Detecting an Odd Restless Markov Arm with a Trembling Hand

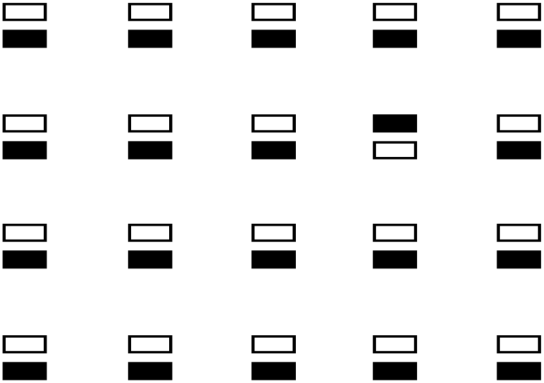
P. N. Karthik and Rajesh Sundaresan
Department of Electrical Communication Engineering,
Indian Institute of Science, Bangalore
Email: (periyapatna, rajeshs)@iisc.ac.in

2020 IEEE International Symposium on Information Theory

Outline

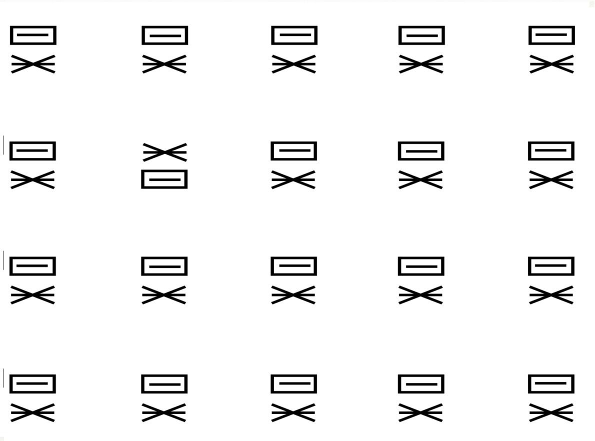
Motivation

Visual Search Experiment 1



Identify the location of the odd image. No guessing allowed.

Visual Search Experiment 2



Identify the location of the odd image. No guessing allowed.

Visual Search with Static Images

<u>Goal</u>: to identify the location of the odd image as quickly and accurately as possible

Visual Search with Static Images

Goal: to identify the location of the odd image as quickly and accurately as possible

The time to identify the location of the odd image depends on

(a) the prescribed accuracy (or error tolerance) level

(b) the "closeness" between the odd and the non-odd images

Visual Search with Static Images

Goal: to identify the location of the odd image as quickly and accurately as possible

> The time to identify the location of the odd image depends on
> (a) the prescribed accuracy (or error tolerance) level
> (b) the "closeness" between the odd and the non-odd images

- Vaidhiyan et al. [1, 2] showed that given an error tolerance level $\epsilon > 0$, the time to identify the location of the odd image grows as $\log\left(\frac{1}{\epsilon}\right) \cdot \frac{1}{D^*}$, where D^* is a measure of the closeness between the odd and the non-odd images
- Vaidhiyan et al. also demonstrated that the growth rate of $\log\left(\frac{1}{\epsilon}\right) \cdot \frac{1}{D^*}$ is tight in the limit as $\epsilon \downarrow 0$

From Static Images to Movies

- A MATLAB® demo
- A total of 8 drifting-dots moving images (movies)
- The drift in one of the movies (the "odd" movie) is different from the common drift of all the other movies
- **Goal: to identify the "odd" movie as quickly and accurately as possible**

From Static Images to Movies

- A MATLAB® demo
- A total of 8 drifting-dots moving images (movies)
- The drift in one of the movies (the "odd" movie) is different from the common drift of all the other movies
- **Goal: to identify the "odd" movie as quickly and accurately as possible**
- **How does the time to identify the odd movie grow as a function of (a) the error tolerance, and (b) the "closeness" between the odd and the non-odd movies?**

From Static Images to Movies

- A MATLAB® demo
- A total of 8 drifting-dots moving images (movies)
- The drift in one of the movies (the "odd" movie) is different from the common drift of all the other movies
- **Goal: to identify the "odd" movie as quickly and accurately as possible**
- **How does the time to identify the odd movie grow as a function of (a) the error tolerance, and (b) the "closeness" between the odd and the non-odd movies?**

A systematic analysis of this question, along the lines of [1, 2], requires an understanding of the *odd restless Markov arm problem*, which is the subject of this paper

The Notion of Trembling Hand

Trembling Hand in Visual Search

- Typically, in visual search experiments, the subject (or decision maker) intends to focus at a certain location, but the actual focus location differs from the intended focus location with a small probability

Trembling Hand in Visual Search

- Typically, in visual search experiments, the subject (or decision maker) intends to focus at a certain location, but the actual focus location differs from the intended focus location with a small probability

- Suppose B_t is the subject's intended focus location, and A_t is the actual focus location at time t. Then,

$$A_t = \begin{cases} B_t & \text{w.p. } 1-\eta, \\ \text{unif. randomly chosen location} & \text{w.p. } \eta, \end{cases}$$

for some $\eta > 0$

Trembling Hand in Visual Search

- Typically, in visual search experiments, the subject (or decision maker) intends to focus at a certain location, but the actual focus location differs from the intended focus location with a small probability

