin{cases} -\frac{1 - \lambda' q_t^i}{1 - \beta}, & \text{if } \alpha \leq p_t^i \\ -\frac{\lambda p_t^i}{1 - \beta}, & \text{if } \alpha > p_t^i. \end{cases}$$

or equivalently

$$w^i(x_t^i) = -\frac{1}{1 - \beta} \max \{ \lambda p_t^i, 1 - \lambda' q_t^i \},$$

which completes the proof. $\square$

Note that the transformations $\mathfrak{D}$, $\mathfrak{D}_1$, and $\mathfrak{D}_2$ are monotone contractions, and successive approximations will converge to their unique fixed points for any initial points of $v^{i,(1)}$, $u_T^{i,(1)}$, and $U_T^{i,(1)}$ (see Lemma 1 and Lemma 2 of [39]). Moreover, as shown in Theorem 5 of [39], if $v^{i,(1)}$ is set as the lowerbound for the optimal value function, then the sequence $v^{i,(k)}$, $k \in \{1, 2, \dots\}$ is nondecreasing, while for the initial point set as the upperbound, the sequence is nonincreasing. At this point, based on the theory and presentation of [60] and to be able to use (5.17) and (5.18) in solving Eq. (5.14), we introduce the following propositions.

Proposition 5.3.3. *We have*

$$-\frac{1}{1-\beta} \leq v^i(x_t^i) \leq w^i(x_t^i), \text{ for all } x_t^i \in \mathcal{X}^i \quad (5.23)$$

and

$$u_T^i(x_t^i) \leq v^i(x_t^i) \leq U_T^i(x_t^i), \text{ for all } x_t^i \in \mathcal{X}_T^i. \quad (5.24)$$

Proof. The first inequality in (5.23) holds as the left hand side is the cost one can pay by always winning. The second inequality in (5.23) holds as the left hand side is the risk-adjusted discounted cost achieved by the suboptimal policy of never restarting the process in state x_1^i when the initial state is x_t^i . Moreover, inequalities (5.24) follow from (5.23), the monotonicity of these transformations, as well as the fact that successive approximations will converge to their unique fixed points for any initial points of $v^{i,(1)}$, $u_T^{i,(1)}$, and $U_T^{i,(1)}$. $\square$

Proposition 5.3.4. *For any $\epsilon > 0$, there exist a $T_0 = T(\epsilon)$ such that*

$$U_T^i(x_1^i) - u_T^i(x_1^i) < \epsilon, \text{ for all } T \geq T_0. \quad (5.25)$$

Proof. Due to Eq. (5.24), it would be enough to show that for any positive constants ϵ_1 and ϵ_2 there exist $T_1 = T(\epsilon_1)$ and $T_2 = T(\epsilon_2)$ such that

$$v^i(x_1^i) - u_T^i(x_1^i) < \epsilon_1, \text{ for all } T \geq T_1 \quad (5.26)$$

and

$$U_T^i(x_1^i) - v^i(x_1^i) < \epsilon_2, \text{ for all } T \geq T_2. \quad (5.27)$$

In case we take $v^{i,(1)} = u_T^{i,(1)} = -\frac{1}{1-\beta}$, for all T and $n < T$, we have $v^{i,(n)}(x_1^i) = u_T^{i,(n)}(x_1^i)$. With respect to Lemma 2 of [39], we know that $\mathfrak{D}$ is a contraction mapping with modulus β on the L_∞ space. Given these and the fact that $v^i(x_1^i) \leq 0$, we have

$$\begin{aligned} v^i(x_1^i) - v^{i,(n)}(x_1^i) &\leq \beta \|v^i - v^{i,(n-1)}\|_\infty \leq \dots \\ &\leq \beta^n \|v^i - v^{i,(1)}\|_\infty \leq \frac{\beta^n}{1-\beta}. \end{aligned}$$

Therefore, it follows that

$$v^i(x_1^i) - u_T^{i,(n)}(x_1^i) \leq \frac{\beta^n}{1-\beta}, \text{ for all } n < T. \quad (5.28)$$

Similar arguments imply that

$$u_T^i(x_1^i) - u_T^{i,(n)}(x_1^i) \leq \frac{\beta^n}{1-\beta}, \text{ for all } n < T. \quad (5.29)$$

So, based on (5.28) and (5.29), Eq. (5.26) can be achieved. Using similar approach, we can show that Eq. (5.27) also holds, which completes the proof. $\square$

Now, we can provide the approach we will use for computing the risk-averse allocation indices for each state. With the initial state $x_1^i = (s_1^i, f_1^i)$, we compute $M^i(x_1^i) = v^i(x_1^i)$ with $\mathfrak{D}_1$ and $\mathfrak{D}_2$ transformations with the initial value functions of $u_T^{i,(1)}(x_t^i) = -1/(1-\beta)$ and $U_T^{i,(1)}(x_t^i) = w^i(x_t^i)$. We also take T large enough so that with respect to Proposition 5.3.4 and by iterating until reaching to $U_T^{i,(n)}(x_1^i) - u_T^{i,(n)}(x_1^i) < \epsilon$. Then, we find an approximation of $v^i(x_1^i)$ as the midpoint of the final interval.

5.4 Numerical Study

In this section, we first provide the allocation index values and their interpretation for a risk-averse clinical trial setting with risk measures of first-order mean-semideviation and mean-AVaR (Section 5.4.1) and then conduct a simulation study based on the extended results provided in the previous part (Section 5.4.2).

5.4.1 Allocation Index Values

In order to elaborate more on the index policy for solving a BB-MAB in a clinical trial setting, we consider three cases: risk-neutral; risk-averse with first-order mean-semideviation risk measure, given in (5.4), with $\kappa \in \{0.25, 0.50, 0.75, 1\}$; and risk-averse with mean-AVaR risk measure, given in (5.5), with parameters of $\alpha \in \{0.80, 0.90, 0.95\}$ and $\lambda \in \{0, 0.50\}$. Figures 5.2, 5.3, and 5.4 illustrate the (approximate) index values for the mentioned cases respectively, where the discount factor is considered to be 0.90. Moreover, Tables A.1-A.6 report the index values for each risk measure with discount factors of 0.50, 0.75, and 0.90. In these tables, the index related a treatment with prior $x_1^i = (s_1^i, f_1^i)$ and state (s_t^i, f_t^i) is the entry in cell $x_t^i = (s_1^i + s_t^i, f_1^i + f_t^i)$. To compute these indices, we consider the index precision of 3 digits and the truncation to $T = 12, 28$, and 75 for discount factor values of 0.50, 0.75, and 0.90, respectively. The reason we choose these truncation values is that 0.50^{12} , 0.75^{28} , and 0.90^{75} are all less than 10^{-3} , so the treated patients after these periods will result in a negligible risk-adjusted discounted outcome and are therefore ignored.

Similar to [18], our risk-averse index policy assigns a number to each treatment (retrieved from an extended version of Tables A.1-A.6) based on the observed values of $x_t^i = (s_1^i + s_t^i, f_1^i + f_t^i)$, and then dictate assigning the treatment with the highest value. As a result, if we account for each treatment prior, the same table will be used to make the decisions in the trial. For instance, considering the discount factor of 0.90 and first-order mean-semideviation risk measure with $\kappa = 0.50$, in case we have two available treatments that start with a common uniform prior, the initial state of both will be $x_1^i = (1, 1)$ with a common index value of 5.426. However, if the beta prior parameters of the first treatment and the second treatment is $(1, 3)$ and $(2, 1)$, respectively, their initial states will be $x_1^1 = (1, 3)$ and $x_1^2 = (2, 1)$, and their corresponding index values will be 2.138 and 6.871. The same argument holds if priors combine with the data would result in $x_1^1 = (1, 3)$ and $x_1^2 = (2, 1)$. Further, the exploration-exploitation trade-off which exists in the multi-arm bandits can also be clearly seen in these tables. Among the two treatments with $x_t^1 = (2, 2)$ and $x_t^2 = (4, 4)$ which have the

same posterior success mean of 0.5, the treatment with the fewest number of observations will be chosen with respect to our risk-averse index policy: $4.705 > 4.254$. This behavior is observed with any extent of risk-aversion and highlights the exploration dimension of MAB. Moreover, consider the two treatments with $x_t^1 = (1, 1)$ and $x_t^2 = (50, 30)$ where the posterior means of success indicate that the second treatment is superior to the first one, but their indices suggest the reverse, $5.426 > 5.141$, indicating that the least observed population is preferred. That being said, as the degree of risk aversion increases (i.e., as κ rises), the second treatment will be preferred in terms of our index policy, as $3.637 < 3.946$ with $\kappa = 1$. This indicates that as the decision maker's risk-aversion level increases, the decision maker will be more inclined to exploit rather than explore. This can also be visually verified in Figures 5.3 and 5.4.

