Risk-Averse Multiarmed Bandit Problem with Switching Penalties

Milad Malekipirbazari

Data Science and AI Division,
Department of Computer Science and Engineering,
Chalmers University of Technology

November 20, 2023

Markovian Bandits with Switching Penalties

- Each arm is an independent Markov machine
- Playing the chosen arm $\rightarrow$ payoff & transition
- Switching between arms $\rightarrow$ penalties
- **Goal**: maximizing the total expected reward obtained over an infinite horizon while minimizing the total switching penalty

Solution Methods

- **One solution method**: Model as a Markov decision process (MDP) and apply Markov decision theory to find an optimal policy → `computationally challenging`
- **Gittins and Jones (1974)**: Provide an optimal solution for the problem by proposing decomposition to each arm and an index policy approach
- **Banks and Sundaram (1994)**: In the presence of switching costs, it is not possible to define an index for each arm such that the resulting strategy always produces the maximized payoffs.
- **Asawa and Teneketzis (1996)**: Introduce an index-based heuristic.
- There is no research on risk-averse bandits with switching costs.

Solution Methods

- **One solution method**: Model as a Markov decision process (MDP) and apply Markov decision theory to find an optimal policy → computationally challenging
- **Gittins and Jones (1974)**: Provide an optimal solution for the problem by proposing decomposition to each arm and an index policy approach
- **Banks and Sundaram (1994)**: In the presence of switching costs, it is not possible to define an index for each arm such that the resulting strategy always produces the maximized payoffs.
- **Asawa and Teneketzis (1996)**: Introduce an index-based heuristic.
- There is no research on risk-averse bandits with switching costs.

Solution Methods

- **One solution method**: Model as a Markov decision process (MDP) and apply Markov decision theory to find an optimal policy → computationally challenging
- **Gittins and Jones (1974)**: Provide an optimal solution for the problem by proposing decomposition to each arm and an index policy approach
- **Banks and Sundaram (1994)**: In the presence of switching costs, it is not possible to define an index for each arm such that the resulting strategy always produces the maximized payoffs.
- **Asawa and Teneketzis (1996)**: Introduce an index-based heuristic.
- There is no research on risk-averse bandits with switching costs.

Solution Methods

- **One solution method**: Model as a Markov decision process (MDP) and apply Markov decision theory to find an optimal policy → `computationally challenging`
- **Gittins and Jones (1974)**: Provide an optimal solution for the problem by proposing decomposition to each arm and an index policy approach
- **Banks and Sundaram (1994)**: In the presence of switching costs, it is not possible to define an index for each arm such that the resulting strategy always produces the maximized payoffs.
- **Asawa and Teneketzis (1996)**: Introduce an index-based heuristic.
- There is no research on risk-averse bandits with switching costs.

Risk-Averse Optimization

- **Risk-neutral opt.**: highest expected total reward
- **Risk-averse opt.**: high expected total reward & low reward variability

Example 1 (Clinical Trials)

- Optimal risk-neutral treatment: significant variability in patients' satisfaction with the treatment (side effects)
- Optimal risk-averse treatment: maybe not the most effective on average but more reliable and consistent

Example 2 (Deploying Ambulances During Emergency)

- Optimal risk-neutral plan: minimizing the average of arrival times to possible calls
- Optimal risk-averse plan: also minimizing the variance of response times, resulting in a greater number of saved lives

Risk-Averse Optimization

- **Risk-neutral opt.**: highest expected total reward
- **Risk-averse opt.**: high expected total reward & low reward variability

Example 1 (Clinical Trials)

- Optimal risk-neutral treatment: significant variability in patients' satisfaction with the treatment (side effects)
- Optimal risk-averse treatment: maybe not the most effective on average but more reliable and consistent

Example 2 (Deploying Ambulances During Emergency)

- Optimal risk-neutral plan: minimizing the average of arrival times to possible calls
- Optimal risk-averse plan: also minimizing the variance of response times, resulting in a greater number of saved lives

