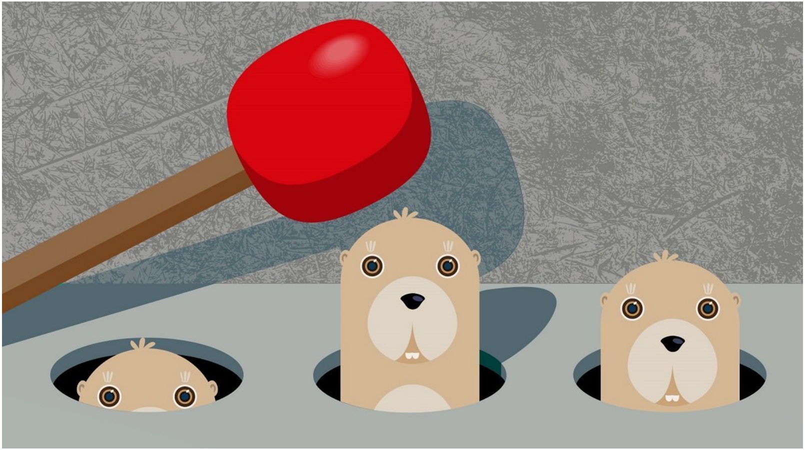
Dynamic spectrum access under partial observations: A restless bandit approach

Nima Akbarzadeh, Aditya Mahajan

McGill University, Electrical and Computer Engineering Department

June 3, 2019

Restless Bandits Example



Channel Scheduling Problem

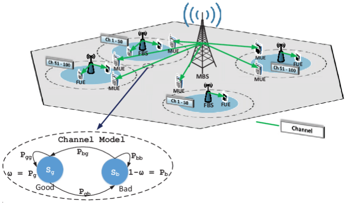
At which time, which channel and which resource should be used?

Features:

- Time-varying channels
- Partially-observable environment
- Resource Allocation

Examples:

- Cognitive radio networks
- Resource constraint jamming



Model (Channel)

- n finite state Markov channels, $\mathcal{N} = \{1, \dots, n\}$.
- State space is finite ordered set $\mathcal{S}^i$, $i \in \mathcal{N}$
  - Markov state process: $\{S_t^i\}_{t \geq 0}$
  - Transition Probability Matrix: P^i
- Resource: rate, power, bandwidth, etc., $\mathcal{R} = \{\emptyset, r_1, \dots, r_k\}$
- Payoff: $\rho^i(s, r)$, $s \in \mathcal{S}^i$, $r \in \mathcal{R}$
  - $\rho^i(s, r) = 0$ if $r = \emptyset$

Example: $\mathcal{S}^i = \{s_{\text{bad}}, s_{\text{good}}\}$, $\mathcal{R} = \{r_{\text{low}}, r_{\text{high}}\}$

$$\rho^i(s, r) = \begin{cases} r_{\text{low}}, & \text{if } r = r_{\text{low}} \\ r_{\text{high}}, & \text{if } r = r_{\text{high}} \text{ and } s = s_{\text{good}} \\ 0, & \text{if } r = r_{\text{high}} \text{ and } s = s_{\text{bad}} \end{cases}$$

Model (Channel)

- n finite state Markov channels, $\mathcal{N} = \{1, \ldots, n\}$.
- State space is finite ordered set $\mathcal{S}^i$, $i \in \mathcal{N}$
  - Markov state process: $\{S_t^i\}_{t \geq 0}$
  - Transition Probability Matrix: P^i
- Resource: rate, power, bandwidth, etc., $\mathcal{R} = \{\emptyset, r_1, \ldots, r_k\}$
- Payoff: $\rho^i(s, r)$, $s \in \mathcal{S}^i$, $r \in \mathcal{R}$
  - $\rho^i(s, r) = 0$ if $r = \emptyset$

Example: $\mathcal{S}^i = \{s_{\text{bad}}, s_{\text{good}}\}$, $\mathcal{R} = \{r_{\text{low}}, r_{\text{high}}\}$

$$\rho^i(s, r) = \begin{cases} r_{\text{low}}, & \text{if } r = r_{\text{low}} \\ r_{\text{high}}, & \text{if } r = r_{\text{high}} \text{ and } s = s_{\text{good}} \\ 0, & \text{if } r = r_{\text{high}} \text{ and } s = s_{\text{bad}} \end{cases}$$

