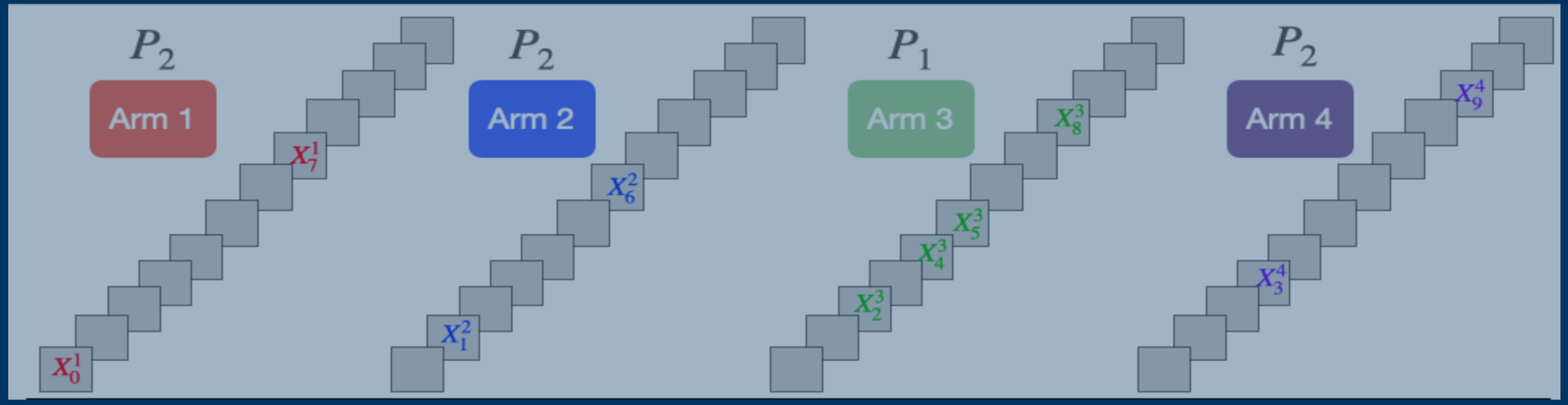


Learning to Detect an Odd Restless Markov Arm

2021 IEEE International Symposium on Information Theory

P. N. Karthik and Rajesh Sundaresan, Indian Institute of Science

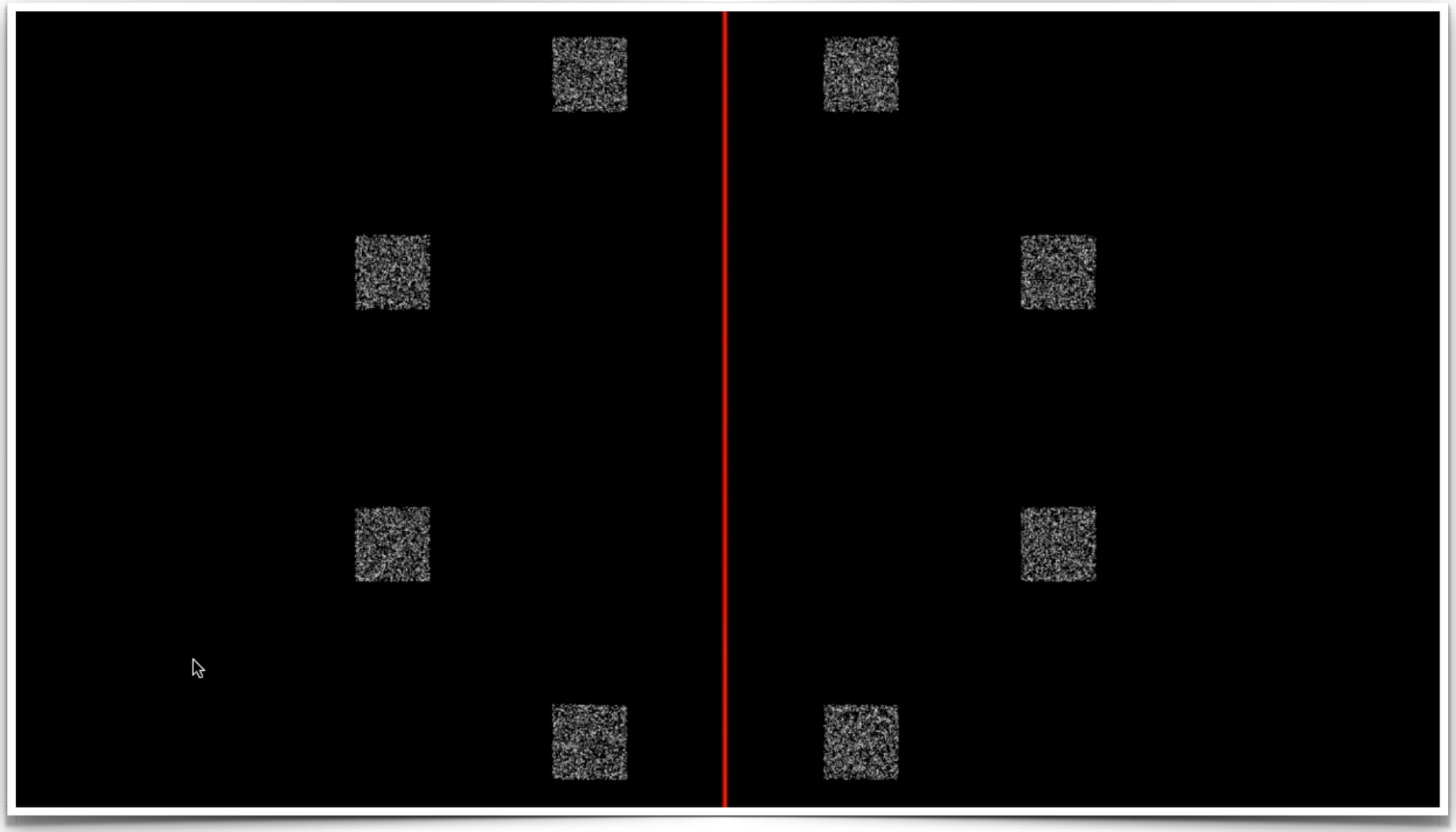
विज्ञान एवं प्रौद्योगिकी विभाग
DEPARTMENT OF
SCIENCE & TECHNOLOGY
सत्यमेव जयते