- Suppose B_t is the subject's intended focus location, and A_t is the actual focus location at time t. Then,

$$A_t = \begin{cases} B_t & \text{w.p. } 1-\eta, \\ \text{unif. randomly chosen location} & \text{w.p. } \eta, \end{cases}$$

for some $\eta > 0$

- **We refer to the above phenomenon as the decision maker having a trembling hand, with η being the corresponding trembling hand parameter**

The Odd Restless Markov Arm Problem

Visual Search with Movies and Multi-armed Bandits

Visual Search with Movies	Multi-armed Bandits
Movie	Arm
Movie frame	Observation
Positions of dots in two successive frames of a movie are related to one another	Successive observations from an arm form a Markov process
The drift of one of the movies is different from the common drift of the other movies	The TPM of one of the Markov processes is different from the common TPM of the others
Each movie **continues to play** whether or not the movie is observed	The arms are **restless** (terminology from Whittle [3])
A movie is **paused** when not observed	The arms are **rested**
Identifying the odd movie	Identifying the odd arm

TPM: transition probability matrix

Odd Arm Identification with Restless Arms

- A multi-armed bandit with $K \geq 3$ independent arms

Odd Arm Identification with Restless Arms

- A multi-armed bandit with $K \geq 3$ independent arms
- Each arm is a time-homogeneous and ergodic **Markov** process on a common finite state space $\mathcal{S}$

Odd Arm Identification with Restless Arms

- A multi-armed bandit with $K \geq 3$ independent arms
- Each arm is a time-homogeneous and ergodic **Markov** process on a common finite state space $\mathcal{S}$
- The transition probability matrix (TPM) of one of the arms (the **odd** arm) is P_1, while the transition probability matrix of each of the remaining non-odd arms is P_2, where $P_2 \neq P_1$

Odd Arm Identification with Restless Arms

- A multi-armed bandit with $K \geq 3$ independent arms
- Each arm is a time-homogeneous and ergodic **Markov** process on a common finite state space $\mathcal{S}$
- The transition probability matrix (TPM) of one of the arms (the **odd** arm) is P_1, while the transition probability matrix of each of the remaining non-odd arms is P_2, where $P_2 \neq P_1$
- A decision maker who knows P_1 and P_2 wishes to identify the odd arm as quickly and accurately as possible

Odd Arm Identification with Restless Arms

- A multi-armed bandit with $K \geq 3$ independent arms
- Each arm is a time-homogeneous and ergodic **Markov** process on a common finite state space $\mathcal{S}$
- The transition probability matrix (TPM) of one of the arms (the **odd** arm) is P_1, while the transition probability matrix of each of the remaining non-odd arms is P_2, where $P_2 \neq P_1$
- A decision maker who knows P_1 and P_2 wishes to identify the odd arm as quickly and accurately as possible
- The decision maker is allowed to observe the state of only one arm at any given time. The unobserved arms continue to evolve (**restless** arms)

Odd Arm Identification with Restless Arms

- A multi-armed bandit with $K \geq 3$ independent arms
- Each arm is a time-homogeneous and ergodic **Markov** process on a common finite state space $\mathcal{S}$
- The transition probability matrix (TPM) of one of the arms (the **odd** arm) is P_1, while the transition probability matrix of each of the remaining non-odd arms is P_2, where $P_2 \neq P_1$
- A decision maker who knows P_1 and P_2 wishes to identify the odd arm as quickly and accurately as possible
- The decision maker is allowed to observe the state of only one arm at any given time. The unobserved arms continue to evolve (**restless** arms)

Our Contributions

Our Contributions - 1

- Given an error tolerance level $\epsilon > 0$, a TPM P_1 for the odd movie and a TPM P_2 for the non-odd movies, we show that the average time to identify the odd movie grows as
$$\log\left(\frac{1}{\epsilon}\right) \cdot \frac{1}{R^*(P_1, P_2)}$$
where $R^*(P_1, P_2)$ is a measure of closeness between the odd movie and the non-odd movies

Our Contributions - 1

- Given an error tolerance level $\epsilon > 0$, a TPM P_1 for the odd movie and a TPM P_2 for the non-odd movies, we show that the average time to identify the odd movie grows as

$$\log\left(\frac{1}{\epsilon}\right) \cdot \frac{1}{R^*(P_1, P_2)}$$

  where $R^*(P_1, P_2)$ is a measure of closeness between the odd movie and the non-odd movies

- We provide an explicit characterisation for $R^*(P_1, P_2)$ – the first known characterisation of this constant for the setting of restless arms

Our Contributions - 1

- Given an error tolerance level $\epsilon > 0$, a TPM P_1 for the odd movie and a TPM P_2 for the non-odd movies, we show that the average time to identify the odd movie grows as
$$\log\left(\frac{1}{\epsilon}\right) \cdot \frac{1}{R^*(P_1, P_2)}$$
where $R^*(P_1, P_2)$ is a measure of closeness between the odd movie and the non-odd movies
- We provide an explicit characterisation for $R^*(P_1, P_2)$ – the first known characterisation of this constant for the setting of restless arms
- We show that the above growth rate of $\log\left(\frac{1}{\epsilon}\right) \cdot \frac{1}{R^*(P_1,P_2)}$ is tight in the asymptotic limit as $\epsilon \downarrow 0$