Model (Transmitter)

Two decisions to make at each time t:

- Select L channels indexed by $\mathcal{L}_t$
  $A_t^i = 1$ if $i \in \mathcal{L}_t$ and 0 otherwise
- Select resources denoted by R_t^i
  $R_t^i = \emptyset$ if $i \notin \mathcal{L}_t$

Observation Process:

$$Y_t^i = \begin{cases} S_t^i, & \text{if } A_t^i = 1 \\ \mathfrak{E}, & \text{if } A_t^i = 0. \end{cases}$$

Strategies:

$$\mathbf{A}_t = f_t(\mathbf{Y}_{0:t-1}, \mathbf{R}_{0:t-1}, \mathbf{A}_{0:t-1}),$$
$$\mathbf{R}_t = g_t(\mathbf{Y}_{0:t-1}, \mathbf{R}_{0:t-1}, \mathbf{A}_{0:t-1}, \mathbf{A}_t).$$

Model (Transmitter)

Two decisions to make at each time t:

- Select L channels indexed by $\mathcal{L}_t$
  $A_t^i = 1$ if $i \in \mathcal{L}_t$ and 0 otherwise
- Select resources denoted by R_t^i
  $R_t^i = \emptyset$ if $i \notin \mathcal{L}_t$

Observation Process:

$$Y_t^i = \begin{cases} S_t^i, & \text{if } A_t^i = 1 \\ \mathfrak{E}, & \text{if } A_t^i = 0. \end{cases}$$

Strategies:

$$\mathbf{A}_t = f_t(\mathbf{Y}_{0:t-1}, \mathbf{R}_{0:t-1}, \mathbf{A}_{0:t-1}),$$
$$\mathbf{R}_t = g_t(\mathbf{Y}_{0:t-1}, \mathbf{R}_{0:t-1}, \mathbf{A}_{0:t-1}, \mathbf{A}_t).$$

Model (Transmitter)

Two decisions to make at each time t:

- Select L channels indexed by $\mathcal{L}_t$
  $A_t^i = 1$ if $i \in \mathcal{L}_t$ and 0 otherwise
- Select resources denoted by R_t^i
  $R_t^i = \emptyset$ if $i \notin \mathcal{L}_t$

Observation Process:

$$Y_t^i = \begin{cases} S_t^i, & \text{if } A_t^i = 1 \\ \mathfrak{E}, & \text{if } A_t^i = 0. \end{cases}$$

Strategies:

$$\mathbf{A}_t = f_t(\mathbf{Y}_{0:t-1}, \mathbf{R}_{0:t-1}, \mathbf{A}_{0:t-1}),$$
$$\mathbf{R}_t = g_t(\mathbf{Y}_{0:t-1}, \mathbf{R}_{0:t-1}, \mathbf{A}_{0:t-1}, \mathbf{A}_t).$$

Model (Optimization Problem)

Problem

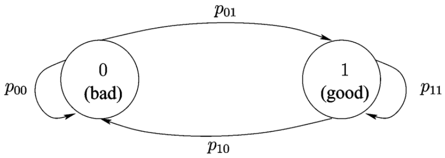
Given a discount factor $\beta \in (0,1)$, *a set of resources* $\mathcal{R}$, *and the state space, transition probability, and reward function* $(\mathcal{S}^i, P^i, \rho^i)_{i \in \mathcal{N}}$ *for all channels, choose a communication strategy* $(\mathbf{f}, \mathbf{g})$ *to maximize*

$$J(\mathbf{f}, \mathbf{g}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \sum_{i \in \mathcal{N}} \rho^i(S_t^i, R_t^i) A_t^i\right].$$

Literature Review and Approaches

Partially Observable Markov Decision Process (POMDP).