iNSERB
INDIA

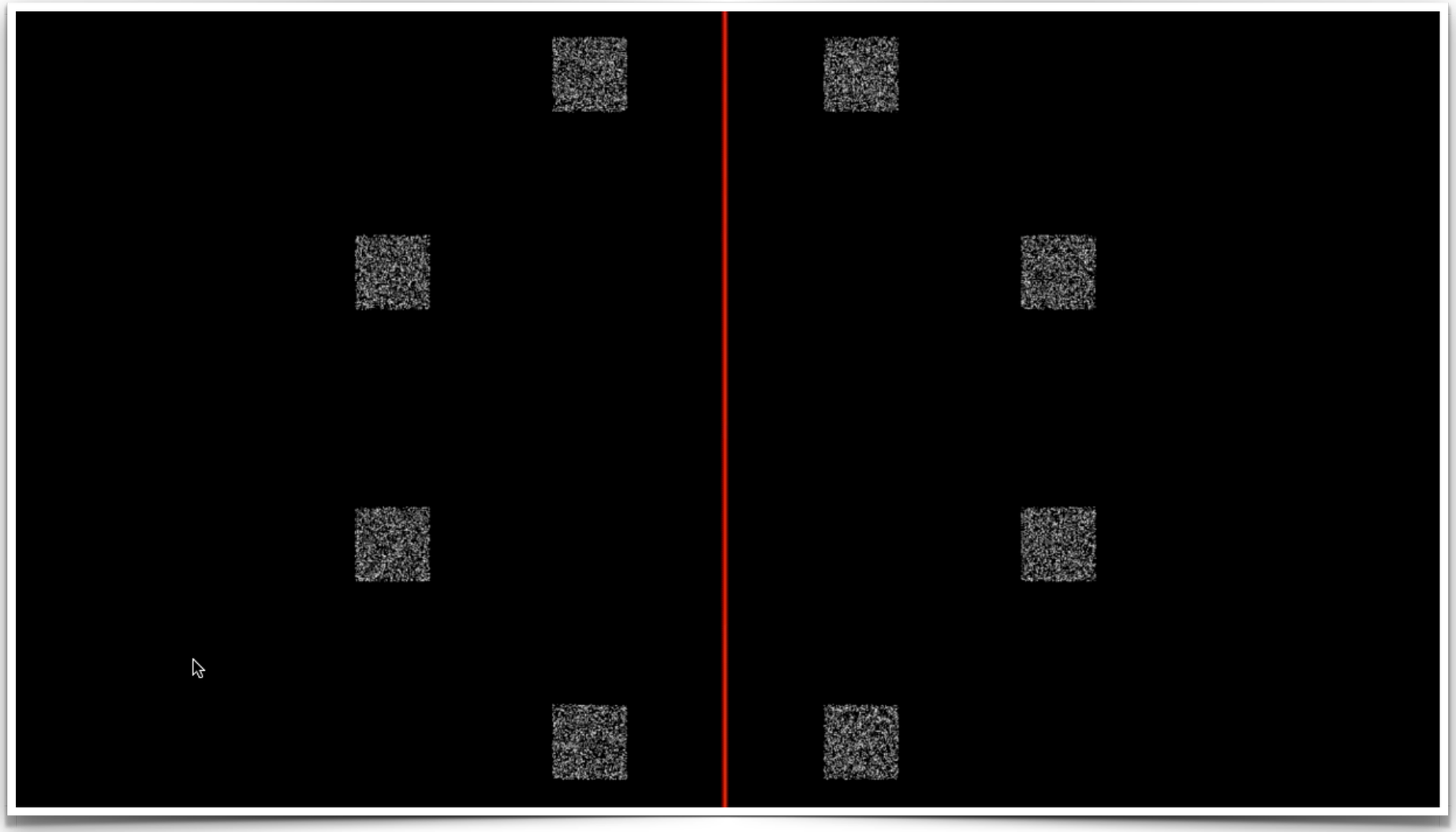
Motivation

Visual Search