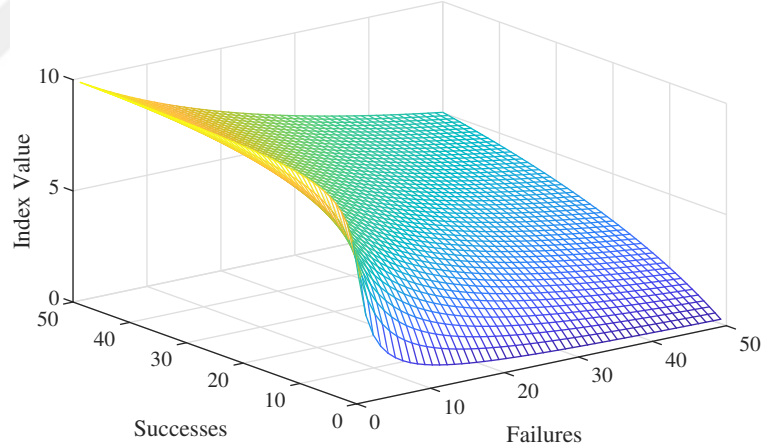


Figure 5.2: The (approximate) Gittins index values for $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures.

5.4.2 Simulation Study

In this section, we assess the performance of several patient allocation rules in a clinical trial setting, including the risk-averse index-based solution of Section 5.3. We focus on the expectation and standard deviation of number of patient successes. In particular, we consider the following patient allocation methods:

- a fixed randomized design which is created by allocating patients to each

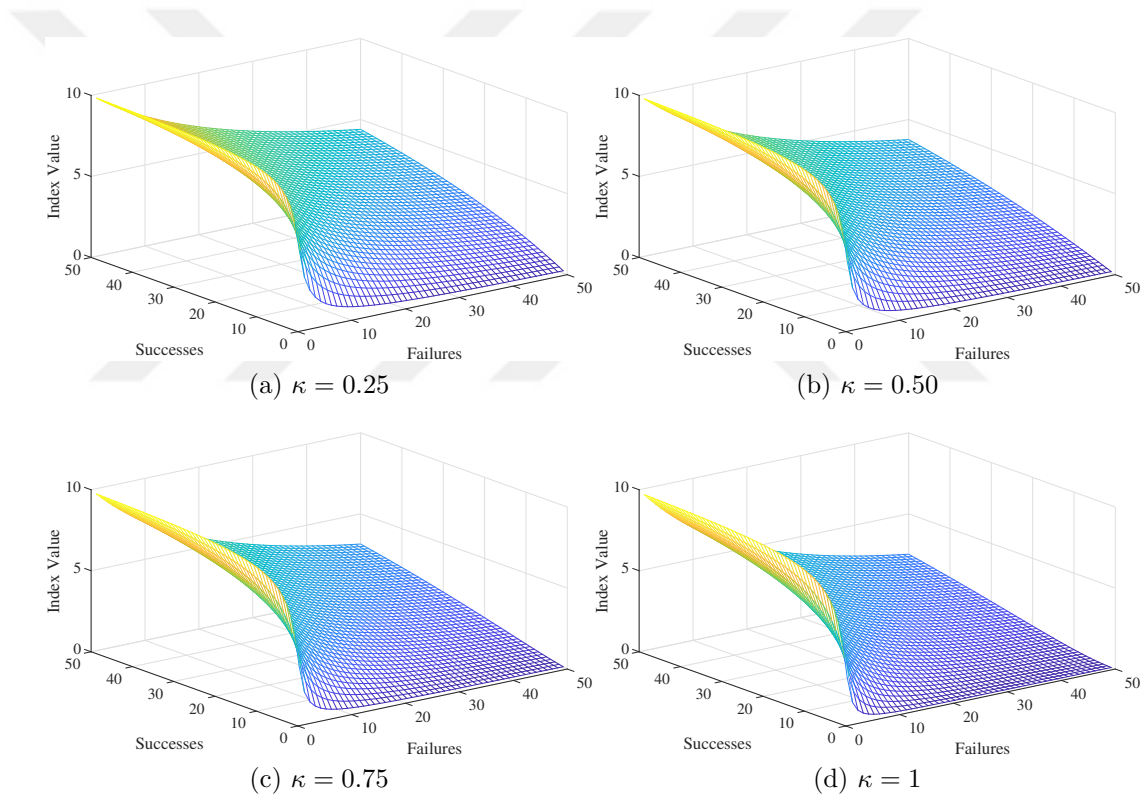


Figure 5.3: The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for first-order mean-semideviation risk measure with different parameters of risk aversion.