Our Contributions - 2

- We identify a family of Markov decision problems (MDPs) and stitch together the solutions to these MDPs to arrive at our asymptotic lower and upper bounds

Our Contributions - 2

- We identify a family of Markov decision problems (MDPs) and stitch together the solutions to these MDPs to arrive at our asymptotic lower and upper bounds

- The presence of the trembling hand ($\eta > 0$) translates to a key ergodicity property for the MDPs. We leverage this in our analysis of the lower and the upper bounds

Our Contributions - 2

- We identify a family of Markov decision problems (MDPs) and stitch together the solutions to these MDPs to arrive at our asymptotic lower and upper bounds

- The presence of the trembling hand ($\eta > 0$) translates to a key ergodicity property for the MDPs. We leverage this in our analysis of the lower and the upper bounds

- Traditional works on MDPs deal with reward maximisation, whereas our work is based on the theme of optimal stopping

Our Contributions - 2

- We identify a family of Markov decision problems (MDPs) and stitch together the solutions to these MDPs to arrive at our asymptotic lower and upper bounds

- The presence of the trembling hand ($\eta > 0$) translates to a key ergodicity property for the MDPs. We leverage this in our analysis of the lower and the upper bounds

- Traditional works on MDPs deal with reward maximisation, whereas our work is based on the theme of optimal stopping

- The framework of MDPs provides us with the right 'global' perspective to solve the odd restless Markov arm problem. This is in contrast to the 'local' perspectives offered by the prior works

Arm Delays and Last Observed States

Arm Delays and Last Observed States

- **The continued evolution of the unobserved arms makes it necessary to keep track of**
  - the time elapsed since each arm was last selected (the arm's **delay**)
  - the **last observed state** of each arm

Arm Delays and Last Observed States

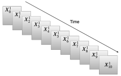
- **The continued evolution of the unobserved arms makes it necessary to keep track of**
  - the time elapsed since each arm was last selected (the arm's **delay**)
  - the **last observed state** of each arm

- The notion of arm delays is superfluous in the case when
  - each arm yields independent and identically distributed (iid) observations as in [1, 2]
  - each arm yields Markov observations and the arms are rested as in [4]

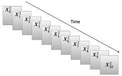
Arm Delays and Last Observed States

- **The continued evolution of the unobserved arms makes it necessary to keep track of**
  - the time elapsed since each arm was last selected (the arm's **delay**)
  - the **last observed state** of each arm
- The notion of arm delays is superfluous in the case when
  - each arm yields independent and identically distributed (iid) observations as in [1, 2]
  - each arm yields Markov observations and the arms are rested as in [4]

Understanding Arm Delays and Last Observed States

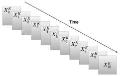


$\{X_t^1 : t = 0, 1, 2, \ldots\}$



$\{X_t^2 : t = 0, 1, 2, \ldots\}$

$\cdots\cdots$

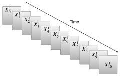


$\{X_t^K : t = 0, 1, 2, \ldots\}$

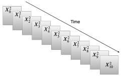
Time	**Delay** of arm 1	**Delay** of arm 2	$\cdots$	**Delay** of arm K
	LOS of arm 1	**LOS** of arm 2	$\cdots$	**LOS** of arm K
$t = K$	$d_1(t) = K$	$d_2(t) = K - 1$	$\cdots$	$d_K(t) = 1$
	$i_1(t) = X_0^1$	$i_2(t) = X_1^2$	$\cdots$	$i_K(t) = X_{K-1}^K$
$t = K + 1$				

Delay: Time since last observation **LOS: Last Observed State**

Understanding Arm Delays and Last Observed States

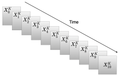


$\{X_t^1 : t = 0, 1, 2, \ldots\}$



$\{X_t^2 : t = 0, 1, 2, \ldots\}$

$\cdots\cdots$

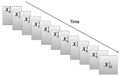


$\{X_t^K : t = 0, 1, 2, \ldots\}$

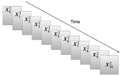
Time	**Delay** of arm 1 **LOS** of arm 1	**Delay** of arm 2 **LOS** of arm 2	$\cdots$ $\cdots$	**Delay** of arm K **LOS** of arm K
$t = K$	$d_1(t) = K$ $i_1(t) = X_0^1$	$d_2(t) = K - 1$ $i_2(t) = X_1^2$	$\cdots$ $\cdots$	$d_K(t) = 1$ $i_K(t) = X_{K-1}^K$
$t = K + 1$	$d_1(t) = K + 1$ $i_1(t) = i_1(t - 1)$	$d_2(t) = 1$ $i_2(t) = X_K^2$	$\cdots$ $\cdots$	$d_K(t) = 2$ $i_K(t) = i_K(t - 1)$
$t = K + 2$				