- POMDP models suffer from curse of dimensionality:
  - The state space size is exponential in the number of channels
- Simplified modelling assumptions:
  - Two state Gilbert-Elliot channels
  - Multi-state channels but identical
  - Fully-observable Markov Decision Process (MDP)



Our contributions

- Multi-state non-identical channels
- Restless Bandit approach
- Convert the POMDP into a countable MDP
- Finite-state Approximation of the MDP

POMDP (Belief State)

Belief state:

$$\Pi_t^i(s) = \mathbb{P}(S_t^i = s \mid Y_{0:t-1}^i, R_{0:t-1}^i, A_{0:t-1}^i).$$

Proposition

Let $\mathbf{\Pi}_t$ *denote* $(\Pi_t^1, \ldots, \Pi_t^n)$*. Then, without loss of optimality,*

$$\mathbf{A}_t = f_t(\mathbf{\Pi}_t)$$
$$\mathbf{R}_t = g_t(\mathbf{\Pi}_t, \mathbf{A}_t).$$

Recall: $\mathbf{f}$ *is chennel selection policy and* $\mathbf{g}$ *is resource selection policy.*

Optimal Resource Allocation Strategy

No need for joint optimization of $(\mathbf{f}, \mathbf{g})$.
Let

$$\bar{\rho}^i(\pi) := \max_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}^i} \pi(s) \rho^i(s, r),$$

$$r^{i,*}(\pi) := \arg\max_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}^i} \pi(s) \rho^i(s, r).$$

Proposition

Define $g^{i,*} \colon \Delta(\mathcal{S}^i) \times \{0, 1\} \to \mathcal{R}$ as follows

$$g^{i,*}(\pi, 0) = \emptyset,$$
$$g^{i,*}(\pi, 1) = r^{i,*}(\pi).$$

For any channel selection policy, $(\mathbf{g}^*, \mathbf{g}^*, \dots)$ is an optimal resource allocation strategy.

Restless Bandit Model

(1) Each $\{\Pi^i_t\}_{t\geq 0}$, $i \in \mathcal{N}$, is a bandit process.

(2) The transmitter can *activate* L of these processes.

(3) Belief state evolution:

$$\Pi^i_{t+1} = \begin{cases} \delta_{S^i_t}, & \text{if process } i \text{ is activated, } A^i_t = 1, \\ \Pi^i_t \cdot P^i, & \text{if process } i \text{ is passive, } A^i_t = 0. \end{cases}$$

(4) Expected reward:

$$\rho^i_t = \begin{cases} \bar{\rho}^i(\Pi^i_t), & \text{if process } i \text{ is activated, } A^i_t = 1, \\ 0, & \text{if process } i \text{ is passive, } A^i_t = 0. \end{cases}$$

Process:

Dynamics:

$$\dots \to \underbrace{\Pi^i_t \xrightarrow{\mathbf{f}} A^i_t \xrightarrow{\mathbf{g}^*} R^i_t \to Y^i_t \to \rho^i_t}_{\text{time } t} \to \Pi^i_{t+1} \to \dots$$

Restless Bandit Model

(1) Each $\{\Pi_t^i\}_{t\geq 0}$, $i \in \mathcal{N}$, is a bandit process.
(2) The transmitter can *activate* L of these processes.
(3) Belief state evolution:

$$\Pi_{t+1}^i = \begin{cases} \delta_{S_t^i}, & \text{if process } i \text{ is activated, } A_t^i = 1, \\ \Pi_t^i \cdot P^i, & \text{if process } i \text{ is passive, } A_t^i = 0. \end{cases}$$

(4) Expected reward:

$$\rho_t^i = \begin{cases} \bar{\rho}^i(\Pi_t^i), & \text{if process } i \text{ is activated, } A_t^i = 1, \\ 0, & \text{if process } i \text{ is passive, } A_t^i = 0. \end{cases}$$

Process:

Dynamics:

$$\dots \to \underbrace{\Pi_t^i \xrightarrow{\mathbf{f}} A_t^i \xrightarrow{\mathbf{g}^*} R_t^i \to Y_t^i \to \rho_t^i}_{\text{time } t} \to \Pi_{t+1}^i \to \dots$$