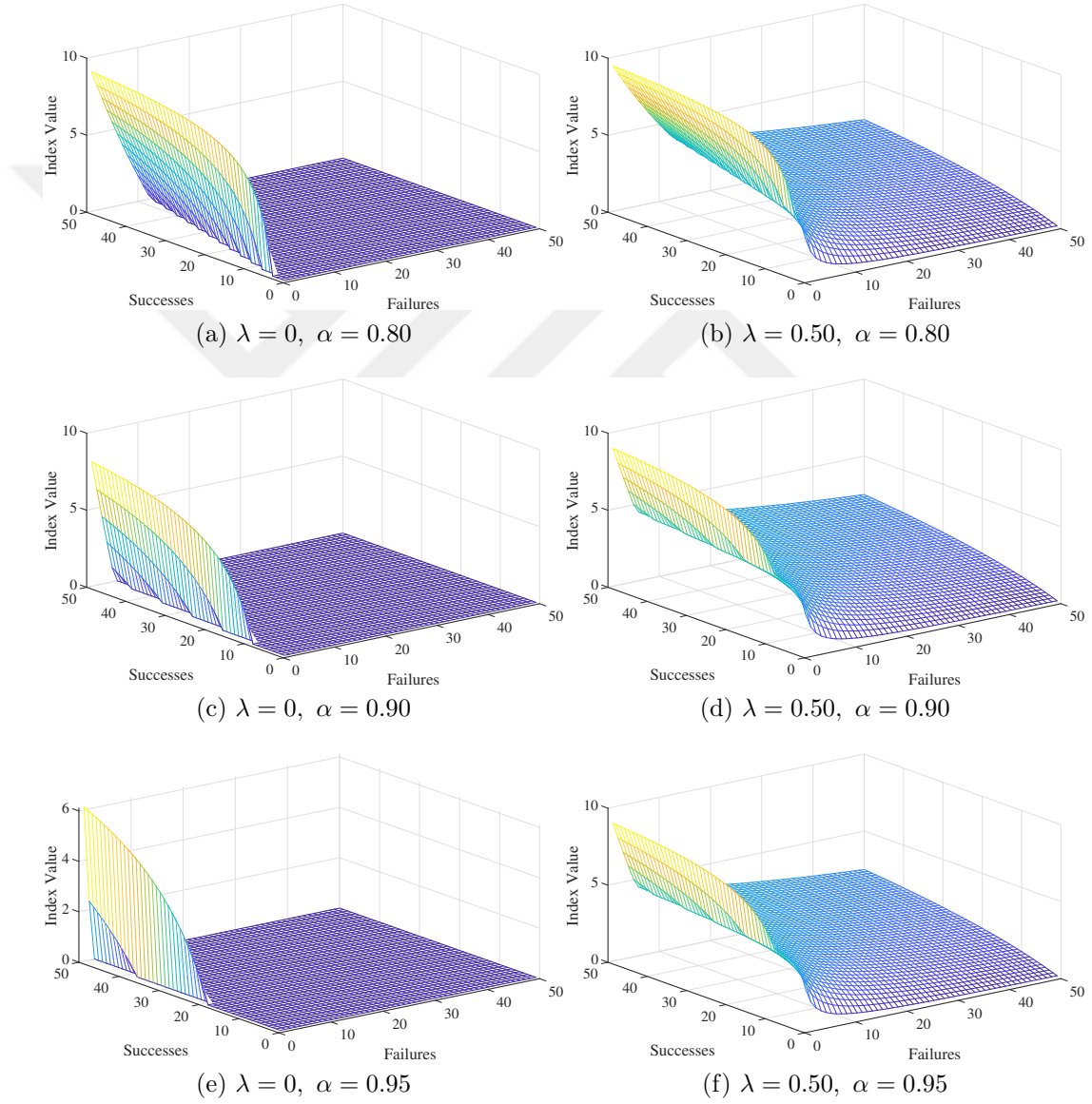


Figure 5.4: The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for mean-AVaR risk measure with different parameters of risk aversion.

arm with an equal probability throughout the trial,

- the risk-neutral policy obtained via Gittins indices,
- the risk-averse index policy described in Definition 5.2.1.

We utilize simulation to examine the performance of different allocation methods. Following [55], we assume that the decision maker can choose one of $K = 3$ arms at each step. We will look at the situation where the decision maker doesn't know which arm has the highest mean before the data is gathered. To do so, we use equal (unit) values for the beta prior's shape parameters. A discount factor of 0.90 is adopted in all of the experiments. After determining these values, we employ Monte Carlo simulation to calculate the performance of the randomized design as well as the risk-averse and risk-neutral index policies. We repeat this experiment 1,000 times. The results are shown in Table 5.1. The performance of the randomized design is lower than that of the index-based approaches. The two index-based approaches, on the other hand, exhibit the best performance. It is worth noting that, even though the performance of the two index policies are comparable, the risk-averse policy reduces the standard deviation of rewards. For instance, for mean-AVaR risk measure with $\lambda = 0$ and $\alpha = 0.95$, the standard deviation of rewards for risk-neutral case is 54.227 while it is 53.019 for the risk-averse policy. This indicates that the risk-averse policy provides approximately 2.25% improvement over the risk-neutral policy in the variation of the results. Moreover, for first-order mean-semideviation risk measure with $\kappa = 1$, we observe both expected performance increase and performance variability reduction, which corresponds to more than 1.7% improvement in the coefficient of variation.

5.5 Conclusion

The vast majority of multi-armed bandit literature in the clinical trial context focuses on finding the treatment with the highest total success in expectation. In this chapter, we adopt a novel multi-armed bandit setting where the objective

Table 5.1: Performance comparison of different designs for BB-MAB.

Policy			Average performance	Sample standard deviation	Coefficient of variation
Fixed Randomization			123.583	41.503	0.3358
Gittins Index Policy			173.345	54.227	0.3128
Risk-averse Index Policy (First-order Mean-semideviation)	κ				
	0.25		173.497	53.639	0.3092
	0.50		173.142	53.646	0.3098
	0.75		173.375	53.549	0.3089
	1		173.359	53.283	0.3074
Risk-averse Index Policy (Mean-AVaR)	λ	α			
	0	0.80	171.735	53.019	0.3087
		0.90	171.735	53.019	0.3087
		0.95	171.735	53.019	0.3087
	0.50	0.80	173.279	53.461	0.3085
		0.90	173.182	53.520	0.3090
		0.95	173.182	53.520	0.3090

is to find the best risk–return trade–off introduced in Chapter 4 and propose a framework to solve the risk-averse Bayesian Bernoulli multi–armed bandits. In particular, we rely on two well-known dynamic risk measures to measure the performance of the designs, namely first-order mean-semideviation and mean-AVaR. Based on our numerical results, the proposed risk-averse index policy provides improvements in terms of the performance variability with minimal degradation in the expected performance over the standard Gittins policy. The reduced variance indicates that the risk-averse policy has the ability to choose the treatment with the lowest risk.

Chapter 6

Conclusion

In this dissertation, we consider risk-averse multi-armed bandit problems with dynamic risk measures, which falls into the broad category of risk-averse dynamic programming. In Chapter 3, we introduce a new approach in incorporating risk into multi-armed bandit problem by using dynamic coherent measures of risk and propose a priority-index heuristic, analogous to the structure of Gittins index. The proposed heuristic is based on Lagrangian duality and the indices are expressed as the Lagrangian multiplier corresponding to the activation constraint. We also characterize optimal policies for the risk-averse one-armed bandit problem and show that the proposed heuristic provides an optimal policy for this problem. We computationally test the performance of our index policy using two famous coherent risk measures of *first-order mean-semideviation* and *mean-AVaR*. These experiments indicate the excellence of our index policy in all evaluated criteria. Based on our results, the proposed index policy provides much lower suboptimality percentages and it is more similar to the optimal policy compared to the policy derived from Gittins index. More specifically, regarding *first-order mean-semideviation* and *mean-AVaR* risk measures, respectively, our index policy is optimal in 90.65% and 97.67% of cases and in no occasion it is more than 5% and 7.6% suboptimal, whereas Gittins index policy is optimal only in 75.6% and 58.4% of the cases with more than 10% and 40% suboptimality percentage, respectively. Our experiments comparing the computational complexity

of our index heuristic to solving the corresponding MDP reveals that our method is time efficient and the computation time grows linearly with the problem size, whereas the risk-averse MDP computation time shows an exponential growth.

In Chapter 4, we present a new priority-index heuristic for the risk-incorporated multi-arm bandit problem with dynamic risk measures introduced in Chapter 3. In respect of this specific setting, we present a theoretical analysis based on Whittle’s retirement problem and show that the proposed index policy reduces to Gittins index when the level of risk-aversion converges to zero. We also generalize the restart-in-state formulation of the Gittins index to effectively compute these risk-averse allocation indices. In our numerical experiments, we again use the risk measures of *first order mean-semideviation* and *mean-AVaR* to evaluate the performance of our index policy. The experiments reveal that our index-based policy has much lower percentages of suboptimality and is closer to the optimal policy relative to the Gittins index policy. More precisely, with regard to the risk measures of *first-order mean-semideviation* and *mean-AVaR*, respectively, our index policy achieves optimality in 95% and 91% of cases with no more than 0.13% and 1.2% optimality gap, while the policy derived from Gittins index is optimal only in 62% and 41.5% of the cases with some instances more than 10% and 48.5% suboptimal, respectively.

In Chapter 5, we adopt the multi-armed bandit setting with the objective of finding the best risk–return trade–off introduced in Chapter 4 and propose a framework to solve the risk-averse Bayesian Bernoulli multi–armed bandits. We consider an application of risk-averse Bayesian Bernoulli MAB to assign patients to treatments in a clinical trial. In our analysis, we rely on the risk measures of *first-order mean-semideviation* and *mean-AVaR*. When compared to the standard fixed randomization methodology, our findings reveal that index-based allocation methods perform exceedingly well. Based on our numerical results, the proposed risk-averse index policy provides improvements in terms of the performance variability with minimal degradation in the expected performance. The reduced variance indicates that the risk-averse policy has the ability to choose the treatment with the lowest risk.