Delay: Time since last observation **LOS: Last Observed State**

Understanding Arm Delays and Last Observed States

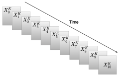


$\{X_t^1 : t = 0, 1, 2, \ldots\}$



$\{X_t^2 : t = 0, 1, 2, \ldots\}$

$\cdots\cdots$



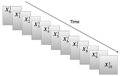
$\{X_t^K : t = 0, 1, 2, \ldots\}$

Time	**Delay** of arm 1 **LOS** of arm 1	**Delay** of arm 2 **LOS** of arm 2	$\cdots$ $\cdots$	**Delay** of arm K **LOS** of arm K
$t = K$	$d_1(t) = K$ $i_1(t) = X_0^1$	$d_2(t) = K - 1$ $i_2(t) = X_1^2$	$\cdots$ $\cdots$	$d_K(t) = 1$ $i_K(t) = X_{K-1}^K$
$t = K + 1$	$d_1(t) = K + 1$ $i_1(t) = i_1(t - 1)$	$d_2(t) = 1$ $i_2(t) = X_K^2$	$\cdots$ $\cdots$	$d_K(t) = 2$ $i_K(t) = i_K(t - 1)$
$t = K + 2$	$d_1(t) = 1$ $i_1(t) = X_{K+1}^1$	$d_2(t) = 2$ $i_2(t) = i_2(t - 1)$	$\cdots$ $\cdots$	$d_K(t) = 3$ $i_K(t) = i_K(t - 1)$
$t = K + 3$				

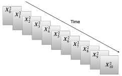
Delay: Time since last observation

LOS: Last Observed State

Understanding Arm Delays and Last Observed States

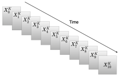


$\{X_t^1 : t = 0, 1, 2, \ldots\}$



$\{X_t^2 : t = 0, 1, 2, \ldots\}$

$\cdots\cdots$

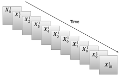


$\{X_t^K : t = 0, 1, 2, \ldots\}$

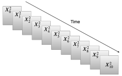
Time	**Delay** of arm 1 **LOS** of arm 1	**Delay** of arm 2 **LOS** of arm 2	$\cdots$ $\cdots$	**Delay** of arm K **LOS** of arm K
$t = K$	$d_1(t) = K$ $i_1(t) = X_0^1$	$d_2(t) = K - 1$ $i_2(t) = X_1^2$	$\cdots$ $\cdots$	$d_K(t) = 1$ $i_K(t) = X_{K-1}^K$
$t = K + 1$	$d_1(t) = K + 1$ $i_1(t) = i_1(t-1)$	$d_2(t) = 1$ $i_2(t) = X_K^2$	$\cdots$ $\cdots$	$d_K(t) = 2$ $i_K(t) = i_K(t-1)$
$t = K + 2$	$d_1(t) = 1$ $i_1(t) = X_{K+1}^1$	$d_2(t) = 2$ $i_2(t) = i_2(t-1)$	$\cdots$ $\cdots$	$d_K(t) = 3$ $i_K(t) = i_K(t-1)$
$t = K + 3$	$d_1(t) = 2$ $i_1(t) = i_i(t-1)$	$d_2(t) = 3$ $i_2(t) = i_2(t-1)$	$\cdots$ $\cdots$	$d_K(t) = 1$ $i_K(t) = X_{K+2}^K$
$t = K + 4$				

Delay: Time since last observation **LOS: Last Observed State**

Understanding Arm Delays and Last Observed States

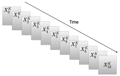


$\{X_t^1 : t = 0, 1, 2, \ldots\}$



$\{X_t^2 : t = 0, 1, 2, \ldots\}$

$\cdots\cdots$



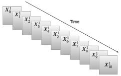
$\{X_t^K : t = 0, 1, 2, \ldots\}$

Time	**Delay** of arm 1 **LOS** of arm 1	**Delay** of arm 2 **LOS** of arm 2	$\cdots$ $\cdots$	**Delay** of arm K **LOS** of arm K
$t = K$	$d_1(t) = K$ $i_1(t) = X_0^1$	$d_2(t) = K - 1$ $i_2(t) = X_1^2$	$\cdots$ $\cdots$	$d_K(t) = 1$ $i_K(t) = X_{K-1}^K$
$t = K + 1$	$d_1(t) = K + 1$ $i_1(t) = i_1(t-1)$	$d_2(t) = 1$ $i_2(t) = X_K^2$	$\cdots$ $\cdots$	$d_K(t) = 2$ $i_K(t) = i_K(t-1)$
$t = K + 2$	$d_1(t) = 1$ $i_1(t) = X_{K+1}^1$	$d_2(t) = 2$ $i_2(t) = i_2(t-1)$	$\cdots$ $\cdots$	$d_K(t) = 3$ $i_K(t) = i_K(t-1)$
$t = K + 3$	$d_1(t) = 2$ $i_1(t) = i_i(t-1)$	$d_2(t) = 3$ $i_2(t) = i_2(t-1)$	$\cdots$ $\cdots$	$d_K(t) = 1$ $i_K(t) = X_{K+2}^K$
$t = K + 4$	$d_1(t) = 1$ $i_1(t) = X_{K+3}^1$	$d_2(t) = 4$ $i_2(t) = i_2(t-1)$	$\cdots$ $\cdots$	$d_K(t) = 2$ $i_K(t) = i_K(t-1)$
$t = K + 5$				