Restless Bandit Model

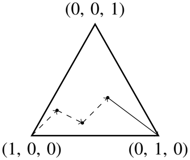
(1) Each $\{\Pi_t^i\}_{t\geq 0}$, $i \in \mathcal{N}$, is a bandit process.
(2) The transmitter can *activate* L of these processes.
(3) Belief state evolution:

$$\Pi_{t+1}^i = \begin{cases} \delta_{S_t^i}, & \text{if process } i \text{ is activated, } A_t^i = 1, \\ \Pi_t^i \cdot P^i, & \text{if process } i \text{ is passive, } A_t^i = 0. \end{cases}$$

(4) Expected reward:

$$\rho_t^i = \begin{cases} \bar{\rho}^i(\Pi_t^i), & \text{if process } i \text{ is activated, } A_t^i = 1, \\ 0, & \text{if process } i \text{ is passive, } A_t^i = 0. \end{cases}$$

Dynamics:



Process:

$$\dots \to \underbrace{\Pi_t^i \xrightarrow{\mathbf{f}} A_t^i \xrightarrow{\mathbf{g}^*} R_t^i \to Y_t^i \to \rho_t^i}_{\text{time } t} \to \Pi_{t+1}^i \to \dots$$

Restless Bandit Model

(1) Each $\{\Pi_t^i\}_{t\geq 0}$, $i \in \mathcal{N}$, is a bandit process.

(2) The transmitter can *activate* L of these processes.

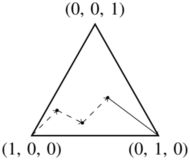
(3) Belief state evolution:

$$\Pi_{t+1}^i = \begin{cases} \delta_{S_t^i}, & \text{if process } i \text{ is activated, } A_t^i = 1, \\ \Pi_t^i \cdot P^i, & \text{if process } i \text{ is passive, } A_t^i = 0. \end{cases}$$

(4) Expected reward:

$$\rho_t^i = \begin{cases} \bar{\rho}^i(\Pi_t^i), & \text{if process } i \text{ is activated, } A_t^i = 1, \\ 0, & \text{if process } i \text{ is passive, } A_t^i = 0. \end{cases}$$

Dynamics:



Process:

$$\ldots \to \underbrace{\Pi_t^i \xrightarrow{\mathrm{f}} A_t^i \xrightarrow{\mathrm{g}^*} R_t^i \to Y_t^i \to \rho_t^i}_{\text{time } t} \to \Pi_{t+1}^i \to \ldots$$

Restless Bandit Model

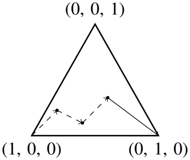
(1) Each $\{\Pi_t^i\}_{t\geq 0}$, $i \in \mathcal{N}$, is a bandit process.
(2) The transmitter can *activate* L of these processes.
(3) Belief state evolution:

$$\Pi_{t+1}^i = \begin{cases} \delta_{S_t^i}, & \text{if process } i \text{ is activated, } A_t^i = 1, \\ \Pi_t^i \cdot P^i, & \text{if process } i \text{ is passive, } A_t^i = 0. \end{cases}$$

(4) Expected reward:

$$\rho_t^i = \begin{cases} \bar{\rho}^i(\Pi_t^i), & \text{if process } i \text{ is activated, } A_t^i = 1, \\ 0, & \text{if process } i \text{ is passive, } A_t^i = 0. \end{cases}$$

Dynamics:



Process:

$$\ldots \to \underbrace{\Pi_t^i \xrightarrow{\mathbf{f}} A_t^i \xrightarrow{\mathbf{g}^*} R_t^i \to Y_t^i \to \rho_t^i}_{\text{time } t} \to \Pi_{t+1}^i \to \ldots$$

Restless Bandit Solution

- The main idea is to decompose the coupled n-channel optimization problem to n independent one-channel problems.
- When the Whittle indexability is satisfied, then one may propose a Whittle index policy.
- The channels with minimum indices are selected.
- The index strategy performs close-to-optimal for many applications in the state-of-arts works.

Goal:
We provide an efficient algorithm to check the indexability and compute the Whittle index.

Problem Decomposition

Modified per-step reward: $(\bar{\rho}^i(\pi) - \lambda)a^i$ where λ can be viewed as the cost for transmitting over channel i.