The risk-averse MAB introduced in Chapters 3 and 4 can be extended with respect to the many variations that exist in the literature for classical MAB. For example, resources may be allocated to more than one arm at a time, new arms may appear, the state of all arms may change, or there may be restrictions linking the arms. Analyzing the results of these variations in a risk-averse setting is an opportunity for future work.

The risk-averse allocation indices derived in Chapters 3 and 4 can be applied to some other important practical applications of multi-armed bandit problem. There are many other applications in which risk incorporation is of great importance, including but not limited to (1) *internet advertising* where we need to figure out which ads to show to specific individuals, (2) *sequential portfolio selection* where we need to find the best combination of assets in the portfolio to achieve the investor’s objectives, as well as (3) *deep learning* where we can employ a bandit optimization process to select fine-tuned parameters. With this, by applying the theory derived in Chapters 3 and 4, one can unveil the importance of risk-averse optimization in the domain of multi-armed bandits by evaluating the quality of the results derived by the proposed algorithms compared to the risk-neutral benchmarks.

In Chapter 5, the results are based on the fact that the outcomes of a treatment can be either success or failure. However, the optimality results of Gittins has also been investigated for other possibilities, such as multinomial distribution [61], normally distributed processes [62], and exponentially distributed populations [63]. Therefore, extending the analysis and conclusions of this work with this respect can be considered as a future research direction.

Bibliography

- [1] H. Robbins, “Some aspects of the sequential design of experiments,” in *Herbert Robbins Selected Papers*, pp. 169–177, Springer, 1985.
- [2] R. Bellman, “A problem in the sequential design of experiments,” *Sankhyā: The Indian Journal of Statistics (1933-1960)*, vol. 16, no. 3/4, pp. 221–229, 1956.
- [3] D. A. Berry and B. Fristedt, *Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)*. Springer, 1985.
- [4] M. K. Warmuth and D. Kuzmin, “Online variance minimization,” in *International Conference on Computational Learning Theory*, pp. 514–528, Springer, 2006.
- [5] E. Even-Dar, M. Kearns, and J. Wortman, “Risk-sensitive online learning,” in *International Conference on Algorithmic Learning Theory*, pp. 199–213, Springer, 2006.
- [6] W. F. Sharpe, “Mutual fund performance,” *The Journal of business*, vol. 39, no. 1, pp. 119–138, 1966.
- [7] H. Markowitz, “Portfolio selection,” *The journal of finance*, vol. 7, no. 1, pp. 77–91, 1952.
- [8] A. Sani, A. Lazaric, and R. Munos, “Risk-aversion in multi-armed bandits,” in *Advances in Neural Information Processing Systems*, pp. 3275–3283, 2012.
- [9] A. Sani, A. Lazaric, and R. Munos, “Risk-aversion in multi-arm bandit,” Tech. Rep. hal-00750298, 2012.

- [10] O.-A. Maillard, “Robust risk-averse stochastic multi-armed bandits,” in *International Conference on Algorithmic Learning Theory*, pp. 218–233, Springer, 2013.
- [11] J. Y. Yu and E. Nikolova, “Sample complexity of risk-averse bandit-arm selection.,” in *IJCAI*, 2013.
- [12] N. Galichet, M. Sebag, and O. Teytaud, “Exploration vs exploitation vs safety: Risk-aware multi-armed bandits.,” in *ACML*, pp. 245–260, 2013.
- [13] N. Galichet, *Contributions to Multi-Armed Bandits : Risk-Awareness and Sub-Sampling for Linear Contextual Bandits*. PhD thesis, Université Paris Sud - Paris XI, Sept. 2015.
- [14] S. Vakili and Q. Zhao, “Risk-averse online learning under mean-variance measures,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1911–1915, IEEE, 2015.
- [15] S. Vakili and Q. Zhao, “Mean-variance and value at risk in multi-armed bandit problems,” in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1330–1335, IEEE, 2015.
- [16] A. Zimin, R. Ibsen-Jensen, and K. Chatterjee, “Generalized risk-aversion in stochastic multi-armed bandits,” *arXiv preprint arXiv:1405.0833*, 2014.
- [17] S. Vakili and Q. Zhao, “Risk-averse multi-armed bandit problems under mean-variance measure,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, pp. 1093–1111, Sept 2016.
- [18] J. C. Gittins, “Bandit processes and dynamic allocation indices,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 148–177, 1979.
- [19] P. Whittle, “Multi-armed bandits and the gittins index,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 143–149, 1980.
- [20] P. Whittle, “Arm-acquiring bandits,” *The Annals of Probability*, pp. 284–292, 1981.

- [21] P. Nash, “A generalized bandit problem,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 165–169, 1980.
- [22] P. Varaiya, J. Walrand, and C. Buyukkoc, “Extensions of the multiarmed bandit problem: The discounted case,” *IEEE transactions on automatic control*, vol. 30, no. 5, pp. 426–439, 1985.
- [23] M. N. Katehakis and A. F. Veinott Jr, “The multi-armed bandit problem: decomposition and computation,” *Mathematics of Operations Research*, vol. 12, no. 2, pp. 262–268, 1987.
- [24] R. Weber, “On the gittins index for multiarmed bandits,” *The Annals of Applied Probability*, pp. 1024–1033, 1992.
- [25] G. Weiss, “Branching bandit processes,” *Probability in the Engineering and Informational Sciences*, vol. 2, no. 3, pp. 269–278, 1988.
- [26] N. El Karoui and I. Karatzas, “Dynamic allocation problems in continuous time,” *The Annals of Applied Probability*, pp. 255–286, 1994.
- [27] J. N. Tsitsiklis *et al.*, “A short proof of the gittins index theorem,” *The Annals of Applied Probability*, vol. 4, no. 1, pp. 194–199, 1994.
- [28] H. Kaspi and A. Mandelbaum, “Multi-armed bandits in discrete and continuous time,” *Annals of Applied Probability*, pp. 1270–1290, 1998.
- [29] L. C. Kallenberg, “A note on mn katehakis’ and y.-r. chen’s computation of the gittins index,” *Mathematics of operations research*, vol. 11, no. 1, pp. 184–186, 1986.
- [30] I. M. Sonin, “A generalized gittins index for a markov chain and its recursive calculation,” *Statistics & Probability Letters*, vol. 78, no. 12, pp. 1526–1533, 2008.
- [31] P. Whittle, “Restless bandits: Activity allocation in a changing world,” *Journal of applied probability*, vol. 25, no. A, pp. 287–298, 1988.

- [32] P. Varaiya, J. Walrand, and C. Buyukkoc, “Extension of the multi-armed bandit problem,” in *Decision and Control, 1983. The 22nd IEEE Conference on*, vol. 22, pp. 1179–1180, IEEE, 1983.
- [33] J.-P. Chancelier, M. De Lara, and A. De Palma, “Risk aversion, road choice, and the one-armed bandit problem,” *Transportation Science*, vol. 41, no. 1, pp. 1–14, 2007.
- [34] J.-P. Chancelier, M. De Lara, and A. de Palma, “Risk aversion in expected intertemporal discounted utilities bandit problems,” *Theory and decision*, vol. 67, no. 4, pp. 433–440, 2009.
- [35] E. V. Denardo, H. Park, and U. G. Rothblum, “Risk-sensitive and risk-neutral multiarmed bandits,” *Mathematics of Operations Research*, vol. 32, no. 2, pp. 374–394, 2007.
- [36] E. V. Denardo, E. A. Feinberg, and U. G. Rothblum, “The multi-armed bandit, with constraints,” *Annals of Operations Research*, vol. 208, no. 1, pp. 37–62, 2013.
- [37] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.
- [38] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, “Coherent measures of risk,” *Mathematical finance*, vol. 9, no. 3, pp. 203–228, 1999.
- [39] A. Ruszczyński, “Risk-averse dynamic programming for markov decision processes,” *Mathematical programming*, vol. 125, no. 2, pp. 235–261, 2010.
- [40] M. Malekipirbazari and Ö. Çavuş, “Risk-averse allocation indices for multi-armed bandit problem,” *IEEE Transactions on Automatic Control*, pp. 1–1, 2021.
- [41] F. Riedel, “Dynamic coherent risk measures,” *Stochastic processes and their applications*, vol. 112, no. 2, pp. 185–200, 2004.
- [42] A. Ruszczyński and A. Shapiro, “Conditional risk mappings,” *Mathematics of operations research*, vol. 31, no. 3, pp. 544–561, 2006.