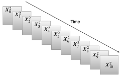
Delay: Time since last observation

LOS: Last Observed State

Understanding Arm Delays and Last Observed States

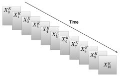


$\{X_t^1 : t = 0, 1, 2, \ldots\}$



$\{X_t^2 : t = 0, 1, 2, \ldots\}$

$\cdots\cdots$



$\{X_t^K : t = 0, 1, 2, \ldots\}$

Time	**Delay** of arm 1	**Delay** of arm 2	$\cdots$	**Delay** of arm K
	LOS of arm 1	**LOS** of arm 2	$\cdots$	**LOS** of arm K
$t = K$	$d_1(t) = K$ $i_1(t) = X_0^1$	$d_2(t) = K - 1$ $i_2(t) = X_1^2$	$\cdots$ $\cdots$	$d_K(t) = 1$ $i_K(t) = X_{K-1}^K$
$t = K + 1$	$d_1(t) = K + 1$ $i_1(t) = i_1(t - 1)$	$d_2(t) = 1$ $i_2(t) = X_K^2$	$\cdots$ $\cdots$	$d_K(t) = 2$ $i_K(t) = i_K(t - 1)$
$t = K + 2$	$d_1(t) = 1$ $i_1(t) = X_{K+1}^1$	$d_2(t) = 2$ $i_2(t) = i_2(t - 1)$	$\cdots$ $\cdots$	$d_K(t) = 3$ $i_K(t) = i_K(t - 1)$
$t = K + 3$	$d_1(t) = 2$ $i_1(t) = i_i(t - 1)$	$d_2(t) = 3$ $i_2(t) = i_2(t - 1)$	$\cdots$ $\cdots$	$d_K(t) = 1$ $i_K(t) = X_{K+2}^K$
$t = K + 4$	$d_1(t) = 1$ $i_1(t) = X_{K+3}^1$	$d_2(t) = 4$ $i_2(t) = i_2(t - 1)$	$\cdots$ $\cdots$	$d_K(t) = 2$ $i_K(t) = i_K(t - 1)$
$t = K + 5$	$d_1(t) = 2$ $i_1(t) = i_1(t - 1)$	$d_2(t) = 1$ $i_2(t) = X_{K+4}^2$	$\cdots$ $\cdots$	$d_K(t) = 3$ $i_K(t) = i_K(t - 1)$

Delay: Time since last observation **LOS: Last Observed State**

A New Notion of State

$$\underbrace{\underline{d}(t) = (d_1(t), \ldots, d_K(t))}_{\text{arm delays}}, \qquad \underbrace{\underline{i}(t) = (i_1(t), \ldots, i_K(t))}_{\text{last observed states of the arms}}$$

A New Notion of State

$$\underbrace{\underline{d}(t) = (d_1(t), \ldots, d_K(t))}_{\text{arm delays}}, \qquad \underbrace{\underline{i}(t) = (i_1(t), \ldots, i_K(t))}_{\text{last observed states of the arms}}$$

$$(B_0, A_0, X_0^{A_0}, B_1, A_1, X_1^{A_1}, \ldots, B_{t-1}, A_{t-1}, X_{t-1}^{A_{t-1}}) \equiv \{B_s,\ (\underline{d}(s), \underline{i}(s)) : K \leq s \leq t-1, \\ (\underline{d}(t), \underline{i}(t))\}$$

A New Notion of State

$$\underbrace{\underline{d}(t) = (d_1(t), \ldots, d_K(t))}_{\text{arm delays}}, \qquad \underbrace{\underline{i}(t) = (i_1(t), \ldots, i_K(t))}_{\text{last observed states of the arms}}$$

$$(B_0, A_0, X_0^{A_0}, B_1, A_1, X_1^{A_1}, \ldots, B_{t-1}, A_{t-1}, X_{t-1}^{A_{t-1}}) \equiv \{B_s,\ (\underline{d}(s), \underline{i}(s)) : K \leq s \leq t-1,$$
$$(\underline{d}(t), \underline{i}(t))\}$$

An interplay of the various variables:

$$\{B_s,\ (\underline{d}(s),\ \underline{i}(s)) : K \leq s \leq t-1, \longrightarrow B_t \xrightarrow{\textbf{TH}} (A_t,\ X_t^{A_t}) \longrightarrow (\underline{d}(t+1),\ \underline{i}(t+1))$$
$$(\underline{d}(t), \underline{i}(t))\}$$

A Controlled Markov Process

$$
\begin{aligned}
&P(\underline{d}(t+1),\ \underline{i}(t+1) \mid \{B_s,\ (\underline{d}(s),\ \underline{i}(s)) : K \le s \le t-1\},\ B_t,\ (\underline{d}(t), \underline{i}(t))) \\
&= P(\underline{d}(t+1),\ \underline{i}(t+1) \mid B_t,\ (\underline{d}(t),\ \underline{i}(t))) \qquad (1)
\end{aligned}
$$