Incorporating Risk into the Markovian Bandits

1. **Concave utility functions**
   - Denardo et al. (2007) & Denardo et al. (2013): exponential utility
   - **Drawback**:
     - ⋆ hard to state a suitable utility function with regard to the level of risk-aversion
     - ⋆ hard to interpret the resulting solutions
2. **Coherent risk measures**
   - Does not inherit the challenges of using utility functions
   - Malekipirbazari and Çavuş (2021) & Malekipirbazari and Çavuş (2024):
     - ⋆ first-order mean-semideviation
     - ⋆ mean-CVaR

Problem Description

- Converting the problem into minimization → considering the negative of rewards (interpreted as costs)
- Each arm i is a Markov chain with a finite state space $\mathcal{X}^i$, $i \in \mathcal{K}$ with $\mathcal{K} = \{1, 2, \ldots, K\}$
- At step $t \in \mathbb{N}$, an action $u_t = [u_t^1 u_t^2 \ldots u_t^K]^T$ is applied:
  - $u_t^i \in \{0, 1\} \rightarrow$ the action applied to arm $i \in \mathcal{K}$
  - $u_t^i = 1$: arm i is played at step t
  - $u_t^i = 0$: arm i is not played at step t
- $c^i(x^i, u^i, y^i)$: Cost incurred by arm i under action u^i by $x^i \rightarrow y^i$
- The switching cost $s = [s^1 s^2 \cdots s^K]^T$ is set to be finite and non-negative, where s^i represents the cost incurred by switching to arm $i \in \mathcal{K}$.
- A stationary (Markov) policy $\pi : \mathcal{X} \rightarrow \{0, 1\}^K$
  - $u_t = \pi(x_t)$: both denote the decision to be taken at state $x_t \in \mathcal{X}$

Problem Description

- Converting the problem into minimization $\rightarrow$ considering the negative of rewards (interpreted as costs)
- Each arm i is a Markov chain with a finite state space $\mathcal{X}^i$, $i \in \mathcal{K}$ with $\mathcal{K} = \{1, 2, \ldots, K\}$
- At step $t \in \mathbb{N}$, an action $u_t = [u_t^1 u_t^2 \ldots u_t^K]^T$ is applied:
  - $u_t^i \in \{0, 1\} \rightarrow$ the action applied to arm $i \in \mathcal{K}$
  - $u_t^i = 1$: arm i is played at step t
  - $u_t^i = 0$: arm i is not played at step t
- $c^i(x^i, u^i, y^i)$: Cost incurred by arm i under action u^i by $x^i \rightarrow y^i$
- The switching cost $s = [s^1 s^2 \cdots s^K]^T$ is set to be finite and non-negative, where s^i represents the cost incurred by switching to arm $i \in \mathcal{K}$.
- A stationary (Markov) policy $\pi : \mathcal{X} \rightarrow \{0, 1\}^K$
  - $u_t = \pi(x_t)$: both denote the decision to be taken at state $x_t \in \mathcal{X}$

Problem Description

- Converting the problem into minimization $\rightarrow$ considering the negative of rewards (interpreted as costs)
- Each arm i is a Markov chain with a finite state space $\mathcal{X}^i$, $i \in \mathcal{K}$ with $\mathcal{K} = \{1, 2, \ldots, K\}$
- At step $t \in \mathbb{N}$, an action $u_t = [u_t^1 u_t^2 \ldots u_t^K]^T$ is applied:
  - $u_t^i \in \{0, 1\} \rightarrow$ the action applied to arm $i \in \mathcal{K}$
  - $u_t^i = 1$: arm i is played at step t
  - $u_t^i = 0$: arm i is not played at step t
- $c^i(x^i, u^i, y^i)$: Cost incurred by arm i under action u^i by $x^i \rightarrow y^i$
- The switching cost $s = [s^1 s^2 \cdots s^K]^T$ is set to be finite and non-negative, where s^i represents the cost incurred by switching to arm $i \in \mathcal{K}$.
- A stationary (Markov) policy $\pi : \mathcal{X} \rightarrow \{0, 1\}^K$
  - $u_t = \pi(x_t)$: both denote the decision to be taken at state $x_t \in \mathcal{X}$