Experiments, Multi-Armed Bandits

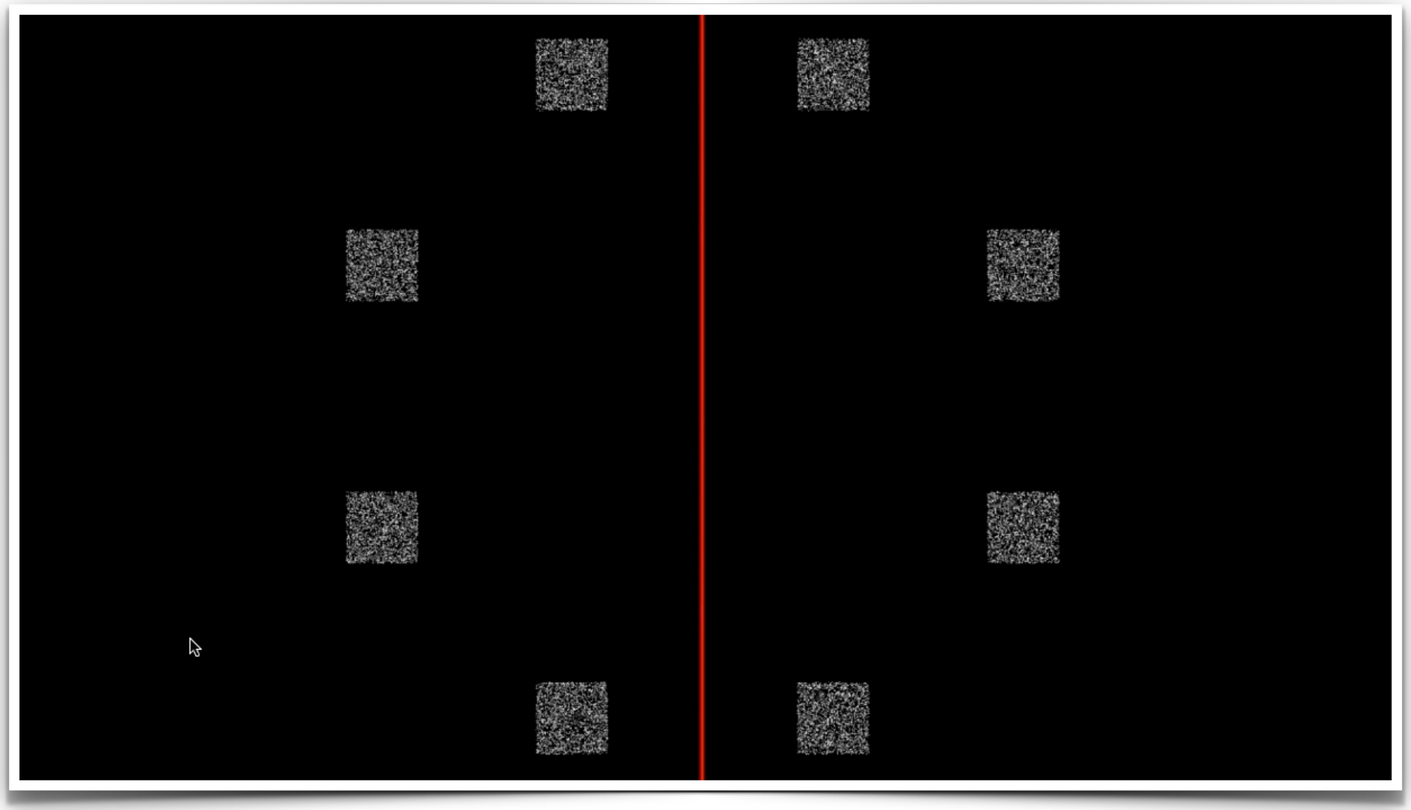