A Controlled Markov Process

$$
\begin{aligned}
&P(\underline{d}(t+1),\ \underline{i}(t+1) \mid \{B_s,\ (\underline{d}(s),\ \underline{i}(s)) : K \leq s \leq t-1\},\ B_t,\ (\underline{d}(t), \underline{i}(t))) \\
&= P(\underline{d}(t+1),\ \underline{i}(t+1) \mid B_t,\ (\underline{d}(t),\ \underline{i}(t)))
\end{aligned} \tag{1}
$$

We have a Markov decision problem with

State space	Set of all possible $(\underline{d},\ \underline{i})$ values
Action space	Set of arms
State at time t	$(\underline{d}(t),\ \underline{i}(t))$
Action at time t	B_t
Observation at time t	$(A_t,\ X_t^{A_t})$
Transition probabilities	As in (1)

Policies

- A policy π prescribes one of the following two actions at each time t:
  - $\{B_s,\ (\underline{d}(s), \underline{i}(s)) : K \leq s \leq t-1,\ (\underline{d}(t), \underline{i}(t))\} \mapsto B_t$
  - stop and declare the odd arm

Policies

- A policy π prescribes one of the following two actions at each time t:
  - $\{B_s,\ (\underline{d}(s), \underline{i}(s)) : K \leq s \leq t-1,\ (\underline{d}(t), \underline{i}(t))\} \mapsto B_t$
  - stop and declare the odd arm

- **SRS policy:** B_t depends only on $(\underline{d}(t), \underline{i}(t))$ for each t, and is chosen according to the randomised rule

$$P(B_t = a \mid \{B_s,\ (\underline{d}(s), \underline{i}(s)) : K \leq s \leq t-1\},\ (\underline{d}(t), \underline{i}(t))) = \lambda(a \mid (\underline{d}(t), \underline{i}(t)))$$

  for some $\lambda(\cdot \mid \cdot)$ that is stationary across time

Policies

- A policy π prescribes one of the following two actions at each time t:
  - $\{B_s,\ (\underline{d}(s), \underline{i}(s)) : K \leq s \leq t-1,\ (\underline{d}(t), \underline{i}(t))\} \mapsto B_t$
  - stop and declare the odd arm

- **SRS policy:** B_t depends only on $(\underline{d}(t), \underline{i}(t))$ for each t, and is chosen according to the randomised rule

$$P(B_t = a \mid \{B_s,\ (\underline{d}(s), \underline{i}(s)) : K \leq s \leq t-1\},\ (\underline{d}(t), \underline{i}(t))) = \lambda(a \mid (\underline{d}(t), \underline{i}(t)))$$

  for some $\lambda(\cdot \mid \cdot)$ that is stationary across time

- Denote an SRS policy associated with $\lambda(\cdot \mid \cdot)$ by π^λ. Let Π_{SRS} be the set of all SRS policies

---

SRS: stationary randomised strategy. The terminology is from Borkar [5].

Ergodicity and the Lower Bound

SRS Policies + Trembling Hand = Ergodicity

A Key Ergodicity Property

Under any $\pi^\lambda \in \Pi_{\mathrm{SRS}}$, the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is a Markov process. Further, this Markov process is ergodic. A unique stationary distribution, call it μ^λ, therefore exists under π^λ.

The proof relies on the hypothesis that the trembling hand parameter $\eta > 0$

SRS Policies + Trembling Hand = Ergodicity

A Key Ergodicity Property

Under any $\pi^\lambda \in \Pi_{\mathsf{SRS}}$, the process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ is a Markov process. Further, this Markov process is ergodic. A unique stationary distribution, call it μ^λ, therefore exists under π^λ.

The proof relies on the hypothesis that the trembling hand parameter $\eta > 0$

Ergodic state action occupancy measure:

$$\nu^\lambda(\underline{d}, \underline{i}, a) = \mu^\lambda(\underline{d}, \underline{i}) \left(\frac{\eta}{K} + (1-\eta)\,\lambda(a \mid \underline{d}, \underline{i}) \right)$$

Lower Bound - 1

- Fix the following quantities:
  - Odd arm location h
  - P_1: TPM of arm h
  - P_2: TPM of arm h' for all $h' \neq h$
  - Error tolerance $\epsilon > 0$

Lower Bound - 1

- Fix the following quantities:
  - Odd arm location h
  - P_1: TPM of arm h
  - P_2: TPM of arm h' for all $h' \neq h$
  - Error tolerance $\epsilon > 0$

$$\Pi(\epsilon) = \{\pi : \text{ Prob. of erroneously declaring the odd arm under } \pi \ \leq \epsilon\}$$

$$P_h^a = \text{ TPM of arm } a \text{ when } h \text{ is the odd arm}$$

$$= \begin{cases} P_1, & a = h, \\ P_2, & a \neq h \end{cases}, \qquad \tau(\pi) = \text{stopping time of policy } \pi$$

Lower Bound - 1

- Fix the following quantities:
  - Odd arm location h
  - P_1: TPM of arm h
  - P_2: TPM of arm h' for all $h' \neq h$
  - Error tolerance $\epsilon > 0$

$$\Pi(\epsilon) = \{\pi : \text{ Prob. of erroneously declaring the odd arm under } \pi \ \leq \epsilon\}$$

$$P_h^a = \text{ TPM of arm } a \text{ when } h \text{ is the odd arm}$$

$$= \begin{cases} P_1, & a = h, \\ P_2, & a \neq h \end{cases}, \qquad \tau(\pi) = \text{stopping time of policy } \pi$$