Problem Description

- Converting the problem into minimization $\rightarrow$ considering the negative of rewards (interpreted as costs)
- Each arm i is a Markov chain with a finite state space $\mathcal{X}^i$, $i \in \mathcal{K}$ with $\mathcal{K} = \{1, 2, \ldots, K\}$
- At step $t \in \mathbb{N}$, an action $u_t = [u_t^1 u_t^2 \ldots u_t^K]^T$ is applied:
  - $u_t^i \in \{0, 1\} \rightarrow$ the action applied to arm $i \in \mathcal{K}$
  - $u_t^i = 1$: arm i is played at step t
  - $u_t^i = 0$: arm i is not played at step t
- $c^i(x^i, u^i, y^i)$: Cost incurred by arm i under action u^i by $x^i \rightarrow y^i$
- The switching cost $s = [s^1 s^2 \cdots s^K]^T$ is set to be finite and non-negative, where s^i represents the cost incurred by switching to arm $i \in \mathcal{K}$.
- A stationary (Markov) policy $\pi : \mathcal{X} \rightarrow \{0, 1\}^K$
  - $u_t = \pi(x_t)$: both denote the decision to be taken at state $x_t \in \mathcal{X}$

Problem Description

- Converting the problem into minimization $\rightarrow$ considering the negative of rewards (interpreted as costs)
- Each arm i is a Markov chain with a finite state space $\mathcal{X}^i$, $i \in \mathcal{K}$ with $\mathcal{K} = \{1, 2, \ldots, K\}$
- At step $t \in \mathbb{N}$, an action $u_t = [u_t^1 u_t^2 \ldots u_t^K]^T$ is applied:
  - $u_t^i \in \{0, 1\} \rightarrow$ the action applied to arm $i \in \mathcal{K}$
  - $u_t^i = 1$: arm i is played at step t
  - $u_t^i = 0$: arm i is not played at step t
- $c^i(x^i, u^i, y^i)$: Cost incurred by arm i under action u^i by $x^i \rightarrow y^i$
- The switching cost $s = [s^1 s^2 \cdots s^K]^T$ is set to be finite and non-negative, where s^i represents the cost incurred by switching to arm $i \in \mathcal{K}$.
- A stationary (Markov) policy $\pi : \mathcal{X} \rightarrow \{0, 1\}^K$
  - $u_t = \pi(x_t)$: both denote the decision to be taken at state $x_t \in \mathcal{X}$

Problem Description

- Converting the problem into minimization $\rightarrow$ considering the negative of rewards (interpreted as costs)
- Each arm i is a Markov chain with a finite state space $\mathcal{X}^i$, $i \in \mathcal{K}$ with $\mathcal{K} = \{1, 2, \ldots, K\}$
- At step $t \in \mathbb{N}$, an action $u_t = [u_t^1 u_t^2 \ldots u_t^K]^T$ is applied:
  - $u_t^i \in \{0, 1\} \rightarrow$ the action applied to arm $i \in \mathcal{K}$
  - $u_t^i = 1$: arm i is played at step t
  - $u_t^i = 0$: arm i is not played at step t
- $c^i(x^i, u^i, y^i)$: Cost incurred by arm i under action u^i by $x^i \rightarrow y^i$
- The switching cost $s = [s^1 s^2 \cdots s^K]^T$ is set to be finite and non-negative, where s^i represents the cost incurred by switching to arm $i \in \mathcal{K}$.
- A stationary (Markov) policy $\pi : \mathcal{X} \rightarrow \{0, 1\}^K$
  - $u_t = \pi(x_t)$: both denote the decision to be taken at state $x_t \in \mathcal{X}$

One-Step Conditional Risk Measure

$\rho_t : \mathcal{Z}_{t+1} \to \mathcal{Z}_t,\ t \in \{1, \ldots, T\}$ satisfying the following axioms is defined as one-step conditional risk measure:

(A1) $\rho_t(\alpha Z + (1-\alpha)W) \leqslant \alpha\rho_t(Z) + (1-\alpha)\rho_t(W), \forall \alpha \in (0,1), Z, W \in \mathcal{Z}_{t+1}$