- [43] W. Ogryczak and A. Ruszczyński, “From stochastic dominance to mean-risk models: Semideviations as risk measures,” *European journal of operational research*, vol. 116, no. 1, pp. 33–50, 1999.
- [44] R. T. Rockafellar and S. Uryasev, “Conditional value-at-risk for general loss distributions,” *Journal of banking & finance*, vol. 26, no. 7, pp. 1443–1471, 2002.
- [45] K. D. Glazebrook and R. Minty, “A generalized gittins index for a class of multiarmed bandits with general resource requirements,” *Mathematics of Operations Research*, vol. 34, no. 1, pp. 26–44, 2009.
- [46] M. Woodroffe, “A one-armed bandit problem with a concomitant variable,” *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 799–806, 1979.
- [47] Ö. Çavuş and A. Ruszczyński, “Computational methods for risk-averse undiscounted transient markov models,” *Operations Research*, vol. 62, no. 2, pp. 401–417, 2014.
- [48] V. Kuleshov and D. Precup, “Algorithms for multi-armed bandit problems,” *arXiv preprint arXiv:1402.6028*, 2014.
- [49] P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, and H. Ku, “Coherent multiperiod risk adjusted values and bellman’s principle,” *Annals of Operations Research*, vol. 152, no. 1, pp. 5–22, 2007.
- [50] P. Cheridito, F. Delbaen, and M. Kupper, “Dynamic monetary risk measures for bounded discrete-time processes,” *Electronic Journal of Probability*, vol. 11, pp. 57–106, 2006.
- [51] H. Föllmer and I. Penner, “Convex risk measures and the dynamics of their penalty functions,” *Statistics & Decisions*, vol. 24, no. 1/2006, pp. 61–96, 2006.
- [52] G. Calafiore and F. Dabbene, *Probabilistic and randomized methods for design under uncertainty*. Springer, 2006.

- [53] Ö. Çavuş and A. Ruszczyński, “Risk-averse control of undiscounted transient markov models,” *SIAM Journal on Control and Optimization*, vol. 52, no. 6, pp. 3935–3966, 2014.
- [54] A. Ruszczyński and A. Shapiro, “Optimization of convex risk functions,” *Mathematics of operations research*, vol. 31, no. 3, pp. 433–452, 2006.
- [55] M. J. Kim and A. E. Lim, “Robust multiarmed bandit problems,” *Management Science*, vol. 62, no. 1, pp. 264–285, 2016.
- [56] F. Caro and A. D. Gupta, “Robust control of the multi-armed bandit problem,” *Annals of Operations Research*, pp. 1–20, 2015.
- [57] J. Nino-Mora, “Restless bandits, partial conservation laws and indexability,” *Advances in Applied Probability*, vol. 33, no. 1, pp. 76–98, 2001.
- [58] D. Robinson, “Algorithms for evaluating the dynamic allocation index,” *Operations Research Letters*, vol. 1, no. 2, pp. 72–74, 1982.
- [59] R. A. Collado, D. Papp, and A. Ruszczyński, “Scenario decomposition of risk-averse multistage stochastic programming problems,” *Annals of Operations Research*, vol. 200, no. 1, pp. 147–170, 2012.
- [60] M. N. Katehakis and C. Derman, “Computing optimal sequential allocation rules in clinical trials,” *Lecture notes-monograph series*, pp. 29–39, 1986.
- [61] K. Glazebrook, “On the optimal allocation of two or more treatments in a controlled clinical trial,” *Biometrika*, vol. 65, no. 2, pp. 335–340, 1978.
- [62] D. Jones, *Search Procedures for Industrial Chemical Research*. PhD thesis, University of Cambridge, 1975.
- [63] J. Ferreira Pena Do Amaral, *Aspects of optimal sequential resource allocation*. PhD thesis, University of Oxford, 1985.

Appendix A

Allocation index values for risk
measures of first-order
mean-semideviation and
mean-AVaR

Table A.1: The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for first-order mean-semideviation risk measure where $\beta = 0.50$.

κ	$s_1^i + s_t^i$	$f_1^i + f_t^i$									
		1	2	3	4	5	10	20	30	40	50
0.25	1	1.118	0.751	0.560	0.444	0.367	0.194	0.099	0.066	0.049	0.039
	2	1.411	1.071	0.859	0.715	0.611	0.351	0.188	0.128	0.097	0.078
	3	1.554	1.257	1.051	0.902	0.789	0.482	0.269	0.186	0.142	0.115
	4	1.639	1.379	1.187	1.040	0.925	0.592	0.342	0.240	0.185	0.150
	5	1.697	1.466	1.288	1.147	1.032	0.688	0.410	0.291	0.226	0.184
	10	1.829	1.683	1.558	1.449	1.354	1.016	0.677	0.507	0.405	0.337
	20	1.908	1.824	1.747	1.675	1.610	1.344	1.008	0.807	0.672	0.576
	30	1.937	1.878	1.822	1.769	1.720	1.507	1.207	1.006	0.862	0.754
	40	1.952	1.906	1.863	1.821	1.781	1.605	1.338	1.148	1.004	0.892
	50	1.961	1.924	1.888	1.854	1.820	1.670	1.433	1.254	1.115	1.002
0.50	1	0.845	0.495	0.342	0.259	0.208	0.103	0.051	0.033	0.025	0.020
	2	1.189	0.804	0.598	0.471	0.388	0.201	0.101	0.067	0.050	0.040
	3	1.372	1.010	0.788	0.641	0.539	0.293	0.150	0.100	0.075	0.060
	4	1.486	1.155	0.934	0.779	0.666	0.378	0.198	0.133	0.100	0.080
	5	1.564	1.263	1.049	0.892	0.773	0.456	0.244	0.165	0.124	0.100
	10	1.750	1.548	1.382	1.245	1.130	0.762	0.450	0.315	0.242	0.195
	20	1.864	1.743	1.635	1.538	1.451	1.121	0.756	0.564	0.447	0.369
	30	1.906	1.820	1.741	1.667	1.598	1.320	0.965	0.754	0.615	0.518
	40	1.928	1.862	1.799	1.739	1.683	1.445	1.116	0.901	0.753	0.644
	50	1.942	1.887	1.835	1.786	1.739	1.532	1.228	1.019	0.867	0.752
0.75	1	0.707	0.369	0.237	0.170	0.132	0.059	0.027	0.017	0.013	0.010
	2	1.075	0.670	0.468	0.352	0.278	0.128	0.058	0.037	0.027	0.021
	3	1.279	0.885	0.656	0.512	0.415	0.200	0.092	0.058	0.042	0.033
	4	1.408	1.042	0.807	0.648	0.537	0.271	0.126	0.080	0.058	0.045
	5	1.497	1.160	0.929	0.765	0.644	0.341	0.162	0.103	0.074	0.058
	10	1.710	1.479	1.293	1.142	1.018	0.634	0.337	0.220	0.161	0.125
	20	1.842	1.703	1.579	1.469	1.371	1.009	0.629	0.443	0.335	0.266
	30	1.891	1.792	1.700	1.616	1.538	1.226	0.845	0.628	0.492	0.400
	40	1.917	1.839	1.767	1.698	1.634	1.365	1.004	0.778	0.627	0.520
	50	1.933	1.869	1.809	1.752	1.698	1.463	1.126	0.901	0.743	0.627
1	1	0.567	0.244	0.134	0.084	0.057	0.016	0.004	0.002	0.001	0.000
	2	0.960	0.536	0.339	0.233	0.170	0.056	0.016	0.007	0.004	0.002
	3	1.185	0.760	0.524	0.383	0.292	0.108	0.034	0.016	0.009	0.006
	4	1.329	0.928	0.680	0.518	0.408	0.166	0.056	0.027	0.016	0.011
	5	1.429	1.057	0.809	0.637	0.515	0.226	0.080	0.041	0.024	0.016
	10	1.671	1.411	1.205	1.039	0.905	0.507	0.224	0.125	0.080	0.055
	20	1.820	1.662	1.523	1.400	1.291	0.897	0.503	0.322	0.223	0.163
	30	1.876	1.763	1.659	1.564	1.477	1.132	0.724	0.502	0.369	0.282
	40	1.905	1.817	1.735	1.658	1.585	1.285	0.893	0.655	0.501	0.396
	50	1.923	1.851	1.783	1.718	1.657	1.393	1.024	0.784	0.619	0.501