Identify the Odd Movie - 1



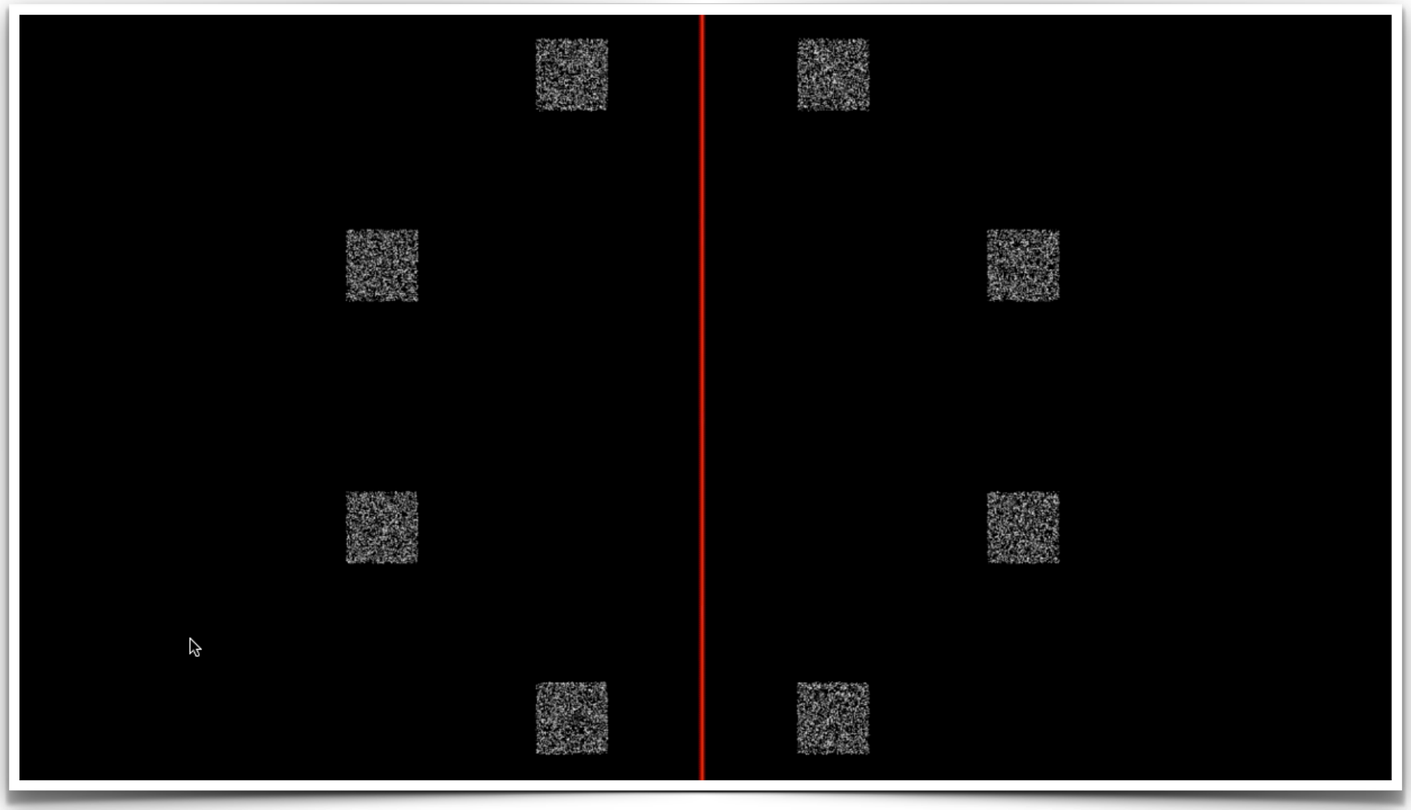
Identify the Odd Movie - 1