Problem

Given channel $i \in \mathcal{N}$, *the discount factor* $\beta \in (0,1)$, *the cost* $\lambda \in \mathbb{R}$, *and the belief state space, transition probability, reward function tuple* $(\Delta(\mathcal{S}^i), P^i, \rho^i)$, *choose a policy* $f^i : \Delta(\mathcal{S}^i) \to \{0,1\}$ *to maximize*

$$J^i_\lambda(f^i) := \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \left(\bar{\rho}^i(\Pi^i_t) - \lambda\right) A^i_t\right].$$

Dynamic Programming (Belief State)

Theorem

Let $V_\lambda^i : \Delta(\mathcal{S}^i) \to \mathbb{R}$ *be the unique fixed point of equation*

$$V_\lambda^i(\pi) = \max_{a \in \{0,1\}} Q_\lambda^i(\pi, a)$$

where

$$Q_\lambda^i(\pi, 0) = \beta V_\lambda^i(\pi \cdot P^i)$$
$$Q_\lambda^i(\pi, 1) = \bar{\rho}_\lambda^i(\pi) - \lambda + \beta \sum_{s \in \mathcal{S}^i} \pi(s) V_\lambda^i(\delta_s).$$

Let $f_\lambda^i(\pi) = 1$ *if* $Q_\lambda^i(\pi, 1) \geq Q_\lambda^i(\pi, 0)$, *and* $f_\lambda^i(\pi) = 0$ *otherwise. Then,* f_λ^i *is optimal for Problem 2.*

Challenge: Continuous state space!

Dynamic Programming (Belief State)

Theorem

Let $V_\lambda^i : \Delta(\mathcal{S}^i) \to \mathbb{R}$ *be the unique fixed point of equation*

$$V_\lambda^i(\pi) = \max_{a \in \{0,1\}} Q_\lambda^i(\pi, a)$$

where

$$Q_\lambda^i(\pi, 0) = \beta V_\lambda^i(\pi \cdot P^i)$$
$$Q_\lambda^i(\pi, 1) = \bar{\rho}_\lambda^i(\pi) - \lambda + \beta \sum_{s \in \mathcal{S}^i} \pi(s) V_\lambda^i(\delta_s).$$

Let $f_\lambda^i(\pi) = 1$ *if* $Q_\lambda^i(\pi, 1) \geq Q_\lambda^i(\pi, 0)$, *and* $f_\lambda^i(\pi) = 0$ *otherwise. Then,* f_λ^i *is optimal for Problem 2.*

Challenge: Continuous state space!

Information State

Let $O_t^i \in \mathcal{S}^i$ denote the last observed state of channel i and $K_t^i \in \mathbb{Z}_{\geq 0}$ denote the time since the last observation. Then, we have

$$(O_{t+1}^i, K_{t+1}^i) = \begin{cases} (S_t^i, 0) & \text{if } A_t^i = 1 \\ (O_t^i, K_t^i + 1) & \text{if } A_t^i = 0. \end{cases}$$

Lemma

At any time t, $\Pi_t^i = \delta_{O_t^i} \cdot (P^i)^{K_t^i}$ almost surely.

Example:

$$\begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

Information State

Let $O_t^i \in \mathcal{S}^i$ denote the last observed state of channel i and $K_t^i \in \mathbb{Z}_{\geq 0}$ denote the time since the last observation. Then, we have

$$(O_{t+1}^i, K_{t+1}^i) = \begin{cases} (S_t^i, 0) & \text{if } A_t^i = 1 \\ (O_t^i, K_t^i + 1) & \text{if } A_t^i = 0. \end{cases}$$

Lemma

At any time t, $\Pi_t^i = \delta_{O_t^i} \cdot (P^i)^{K_t^i}$ almost surely.

Example:

$$\begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

Information State

Let $O_t^i \in \mathcal{S}^i$ denote the last observed state of channel i and $K_t^i \in \mathbb{Z}_{\geq 0}$ denote the time since the last observation. Then, we have

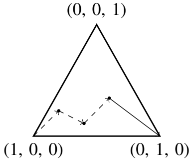
$$(O_{t+1}^i, K_{t+1}^i) = \begin{cases} (S_t^i, 0) & \text{if } A_t^i = 1 \\ (O_t^i, K_t^i + 1) & \text{if } A_t^i = 0. \end{cases}$$

Lemma

At any time t, $\Pi_t^i = \delta_{O_t^i} \cdot (P^i)^{K_t^i}$ almost surely.