(A2) if $Z \preceq W$, then $\rho_t(Z) \leqslant \rho_t(W),\ \forall Z, W \in \mathcal{Z}_{t+1}$

(A3) $\rho_t(Z + W) = Z + \rho_t(W),\ \forall Z \in \mathcal{Z}_t,\ W \in \mathcal{Z}_{t+1}$

(A4) $\rho_t(\alpha Z) = \alpha\rho_t(Z),\ \forall Z \in \mathcal{Z}_{t+1}, \alpha \geqslant 0$

Two Important Conditional Risk Measures

1. *First-order mean-semideviation*:

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

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

2. *Mean-average value-at-risk* (mean-AVaR):

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

where

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

and $\alpha \in (0,1)$ and $\lambda \in [0,1]$

Two Important Conditional Risk Measures

1. *First-order mean-semideviation*:

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

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

2. *Mean-average value-at-risk* (mean-AVaR):

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

where

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

and $\alpha \in (0,1)$ and $\lambda \in [0,1]$

Dynamic Risk Measures

Definition

A dynamic risk measure is a sequence of one-step conditional risk measures.

$$\varrho^{\beta}_{1,T}(Z_2, \dots, Z_{T+1}) :=$$
$$\rho_1\Big(Z_2 + \rho_2\Big(\beta Z_3 + \rho_3\Big(\beta^2 Z_4 + \cdots + \rho_T\Big(\beta^{T-1} Z_{T+1}\Big)\dots\Big)\Big)\Big)$$

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

Risk-Averse MAB Formulation with Switching Costs

- Evaluation of the risk of the cost sequences $c(x_t, u_t, x_{t+1}) \in \mathcal{Z}_{t+1}$, $t \in \mathbb{N}$ in the play with non-negative switching cost s, for policy π and initial state of $x_1 \in \mathcal{X}$:

$$R^{\pi}(x_1, s) = \varrho^{\beta}\big(c(x_1, u_1, x_2) + s^T u_1, c(x_2, u_2, x_3) + s^T u_2 \mathbf{1}_{u_2 \neq u_1}, \\ c(x_3, u_3, x_4) + s^T u_3 \mathbf{1}_{u_3 \neq u_2}, \dots\big)$$

$$R(x_1, s) = \min_{\pi \in \Pi} R^{\pi}(x_1, s)$$

Π: the class of stationary admissible policies for the problem

Ruszczyński (2010, Theorem 4)

An infinite horizon stationary Markov decision process with dynamic risk measures has a stationary optimal policy.

Risk-Averse MAB Formulation with Switching Costs

- Evaluation of the risk of the cost sequences $c(x_t, u_t, x_{t+1}) \in \mathcal{Z}_{t+1}$, $t \in \mathbb{N}$ in the play with non-negative switching cost s, for policy π and initial state of $x_1 \in \mathcal{X}$:

$$R^{\pi}(x_1, s) = \varrho^{\beta}\big(c(x_1, u_1, x_2) + s^T u_1, c(x_2, u_2, x_3) + s^T u_2 \mathbf{1}_{u_2 \neq u_1}, \\ c(x_3, u_3, x_4) + s^T u_3 \mathbf{1}_{u_3 \neq u_2}, \dots\big)$$

$$R(x_1, s) = \min_{\pi \in \Pi} R^{\pi}(x_1, s)$$

Π: the class of stationary admissible policies for the problem

Ruszczyński (2010, Theorem 4)

An infinite horizon stationary Markov decision process with dynamic risk measures has a stationary optimal policy.

Properties of Risk-Averse Bandits with Switching Costs

1. **Switching Costs in Risk-Averse Setting**:
   Every risk-averse bandit problem in which switching away from and switching to an arm incurs costs has an equivalent risk-averse bandit problem in which only switching to an arm incurs costs.

2. **Structure of the Optimal Policy**:
   There is no optimal index for the risk-averse bandits with switching costs.