- For $d \geq 1$, let

$$(P_h^a)^d = d\text{th power of } P_h^a$$

$$(P_h^a)^d(\cdot|i) = i\text{th row of } (P_h^a)^d, \quad i \in \mathcal{S}$$

Lower Bound - 2

Lower Bound: Odd Restless Markov Arm Problem

$$\liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \geq \frac{1}{R^*(P_1, P_2)}$$

where

$$R^*(P_1, P_2) = \sup_{\pi^\lambda \in \Pi_{\mathrm{SRS}}} \min_{h' \neq h} \sum_{(\underline{d}, \underline{i})} \sum_{a=1}^{K} \nu^\lambda(\underline{d}, \underline{i}, a) \underbrace{D((P_h^a)^{d_a}(\cdot|i_a) \| (P_{h'}^a)^{d_a}(\cdot|i_a))}_{\text{Kullback-Leibler divergence}}$$

Lower Bound - 2

Lower Bound: Odd Restless Markov Arm Problem

$$\liminf_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log(1/\epsilon)} \geq \frac{1}{R^*(P_1, P_2)}$$

where

$$R^*(P_1, P_2) = \sup_{\pi^\lambda \in \Pi_{\mathrm{SRS}}} \min_{h' \neq h} \sum_{(\underline{d}, \underline{i})} \sum_{a=1}^{K} \nu^\lambda(\underline{d}, \underline{i}, a) \underbrace{D((P_h^a)^{d_a}(\cdot|i_a) \| (P_{h'}^a)^{d_a}(\cdot|i_a))}_{\text{Kullback-Leibler divergence}}$$

Remarks:

- $R^*(P_1, P_2)$ does not depend on h, the location of odd arm
- The LHS of the lower bound contains *all* policies, whereas the RHS contains *only* SRS policies. This is due to [6, Theorem 8.8.2]
- Computability of $R^*(P_1, P_2)$: Q-learning for restless arms [7]

Upper Bound

Preliminaries

- The expression for $R^*(P_1, P_2)$ has a sup
- It is not clear if this sup is achievable by some SRS policy

Preliminaries

- The expression for $R^*(P_1, P_2)$ has a sup
- It is not clear if this sup is achievable by some SRS policy
- However, the sup may be approached arbitrarily closely:

$$\forall\ \delta > 0,\ \exists\ \lambda_{h,\delta}(\cdot \mid \cdot) \text{ s.t.}$$

$$\min_{h' \neq h} \sum_{(\underline{d},\underline{i})} \sum_{a=1}^{K} \nu^{\lambda_{h,\delta}}(\underline{d}, \underline{i}, a)\ D((P_h^a)^{d_a}(\cdot|i_a) \| (P_{h'}^a)^{d_a}(\cdot|i_a)) > \frac{R^*(P_1, P_2)}{1+\delta}$$

Policy $\pi^\star(L, \delta)$

- Input: Two parameters $L > 1$ and $\delta > 0$

Policy $\pi^\star(L, \delta)$

- Input: Two parameters $L > 1$ and $\delta > 0$
- Select arm 1 at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K - 1$

Policy $\pi^\star(L, \delta)$

- Input: Two parameters $L > 1$ and $\delta > 0$
- Select arm 1 at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K - 1$
- For all $t \geq K$:
  - Maintain guess of odd arm:

$$\hat{\theta}(t) \in \arg\max_{h} \underbrace{\min_{h' \neq h} \log \frac{P_h(B_0, A_0, X_0^{A_0}, \ldots, B_t, A_t, X_t^{A_t})}{P_{h'}(B_0, A_0, X_0^{A_0}, \ldots, B_t, A_t, X_t^{A_t})}}_{M_h(t)}$$

Policy $\pi^\star(L, \delta)$

- Input: Two parameters $L > 1$ and $\delta > 0$
- Select arm 1 at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K - 1$
- For all $t \geq K$:
  - Maintain guess of odd arm:

$$\hat{\theta}(t) \in \arg\max_{h} \underbrace{\min_{h' \neq h} \log \frac{P_h(B_0, A_0, X_0^{A_0}, \ldots, B_t, A_t, X_t^{A_t})}{P_{h'}(B_0, A_0, X_0^{A_0}, \ldots, B_t, A_t, X_t^{A_t})}}_{M_h(t)}$$

  - If $M_{\hat{\theta}(t)}(t) \geq \log((K-1)L)$, stop and declare $\hat{\theta}(t)$ is the odd arm

Policy $\pi^\star(L, \delta)$

- Input: Two parameters $L > 1$ and $\delta > 0$
- Select arm 1 at time $t = 0$, arm 2 at time $t = 1$ and so on until arm K at time $t = K - 1$
- For all $t \geq K$:
  - Maintain guess of odd arm:

$$\hat{\theta}(t) \in \arg\max_{h} \underbrace{\min_{h' \neq h} \log \frac{P_h(B_0, A_0, X_0^{A_0}, \ldots, B_t, A_t, X_t^{A_t})}{P_{h'}(B_0, A_0, X_0^{A_0}, \ldots, B_t, A_t, X_t^{A_t})}}_{M_h(t)}$$