Table A.2: The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for first-order mean-semideviation risk measure where $\beta = 0.75$.

κ	$s_1^i + s_t^i$	$f_1^i + f_t^i$									
		1	2	3	4	5	10	20	30	40	50
0.25	1	2.194	1.400	1.008	0.781	0.634	0.320	0.157	0.104	0.077	0.061
	2	2.766	2.015	1.564	1.272	1.068	0.584	0.301	0.202	0.151	0.121
	3	3.049	2.386	1.938	1.625	1.397	0.810	0.434	0.294	0.222	0.179
	4	3.221	2.635	2.211	1.896	1.657	1.007	0.557	0.383	0.291	0.235
	5	3.338	2.815	2.417	2.111	1.869	1.180	0.671	0.467	0.357	0.289
	10	3.614	3.277	2.989	2.742	2.530	1.812	1.146	0.835	0.656	0.539
	20	3.786	3.588	3.406	3.239	3.087	2.489	1.781	1.383	1.129	0.953
	30	3.851	3.710	3.577	3.453	3.335	2.845	2.187	1.771	1.487	1.280
	40	3.885	3.776	3.671	3.572	3.477	3.064	2.467	2.060	1.766	1.544
	50	3.907	3.817	3.731	3.648	3.568	3.213	2.672	2.283	1.990	1.762
0.50	1	1.894	1.103	0.755	0.567	0.452	0.219	0.105	0.069	0.051	0.041
	2	2.538	1.724	1.276	1.003	0.821	0.421	0.209	0.138	0.102	0.082
	3	2.868	2.125	1.655	1.345	1.127	0.608	0.308	0.205	0.153	0.122
	4	3.071	2.403	1.945	1.619	1.381	0.779	0.405	0.271	0.203	0.162
	5	3.209	2.607	2.169	1.844	1.596	0.936	0.498	0.336	0.253	0.202
	10	3.539	3.143	2.811	2.534	2.301	1.549	0.913	0.639	0.489	0.395
	20	3.744	3.509	3.296	3.102	2.927	2.262	1.524	1.137	0.901	0.743
	30	3.821	3.654	3.497	3.351	3.215	2.656	1.943	1.516	1.237	1.041
	40	3.862	3.732	3.608	3.491	3.380	2.904	2.242	1.812	1.512	1.294
	50	3.888	3.781	3.679	3.581	3.487	3.075	2.467	2.046	1.740	1.510
0.75	1	1.586	0.813	0.514	0.366	0.281	0.123	0.056	0.036	0.026	0.021
	2	2.303	1.432	0.993	0.741	0.583	0.265	0.119	0.076	0.055	0.043
	3	2.682	1.862	1.373	1.068	0.863	0.412	0.187	0.118	0.085	0.067
	4	2.917	2.168	1.677	1.343	1.109	0.557	0.257	0.162	0.117	0.091
	5	3.079	2.397	1.919	1.577	1.325	0.696	0.329	0.208	0.150	0.117
	10	3.463	3.007	2.633	2.325	2.071	1.287	0.681	0.445	0.324	0.253
	20	3.701	3.429	3.184	2.965	2.767	2.036	1.268	0.891	0.674	0.535
	30	3.792	3.597	3.417	3.249	3.093	2.467	1.699	1.262	0.988	0.803
	40	3.840	3.688	3.545	3.410	3.282	2.744	2.018	1.564	1.259	1.044
	50	3.870	3.746	3.627	3.514	3.406	2.936	2.261	1.809	1.491	1.257
1	1	1.273	0.532	0.286	0.177	0.120	0.034	0.009	0.004	0.002	0.001
	2	2.062	1.141	0.715	0.487	0.353	0.115	0.033	0.015	0.009	0.005
	3	2.492	1.597	1.093	0.795	0.603	0.221	0.069	0.033	0.019	0.012
	4	2.761	1.931	1.410	1.070	0.840	0.339	0.113	0.056	0.033	0.022
	5	2.945	2.184	1.668	1.312	1.055	0.461	0.163	0.082	0.049	0.033
	10	3.386	2.870	2.452	2.115	1.841	1.027	0.452	0.253	0.161	0.112
	20	3.659	3.350	3.073	2.826	2.607	1.809	1.013	0.647	0.448	0.329
	30	3.762	3.541	3.336	3.146	2.972	2.278	1.456	1.009	0.740	0.566
	40	3.817	3.645	3.482	3.329	3.185	2.583	1.793	1.316	1.006	0.795
	50	3.851	3.710	3.575	3.446	3.324	2.797	2.056	1.573	1.242	1.005

Table A.3: The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for first-order mean-semideviation risk measure where $\beta = 0.90$.