Index Heuristics for Risk-Averse Bandits

Definition 3 (Malekipirbazari and Çavuş (2021))

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

$$\nu^i(x_1^i) := \sup_{\tau^i > 1} \frac{\varrho^{\beta}_{1,\tau^i-1}\left(c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \ldots, c^i(x^i_{\tau^i-1}, 1, x^i_{\tau^i})\right)}{\varrho^{\beta}_{1,\tau^i-1}(-1, -1, \ldots, -1)}.$$

Definition 4

For $i \in \mathcal{K}$, the risk-averse switching indices (RASI) for each state $x_1^i \in \mathcal{X}^i$ are given by:

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

and

$$\mu^i(x_1^i, 0) := \sup_{\tau^i > 1} \frac{\varrho^{\beta}_{1,\tau^i-1}\left(s^i + c^i(x_1^i, 1, x_2^i), c^i(x_2^i, 1, x_3^i), \ldots, c^i(x^i_{\tau^i-1}, 1, x^i_{\tau^i})\right)}{\varrho^{\beta}_{1,\tau^i-1}(-1, -1, \ldots, -1)}. \tag{2}$$

Definition 5

Select the arm with the highest index value, where the index values are computed for the states in the immediately played arm by (1) and for the states in the other arms by (2).

RASI in Restricted Environments

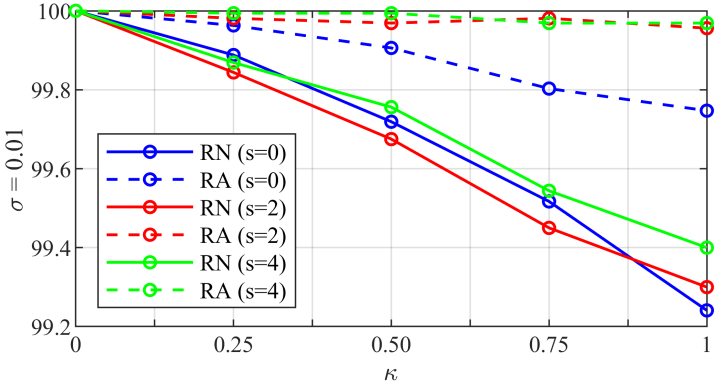
1. **Single-Armed Bandit**:
   In the multiarmed bandits comprised of single-state arms with switching penalties, the index policy described in Definition 5 is optimal.

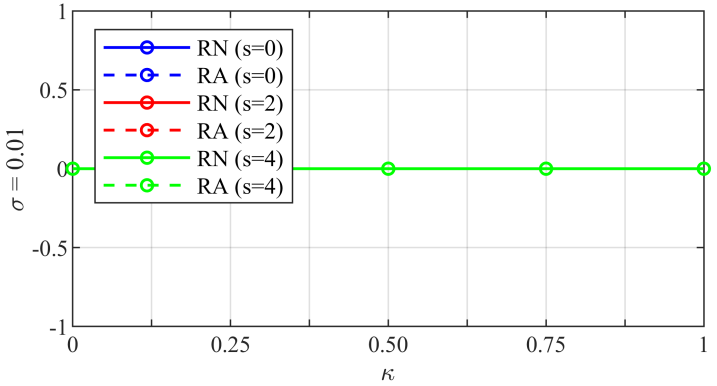
2. **One-Armed Bandit**:
   The risk-averse one-armed bandit problem with switching costs is indexable with *the risk-averse allocation indices* introduced in Definition 5.

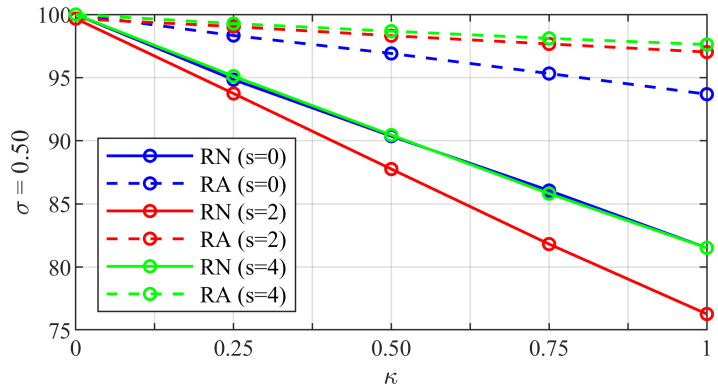
3. **Risk-Neutral Bandits**:
   Decisions about the allocation need to be made only at those steps when the index of the arm that is played is achieved.