  - If $M_{\hat{\theta}(t)}(t) \geq \log((K-1)L)$, stop and declare $\hat{\theta}(t)$ is the odd arm
  - If $M_{\hat{\theta}(t)}(t) < \log((K-1)L)$, select next arm according to $\lambda_{\hat{\theta}(t),\delta}(\cdot \mid \cdot)$

Achievability: Results

- Policy $\pi^\star(L, \delta)$ is *not* an SRS policy
- Policy $\pi^\star(L, \delta)$ stops in finite time w.p. 1
- If $L = 1/\epsilon$, then $\pi^\star(L, \delta) \in \Pi(\epsilon)$ for all $\delta > 0$ (desired error probability)
- **Upper bound:** for $\pi = \pi^\star(L, \delta)$,

$$\limsup_{L \to \infty} \frac{E_h[\tau(\pi)]}{\log L} \leq \frac{1+\delta}{R^*(P_1, P_2)}$$

Achievability: Results

- Policy $\pi^\star(L, \delta)$ is *not* an SRS policy
- Policy $\pi^\star(L, \delta)$ stops in finite time w.p. 1
- If $L = 1/\epsilon$, then $\pi^\star(L, \delta) \in \Pi(\epsilon)$ for all $\delta > 0$ (desired error probability)
- **Upper bound:** for $\pi = \pi^\star(L, \delta)$,

$$\limsup_{L \to \infty} \frac{E_h[\tau(\pi)]}{\log L} \leq \frac{1 + \delta}{R^*(P_1, P_2)}$$

- Stitching together the solutions for various δ, we get

$$\boxed{\limsup_{\delta \downarrow 0} \limsup_{L \to \infty} \frac{E_h[\tau(\pi)]}{\log L} \leq \frac{1}{R^*(P_1, P_2)}}$$

Main Result

Main Result

Main Result: Odd Restless Markov Arm Problem

For the problem of odd arm identification with restless Markov arms in which h is the odd arm, P_1 is the TPM of arm h and P_2 is the common TPM of all arms other than h, where $P_2 \neq P_1$,

$$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E_h[\tau(\pi)]}{\log \frac{1}{\epsilon}} = \frac{1}{R^*(P_1, P_2)}.$$

Conclusions

Concluding Remarks

- Ergodicity of the Markov process $\{(\underline{d}(t), \underline{i}(t)) : t \geq K\}$ under any SRS policy was key to deriving the lower and the upper bounds
- The trembling hand model may be viewed as a regularisation that gives ergodicity of the aforementioned Markov chain for free. When the trembling hand parameter $\eta = 0$, there may be a gap between the resulting upper and lower bounds.
  An analysis of the case $\eta = 0$ may be found in our supplementary manuscript [8]
- Restless arms: $\lambda(\cdot \mid \cdot)$
  IID and rested arms: $\lambda(\cdot)$

Concluding Remarks

- Ergodicity of the Markov process $\{(\underline{d}(t),\ \underline{i}(t)) : t \geq K\}$ under any SRS policy was key to deriving the lower and the upper bounds
- The trembling hand model may be viewed as a regularisation that gives ergodicity of the aforementioned Markov chain for free. When the trembling hand parameter $\eta = 0$, there may be a gap between the resulting upper and lower bounds.
  An analysis of the case $\eta = 0$ may be found in our supplementary manuscript [8]
- Restless arms: $\lambda(\cdot \mid \cdot)$
  IID and rested arms: $\lambda(\cdot)$
- Future work: a study of the case when the transition matrices P_1 and P_2 are not known

References i

- N. K. Vaidhiyan, S. Arun, and R. Sundaresan, “Neural dissimilarity indices that predict oddball detection in behaviour,” *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 4778–4796, 2017.
- N. K. Vaidhiyan and R. Sundaresan, “Learning to detect an oddball target,” *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 831–852, 2017.
- P. Whittle, “Restless bandits: Activity allocation in a changing world,” *Journal of applied probability*, vol. 25, no. A, pp. 287–298, 1988.
- P. N. Karthik and R. Sundaresan, “Learning to Detect an Odd Markov Arm,” 2019. [Online]. Available: https://arxiv.org/abs/1904.11361

- V. S. Borkar, "Control of markov chains with long-run average cost criterion," in *Stochastic Differential Systems, Stochastic Control Theory and Applications*. Springer, 1988, pp. 57–77.
- M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- K. Avrachenkov and V. S. Borkar, "Whittle index based q-learning for restless bandits with average reward," *arXiv preprint arXiv:2004.14427*, 2020.
- P. N. Karthik and R. Sundaresan, "Detecting an odd restless markov arm with a trembling hand (full version)," 2020. [Online]. Available: http://arxiv.org/abs/2005.06255

Acknowledgments

- The Science and Engineering Research Board, Department of Science and Technology (grant no. EMR/2016/002503)
- The Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science
- The Centre for Networked Intelligence, Indian Institute of Science

Thank You!