κ	$s_1^i + s_t^i$	$f_1^i + f_t^i$									
		1	2	3	4	5	10	20	30	40	50
0.25	1	6.257	4.071	2.933	2.268	1.835	0.910	0.436	0.282	0.207	0.163
	2	7.456	5.534	4.304	3.505	2.937	1.592	0.810	0.538	0.400	0.318
	3	8.031	6.379	5.210	4.364	3.756	2.165	1.149	0.775	0.583	0.466
	4	8.377	6.942	5.852	5.027	4.385	2.659	1.461	1.000	0.758	0.609
	5	8.612	7.343	6.333	5.542	4.910	3.094	1.750	1.213	0.925	0.747
	10	9.170	8.367	7.654	7.034	6.497	4.657	2.941	2.138	1.676	1.377
	20	9.524	9.054	8.611	8.200	7.820	6.315	4.519	3.508	2.860	2.413
	30	9.662	9.327	9.005	8.699	8.409	7.182	5.524	4.472	3.753	3.230
	40	9.736	9.476	9.222	8.978	8.744	7.716	6.216	5.191	4.448	3.890
	50	9.783	9.569	9.360	9.157	8.961	8.078	6.723	5.743	5.007	4.433
0.50	1	5.426	3.156	2.138	1.592	1.259	0.598	0.283	0.183	0.135	0.107
	2	6.871	4.705	3.464	2.712	2.209	1.119	0.548	0.360	0.267	0.211
	3	7.582	5.663	4.411	3.568	2.985	1.593	0.803	0.531	0.396	0.315
	4	8.012	6.323	5.118	4.254	3.618	2.027	1.047	0.699	0.523	0.417
	5	8.303	6.799	5.659	4.809	4.157	2.423	1.282	0.862	0.647	0.517
	10	8.993	8.029	7.195	6.489	5.891	3.956	2.323	1.623	1.241	1.002
	20	9.426	8.861	8.334	7.852	7.412	5.728	3.853	2.871	2.273	1.875
	30	9.593	9.191	8.807	8.444	8.104	6.699	4.898	3.819	3.114	2.620
	40	9.683	9.370	9.067	8.777	8.501	7.310	5.643	4.557	3.801	3.252
	50	9.740	9.483	9.233	8.991	8.759	7.729	6.200	5.141	4.372	3.791
0.75	1	4.547	2.281	1.416	0.996	0.756	0.326	0.146	0.093	0.068	0.053
	2	6.248	3.874	2.655	1.968	1.538	0.690	0.308	0.194	0.141	0.110
	3	7.108	4.940	3.625	2.800	2.253	1.063	0.479	0.302	0.218	0.170
	4	7.628	5.690	4.387	3.498	2.876	1.431	0.657	0.413	0.298	0.232
	5	7.980	6.240	4.984	4.087	3.422	1.784	0.838	0.529	0.381	0.297
	10	8.811	7.684	6.730	5.941	5.287	3.271	1.724	1.124	0.819	0.638
	20	9.326	8.665	8.054	7.501	7.001	5.144	3.197	2.244	1.695	1.347
	30	9.523	9.053	8.606	8.187	7.797	6.216	4.276	3.172	2.483	2.017
	40	9.629	9.263	8.911	8.575	8.256	6.903	5.072	3.928	3.160	2.620
	50	9.696	9.396	9.105	8.824	8.555	7.378	5.679	4.542	3.741	3.153
1	1	3.637	1.462	0.766	0.469	0.316	0.088	0.023	0.010	0.006	0.003
	2	5.598	3.056	1.885	1.273	0.916	0.295	0.085	0.039	0.023	0.014
	3	6.611	4.214	2.859	2.063	1.557	0.565	0.175	0.084	0.049	0.032
	4	7.229	5.050	3.665	2.764	2.161	0.864	0.286	0.141	0.083	0.055
	5	7.645	5.674	4.314	3.378	2.708	1.171	0.412	0.208	0.125	0.083
	10	8.624	7.332	6.262	5.393	4.686	2.600	1.140	0.637	0.406	0.281
	20	9.224	8.466	7.771	7.147	6.590	4.564	2.549	1.626	1.125	0.825
	30	9.453	8.914	8.404	7.929	7.488	5.735	3.659	2.532	1.858	1.420
	40	9.575	9.156	8.753	8.371	8.010	6.496	4.504	3.303	2.524	1.993
	50	9.652	9.308	8.975	8.656	8.350	7.027	5.160	3.946	3.114	2.519

Table A.4: The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for mean-AVaR risk measure where $\beta = 0.50$.

λ	α	$s_1^i + s_t^i$	$f_1^i + f_t^i$									
			1	2	3	4	5	10	20	30	40	50
0	0.80	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		5	0.357	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		10	1.119	0.345	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		20	1.536	1.105	0.707	0.339	0.000	0.000	0.000	0.000	0.000	0.000
		30	1.684	1.385	1.101	0.832	0.578	0.000	0.000	0.000	0.000	0.000
		40	1.760	1.530	1.310	1.098	0.896	0.000	0.000	0.000	0.000	0.000
		50	1.807	1.620	1.439	1.265	1.097	0.336	0.000	0.000	0.000	0.000
	0.90	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		10	0.189	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		20	1.063	0.186	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		30	1.365	0.759	0.184	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		40	1.519	1.055	0.610	0.184	0.000	0.000	0.000	0.000	0.000	0.000
		50	1.613	1.237	0.874	0.522	0.183	0.000	0.000	0.000	0.000	0.000
	0.95	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		20	0.097	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		30	0.719	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		40	1.033	0.097	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		50	1.222	0.466	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.50	0.80	1	0.525	0.352	0.263	0.210	0.174	0.093	0.048	0.032	0.024	0.019
		2	0.684	0.515	0.413	0.344	0.294	0.171	0.092	0.063	0.048	0.038
		3	0.769	0.611	0.510	0.438	0.383	0.235	0.132	0.092	0.070	0.057
		4	0.853	0.676	0.580	0.508	0.452	0.290	0.168	0.118	0.091	0.074
		5	1.048	0.722	0.632	0.563	0.506	0.338	0.202	0.144	0.112	0.091
		10	1.479	1.026	0.773	0.718	0.670	0.503	0.335	0.251	0.201	0.167
		20	1.723	1.467	1.231	1.013	0.811	0.668	0.501	0.401	0.334	0.286
		30	1.811	1.632	1.463	1.303	1.152	0.751	0.601	0.501	0.429	0.376
		40	1.856	1.719	1.587	1.461	1.340	0.806	0.667	0.572	0.500	0.445
		50	1.884	1.772	1.664	1.560	1.460	1.005	0.715	0.625	0.556	0.500
	0.90	1	0.524	0.352	0.263	0.210	0.174	0.093	0.048	0.032	0.024	0.019
		2	0.682	0.514	0.413	0.344	0.294	0.171	0.092	0.063	0.048	0.038
		3	0.761	0.611	0.510	0.438	0.383	0.235	0.132	0.092	0.070	0.057
		4	0.807	0.676	0.580	0.508	0.452	0.290	0.168	0.118	0.091	0.074
		5	0.839	0.721	0.632	0.562	0.506	0.338	0.202	0.144	0.112	0.091
		10	1.028	0.836	0.773	0.718	0.670	0.503	0.335	0.251	0.201	0.167
		20	1.489	1.014	0.871	0.835	0.801	0.668	0.501	0.401	0.334	0.286
		30	1.652	1.322	1.009	0.883	0.858	0.751	0.601	0.501	0.429	0.375
		40	1.736	1.483	1.240	1.007	0.889	0.800	0.667	0.572	0.500	0.445
		50	1.787	1.582	1.383	1.191	1.005	0.834	0.715	0.625	0.556	0.500
	0.95	1	0.524	0.352	0.263	0.210	0.174	0.093	0.048	0.032	0.024	0.019
		2	0.682	0.515	0.413	0.344	0.294	0.171	0.092	0.063	0.048	0.038
		3	0.761	0.611	0.510	0.438	0.383	0.235	0.132	0.092	0.070	0.057
		4	0.808	0.676	0.580	0.508	0.452	0.290	0.168	0.118	0.091	0.074
		5	0.839	0.722	0.632	0.563	0.506	0.338	0.202	0.144	0.112	0.091
		10	0.911	0.836	0.773	0.718	0.670	0.503	0.335	0.251	0.201	0.167
		20	1.015	0.910	0.871	0.835	0.802	0.668	0.501	0.401	0.334	0.286
		30	1.333	0.938	0.910	0.883	0.858	0.751	0.601	0.501	0.429	0.376
		40	1.495	1.008	0.931	0.909	0.889	0.801	0.667	0.572	0.500	0.445
		50	1.593	1.199	0.944	0.926	0.909	0.834	0.715	0.625	0.556	0.500

Table A.5: The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for mean-AVaR risk measure where $\beta = 0.75$.