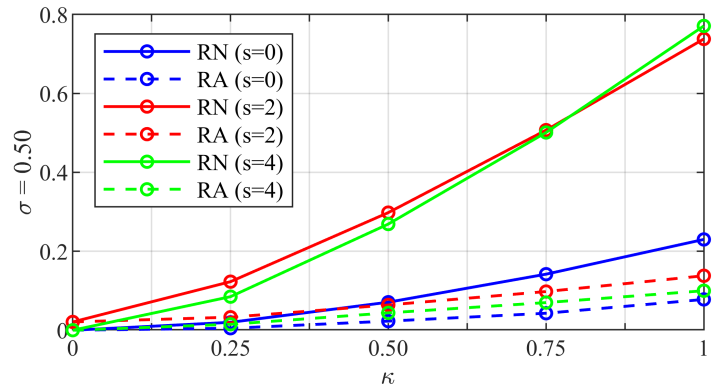
Computational Experiments

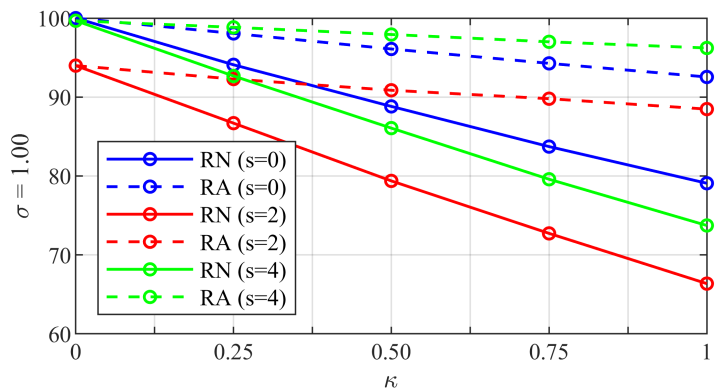
- Three-armed bandits each with four states
- 1000 randomly generated problem instances
- Arm transition probabilities sampled from normalized $U(0, 1)$
- Costs sampled from truncated normal distribution $N(U(-6, -5), \sigma \in \{0.01, 0.5, 1.0\})$ with $\beta \in \{0.50, 0.75, 0.90\}$
- Switching costs sampled from truncated normal distribution $N(s, 0.1s)$ with $s \in \{0, 2, 4\}$

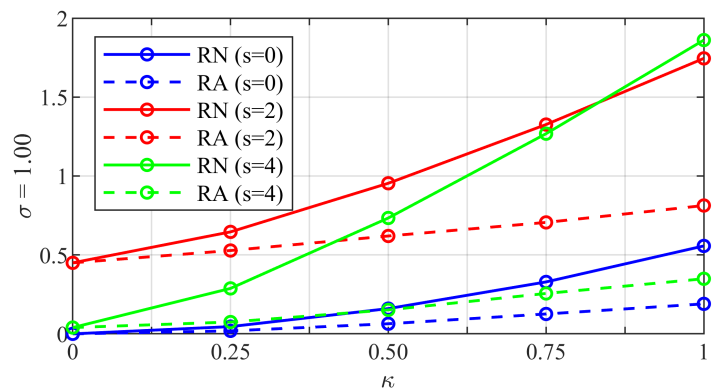












(Optimality Percentage)

(Average of Maximum Suboptimality)

Future Works

1. The conjecture
2. Discussion on the case of the risk-averse bandits where the first switching cost (set-up cost) and remaining switching costs are not the same.
3. Discussion on the case of the risk-averse bandits with switching delays.

References

Asawa, M., Teneketzis, D., 1996. Multi-armed bandits with switching penalties. IEEE transactions on automatic control 41, 328–348.

Banks, J.S., Sundaram, R.K., 1994. Switching costs and the Gittins index. Econometrica: Journal of the Econometric Society , 687–694.

Denardo, E.V., Feinberg, E.A., Rothblum, U.G., 2013. The multi-armed bandit, with constraints. Annals of Operations Research 208, 37–62.