Example:

$$\begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$



Information State

Let $O_t^i \in \mathcal{S}^i$ denote the last observed state of channel i and $K_t^i \in \mathbb{Z}_{\geq 0}$ denote the time since the last observation. Then, we have

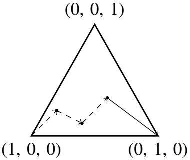
$$(O_{t+1}^i, K_{t+1}^i) = \begin{cases} (S_t^i, 0) & \text{if } A_t^i = 1 \\ (O_t^i, K_t^i + 1) & \text{if } A_t^i = 0. \end{cases}$$

Lemma

At any time t, $\Pi_t^i = \delta_{O_t^i} \cdot (P^i)^{K_t^i}$ almost surely.

Example:

$$\begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$



$$(1,0,0) = (1,0,0) \cdot \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}^0$$

Information State

Let $O_t^i \in \mathcal{S}^i$ denote the last observed state of channel i and $K_t^i \in \mathbb{Z}_{\geq 0}$ denote the time since the last observation. Then, we have

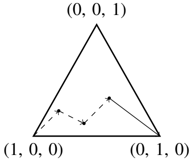
$$(O_{t+1}^i, K_{t+1}^i) = \begin{cases} (S_t^i, 0) & \text{if } A_t^i = 1 \\ (O_t^i, K_t^i + 1) & \text{if } A_t^i = 0. \end{cases}$$

Lemma

At any time t, $\Pi_t^i = \delta_{O_t^i} \cdot (P^i)^{K_t^i}$ almost surely.

Example:

$$\begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$



$$(1,0,0) \cdot \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}^1$$

Information State

Let $O_t^i \in \mathcal{S}^i$ denote the last observed state of channel i and $K_t^i \in \mathbb{Z}_{\geq 0}$ denote the time since the last observation. Then, we have

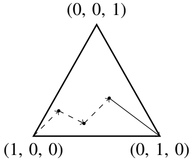
$$(O_{t+1}^i, K_{t+1}^i) = \begin{cases} (S_t^i, 0) & \text{if } A_t^i = 1 \\ (O_t^i, K_t^i + 1) & \text{if } A_t^i = 0. \end{cases}$$

Lemma

At any time t, $\Pi_t^i = \delta_{O_t^i} \cdot (P^i)^{K_t^i}$ almost surely.

Example:

$$\begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$



$$(1, 0, 0) \cdot \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}^2$$

Information State

Let $O_t^i \in \mathcal{S}^i$ denote the last observed state of channel i and $K_t^i \in \mathbb{Z}_{\geq 0}$ denote the time since the last observation. Then, we have

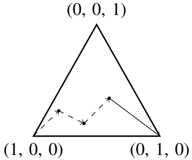
$$(O_{t+1}^i, K_{t+1}^i) = \begin{cases} (S_t^i, 0) & \text{if } A_t^i = 1 \\ (O_t^i, K_t^i + 1) & \text{if } A_t^i = 0. \end{cases}$$

Lemma

At any time t, $\Pi_t^i = \delta_{O_t^i} \cdot (P^i)^{K_t^i}$ almost surely.

Example:

$$\begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$



$$(1, 0, 0) \cdot \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}^3$$

Information State

Let $O_t^i \in \mathcal{S}^i$ denote the last observed state of channel i and $K_t^i \in \mathbb{Z}_{\geq 0}$ denote the time since the last observation. Then, we have

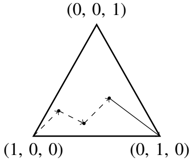
$$(O_{t+1}^i, K_{t+1}^i) = \begin{cases} (S_t^i, 0) & \text{if } A_t^i = 1 \\ (O_t^i, K_t^i + 1) & \text{if } A_t^i = 0. \end{cases}$$

Lemma

At any time t, $\Pi_t^i = \delta_{O_t^i} \cdot (P^i)^{K_t^i}$ almost surely.