λ	α	$s_1^i + s_t^i$	$f_1^i + f_t^i$									
			1	2	3	4	5	10	20	30	40	50
0	0.80	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		5	0.749	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		10	2.290	0.707	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		20	3.103	2.237	1.431	0.686	0.000	0.000	0.000	0.000	0.000	0.000
		30	3.387	2.790	2.219	1.677	1.165	0.000	0.000	0.000	0.000	0.000
		40	3.533	3.076	2.634	2.210	1.802	0.000	0.000	0.000	0.000	0.000
		50	3.622	3.252	2.892	2.542	2.204	0.674	0.000	0.000	0.000	0.000
	0.90	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		10	0.388	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		20	2.152	0.375	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		30	2.751	1.530	0.371	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		40	3.054	2.124	1.228	0.369	0.000	0.000	0.000	0.000	0.000	0.000
		50	3.238	2.486	1.756	1.050	0.368	0.000	0.000	0.000	0.000	0.000
	0.95	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		20	0.197	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		30	1.449	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		40	2.078	0.193	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		50	2.456	0.935	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.50	0.80	1	1.089	0.737	0.554	0.442	0.367	0.196	0.100	0.067	0.050	0.040
		2	1.401	1.052	0.847	0.707	0.607	0.352	0.189	0.129	0.097	0.078
		3	1.589	1.240	1.037	0.892	0.782	0.481	0.270	0.187	0.143	0.115
		4	1.798	1.365	1.174	1.029	0.917	0.590	0.343	0.241	0.185	0.151
		5	2.186	1.456	1.276	1.137	1.024	0.685	0.410	0.292	0.226	0.185
		10	3.013	2.097	1.556	1.442	1.347	1.012	0.676	0.507	0.405	0.337
		20	3.471	2.964	2.490	2.049	1.641	1.340	1.006	0.805	0.671	0.575
		30	3.636	3.284	2.947	2.626	2.321	1.504	1.205	1.004	0.861	0.753
		40	3.722	3.451	3.189	2.937	2.695	1.620	1.337	1.146	1.003	0.892
		50	3.774	3.554	3.340	3.133	2.932	2.019	1.431	1.253	1.114	1.002
	0.90	1	1.086	0.737	0.554	0.442	0.367	0.196	0.100	0.067	0.050	0.040
		2	1.387	1.052	0.847	0.707	0.607	0.352	0.189	0.129	0.097	0.078
		3	1.536	1.240	1.037	0.892	0.782	0.481	0.270	0.187	0.143	0.115
		4	1.626	1.365	1.174	1.029	0.917	0.590	0.343	0.241	0.185	0.151
		5	1.687	1.454	1.276	1.137	1.024	0.685	0.410	0.292	0.226	0.185
		10	2.105	1.677	1.551	1.442	1.347	1.012	0.676	0.507	0.405	0.337
		20	3.009	2.053	1.744	1.672	1.606	1.340	1.006	0.805	0.671	0.575
		30	3.324	2.665	2.036	1.767	1.717	1.504	1.205	1.004	0.861	0.753
		40	3.485	2.982	2.495	2.027	1.780	1.603	1.337	1.146	1.003	0.892
		50	3.584	3.176	2.779	2.394	2.021	1.669	1.431	1.253	1.114	1.002
	0.95	1	1.086	0.737	0.554	0.442	0.367	0.196	0.100	0.067	0.050	0.040
		2	1.387	1.052	0.847	0.707	0.607	0.352	0.189	0.129	0.097	0.078
		3	1.536	1.240	1.037	0.892	0.782	0.481	0.270	0.187	0.143	0.115
		4	1.626	1.365	1.174	1.029	0.917	0.590	0.343	0.241	0.185	0.151
		5	1.686	1.454	1.276	1.137	1.024	0.685	0.410	0.292	0.226	0.185
		10	1.825	1.677	1.551	1.442	1.347	1.012	0.676	0.507	0.405	0.337
		20	2.056	1.822	1.744	1.672	1.606	1.340	1.006	0.805	0.671	0.575
		30	2.686	1.877	1.820	1.767	1.717	1.504	1.205	1.004	0.861	0.753
		40	3.005	2.028	1.862	1.820	1.780	1.603	1.337	1.146	1.003	0.892
		50	3.199	2.409	1.887	1.853	1.819	1.669	1.431	1.253	1.114	1.002

Table A.6: The (approximate) allocation index values of $s_1^i + s_t^i$ successes and $f_1^i + f_t^i$ failures for mean-AVaR risk measure where $\beta = 0.90$.

λ	α	$s_1^i + s_t^i$	$f_1^i + f_t^i$									
			1	2	3	4	5	10	20	30	40	50
0	0.80	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		5	1.942	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		10	5.856	1.796	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		20	7.852	5.655	3.610	1.729	0.000	0.000	0.000	0.000	0.000	0.000
		30	8.534	7.031	5.588	4.221	2.929	0.000	0.000	0.000	0.000	0.000
		40	8.883	7.739	6.625	5.554	4.527	0.000	0.000	0.000	0.000	0.000
		50	9.095	8.172	7.265	6.385	5.534	1.691	0.000	0.000	0.000	0.000
	0.90	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		10	0.988	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		20	5.442	0.947	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		30	6.934	3.851	0.934	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		40	7.684	5.339	3.084	0.927	0.000	0.000	0.000	0.000	0.000	0.000
		50	8.136	6.245	4.408	2.634	0.924	0.000	0.000	0.000	0.000	0.000
	0.95	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		20	0.497	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		30	3.647	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		40	5.224	0.486	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		50	6.171	2.346	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.50	1	2.814	1.923	1.459	1.172	0.976	0.524	0.266	0.176	0.131	0.104
		2	3.596	2.681	2.170	1.819	1.564	0.914	0.492	0.334	0.253	0.202
		3	4.124	3.137	2.631	2.269	1.993	1.233	0.694	0.480	0.366	0.296
		4	4.742	3.442	2.965	2.602	2.322	1.503	0.876	0.616	0.474	0.385
		5	5.708	3.669	3.214	2.866	2.584	1.735	1.043	0.743	0.576	0.470
		10	7.704	5.355	3.916	3.617	3.380	2.544	1.702	1.278	1.022	0.851
		20	8.770	7.500	6.296	5.176	4.140	3.358	2.523	2.021	1.685	1.445
		30	9.149	8.275	7.427	6.616	5.846	3.765	3.017	2.515	2.158	1.888
		40	9.345	8.678	8.023	7.389	6.778	4.069	3.346	2.869	2.511	2.233
		50	9.467	8.926	8.393	7.872	7.366	5.069	3.581	3.135	2.787	2.509
	0.90	1	2.792	1.923	1.459	1.172	0.976	0.524	0.266	0.176	0.131	0.104
		2	3.510	2.680	2.170	1.819	1.564	0.914	0.492	0.334	0.253	0.202
		3	3.870	3.136	2.631	2.269	1.993	1.233	0.694	0.480	0.366	0.296
		4	4.089	3.438	2.965	2.602	2.322	1.503	0.876	0.616	0.474	0.385
		5	4.241	3.655	3.214	2.866	2.584	1.735	1.043	0.743	0.576	0.470
		10	5.379	4.203	3.889	3.617	3.380	2.544	1.702	1.278	1.022	0.851
		20	7.616	5.188	4.364	4.186	4.021	3.358	2.523	2.021	1.685	1.445
		30	8.378	6.716	5.125	4.422	4.297	3.765	3.017	2.515	2.158	1.888
		40	8.765	7.501	6.273	5.093	4.453	4.011	3.346	2.869	2.511	2.233
		50	9.000	7.981	6.981	6.012	5.074	4.174	3.581	3.135	2.787	2.509
	0.95	1	2.792	1.923	1.459	1.172	0.976	0.524	0.266	0.176	0.131	0.104
		2	3.510	2.680	2.170	1.819	1.564	0.914	0.492	0.334	0.253	0.202
		3	3.868	3.136	2.631	2.269	1.993	1.233	0.694	0.480	0.366	0.296
		4	4.085	3.438	2.965	2.602	2.322	1.503	0.876	0.616	0.474	0.385
		5	4.231	3.655	3.214	2.866	2.584	1.735	1.043	0.743	0.576	0.470
		10	4.569	4.203	3.889	3.617	3.380	2.544	1.702	1.278	1.022	0.851
		20	5.197	4.558	4.364	4.186	4.021	3.358	2.523	2.021	1.685	1.445
		30	6.771	4.694	4.554	4.422	4.297	3.765	3.017	2.515	2.158	1.888
		40	7.562	5.098	4.656	4.552	4.452	4.011	3.346	2.869	2.511	2.233
		50	8.039	6.051	4.720	4.634	4.550	4.174	3.581	3.135	2.787	2.509