Denardo, E.V., Park, H., Rothblum, U.G., 2007. Risk-sensitive and risk-neutral multiarmed bandits. Mathematics of Operations Research 32, 374–394.

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

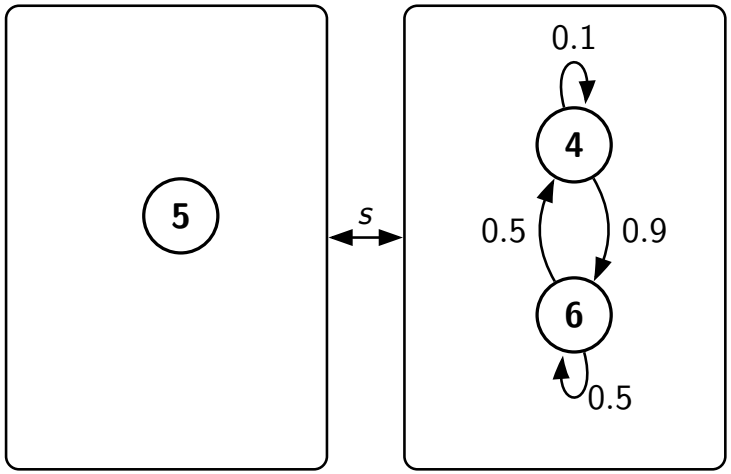
Malekipirbazari, M., Çavuş, Ö., 2021. Risk-averse allocation indices for multiarmed bandit problem. IEEE Transactions on Automatic Control 66, 5522–5529.

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

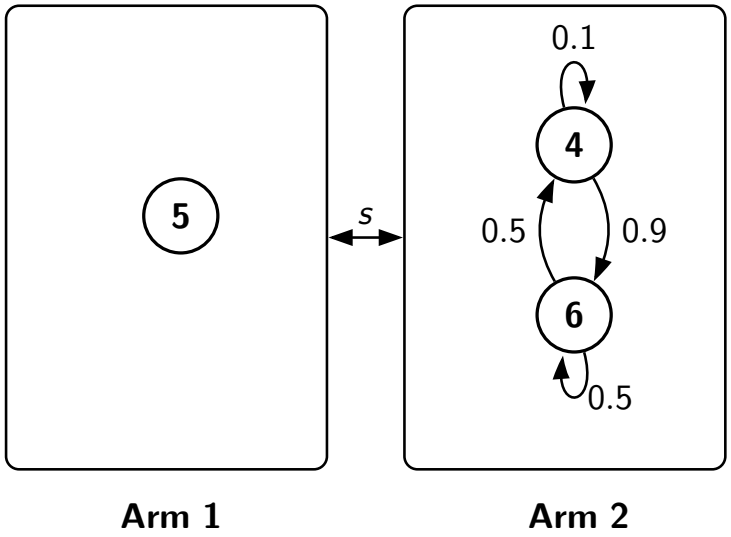
Ruszczyński, A., 2010. Risk-averse dynamic programming for Markov decision processes. Mathematical Programming 125, 235–261.

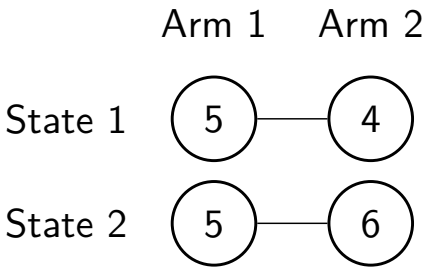
Thank You!