Example:

$$\begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$



$$(0, 1, 0) \cdot \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}^{0}$$

Dynamic Programming (Information State)

- Difficult to solve dynamic programming based on belief state π^i as the state space is $\Delta(\mathcal{S}^i)$.
- A new dynamic programming can be written considering the information state (o^i, k^i) where the state space is $\mathcal{S}^i \times \mathbb{Z}_{\geq 0}$.

Pros and cons:

The state space is countable but still dynamic programming is computationally infeasible.

Finite-State Approximation

Dynamic Programming (Finite state space & Computable!)

Given $m \in \mathbb{N}$, let $\mathbb{N}_m := \{0, \ldots, m\}$ and $V^i_{\lambda,m} : \mathcal{S}^i \times \mathbb{N}_m \to \mathbb{R}$ denote the unique fixed point of

$$\begin{aligned}
V^i_{\lambda,m}(o,k) &= \max_{a \in \{0,1\}} \{Q^i_{\lambda,m}(o,k,a)\} \\
Q^i_{\lambda,m}(o,k,0) &= \beta V^i_{\lambda,m}(o, k+1 \wedge m) \\
Q^i_{\lambda,m}(o,k,1) &= \bar{\rho}^i(o,k) - \lambda + \beta \sum_{s \in \mathcal{S}^i} (P^i)^k_{os} V^i_{\lambda,m}(s,0).
\end{aligned}$$

Let $f^i_{\lambda,m}(o,k) = 1$ if $Q^i_{\lambda,m}(o,k,1) \geq Q^i_{\lambda,m}(o,k,0)$, and $f^i_{\lambda,m}(o,k) = 0$ o.w.

Finite-State Approximation

Approximation Limits

(i) $\lim_{m\to\infty} V^i_{\lambda,m}(o,k) = V^i_\lambda(o,k)$, $\forall (o,k) \in \mathcal{S}^i \times \mathbb{Z}_{\geq 0}$.

(ii) Let $f^{i,*}_\lambda(o,k)$ be any fixed point of $\{f^i_{\lambda,m}(o,k)\}_{m\geq 1}$. Then, the policy $f^{i,*}_\lambda(o,k)$ is optimal for sub-problem i.

Indexability

Let passive set for process i be

$$\mathcal{P}^i_\lambda = \left\{(o,k) \in \mathcal{S}^i \times \mathbb{N}_m : f^i_{\lambda,m}(o,k) = 0\right\}.$$

Recall: $f^i_{\lambda,m}$ is the policy obtained by dynamic programming.

Definition (Indexability)

For any $\lambda_1, \lambda_2 \in \mathbb{R}$ process i is indexable if

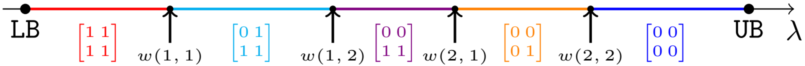
$$\lambda_1 \leq \lambda_2 \implies \mathcal{P}^i_{\lambda_1} \subseteq \mathcal{P}^i_{\lambda_2}.$$

Definition (Whittle index)

The Whittle index of information state (o,k) of process i is defined as

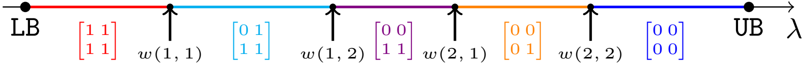
$$w^i(o,k) = \inf\left\{\lambda \in \mathbb{R} : (o,k) \notin \mathcal{P}^i_\lambda\right\}.$$

Algorithms

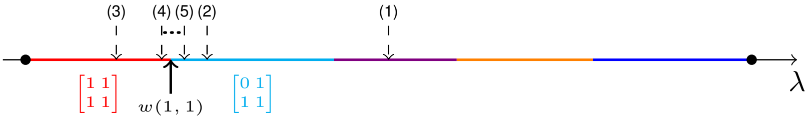


Procedure:

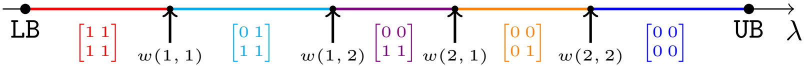
Algorithms



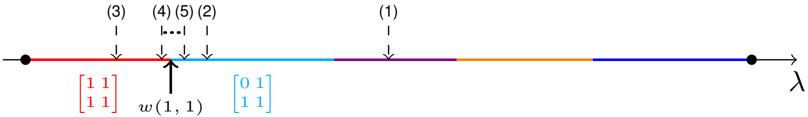
Procedure:

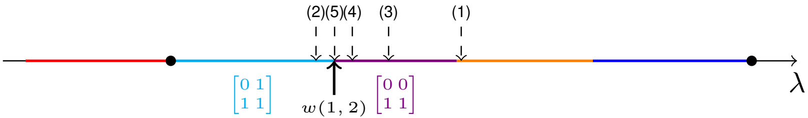


Algorithms

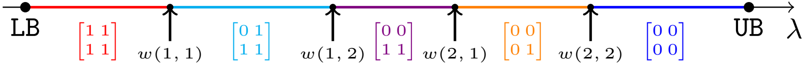


Procedure:

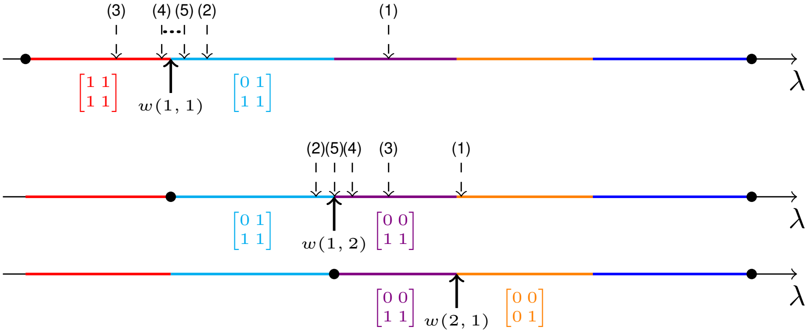




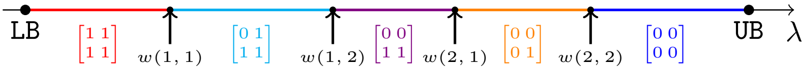
Algorithms



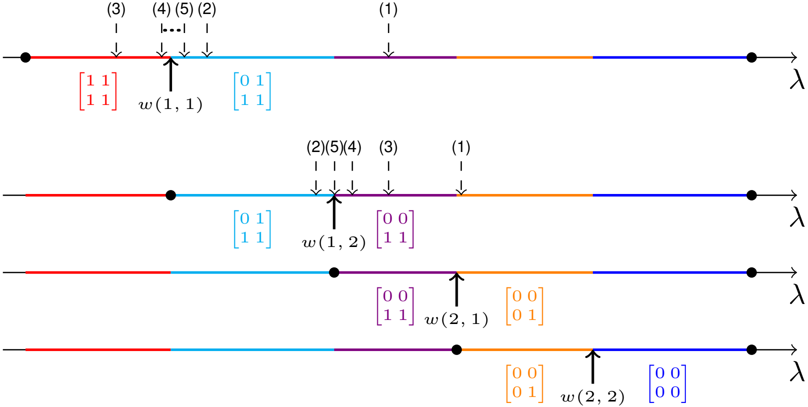
Procedure:



Algorithms



Procedure:



Algorithms

Whittle Index Policy:

At each time,

- Obtain the Whittle index corresponding to current information state of all channels.
- Transmit over the L channels with the smallest Whittle indices.

Conclusion

- Dynamic spectrum access problem for transmitting over multiple channels with partially observed channel state.
- Resource allocation strategy can be computed offline and is not affecting the channel selection strategy.
- To circumvent the curse of dimensionality, we considered the problem as a restless bandit and use the Whittle index heuristic.
- By reachable set of beliefs, the problem is converted from the belief-valued processes into a countable-state process.
- We developed low-complexity algorithms to check whether each channel is indexable and if so, compute the Whittle index for each information state.

Q&A

Thank you!