Example of the Optimal Policy

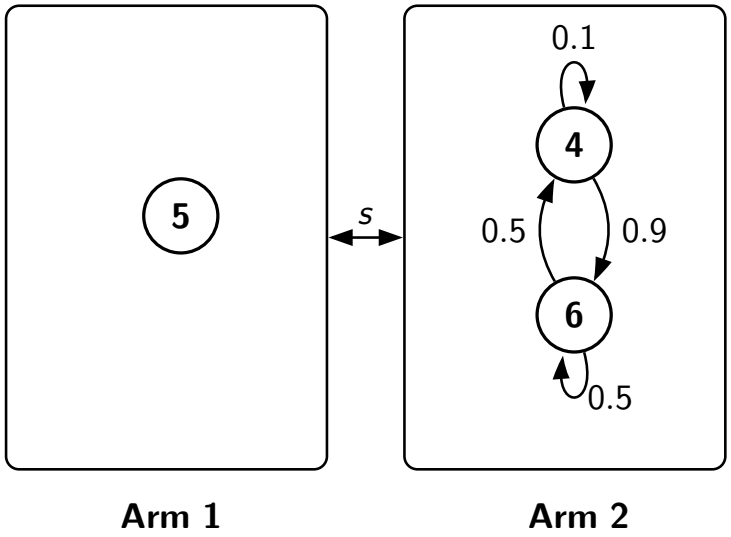


Example of the Optimal Policy

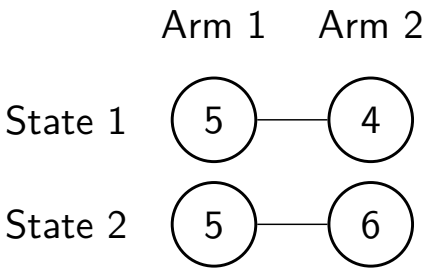




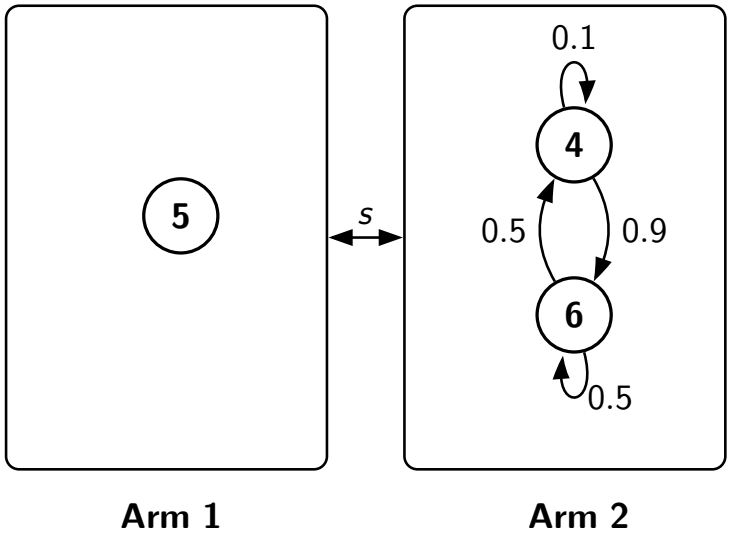
Example of the Optimal Policy

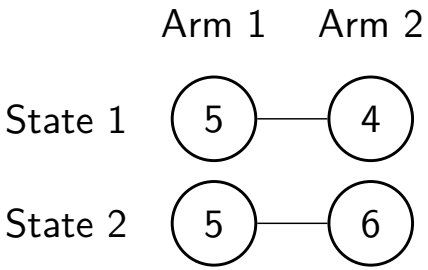


- Switching Costs $= 0$:
  - Optimal Policy for RN: $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$
  - Optimal Policy for RA ($\kappa = 1$): $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$
  - Optimal Policy for RA ($\lambda = 0,\ \alpha = 0.95$): $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$



Example of the Optimal Policy





- Switching Costs $= 0$:
  - Optimal Policy for RN: $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$
  - Optimal Policy for RA ($\kappa = 1$): $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$
  - Optimal Policy for RA ($\lambda = 0,\ \alpha = 0.95$): $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$
- Switching Costs $= +2$:
  - Optimal Policy for RN: $\begin{bmatrix} 2 \\ 2 \end{bmatrix}\begin{bmatrix} 2 \\ 2 \end{bmatrix}$
  - Optimal Policy for RA ($\kappa = 1$): $\begin{bmatrix} 1 \\ 1 \end{bmatrix}\begin{bmatrix} 2 \\ 2 \end{bmatrix}$
  - Optimal Policy for RA ($\lambda = 0,\ \alpha = 0.95$): $\begin{bmatrix} 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix}$