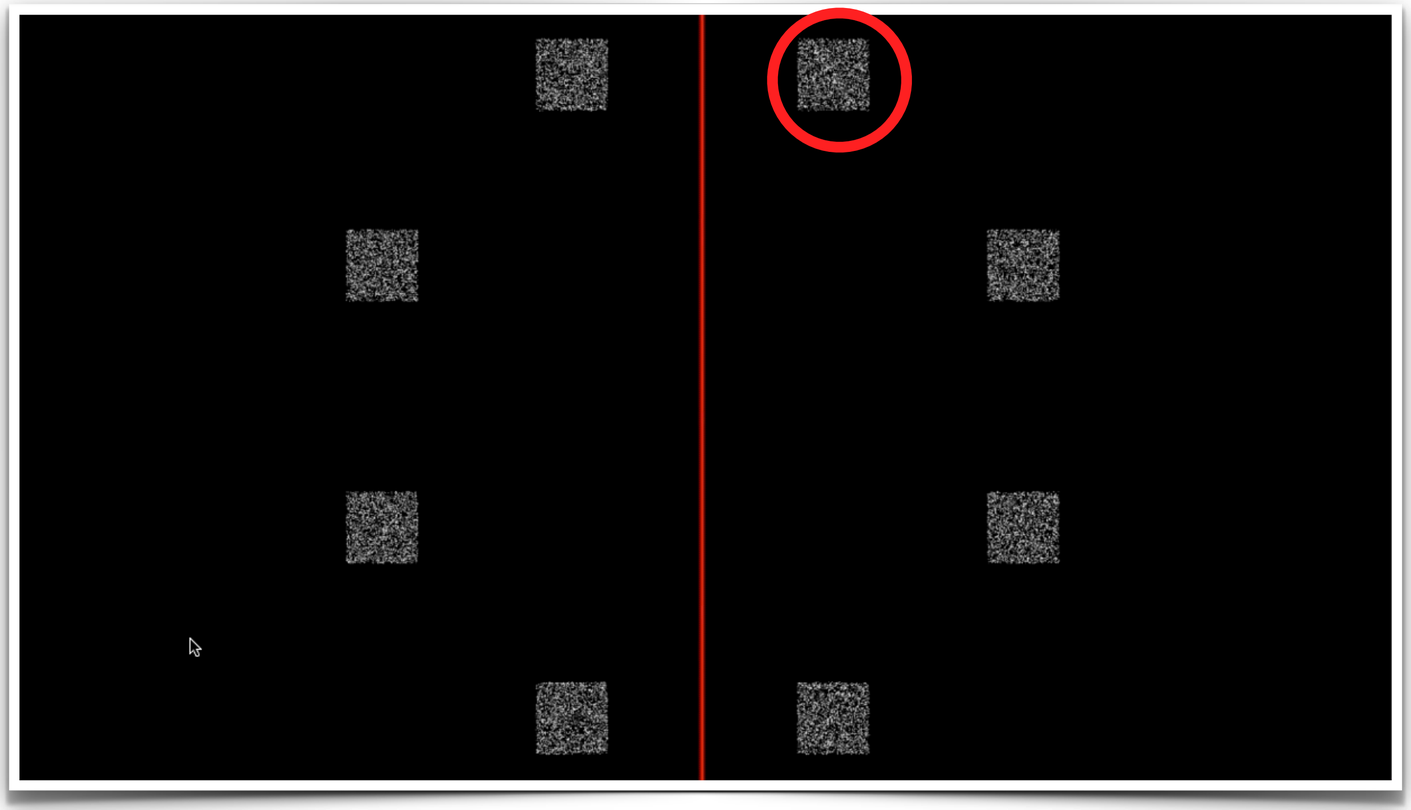
Identify the Odd Movie - 1



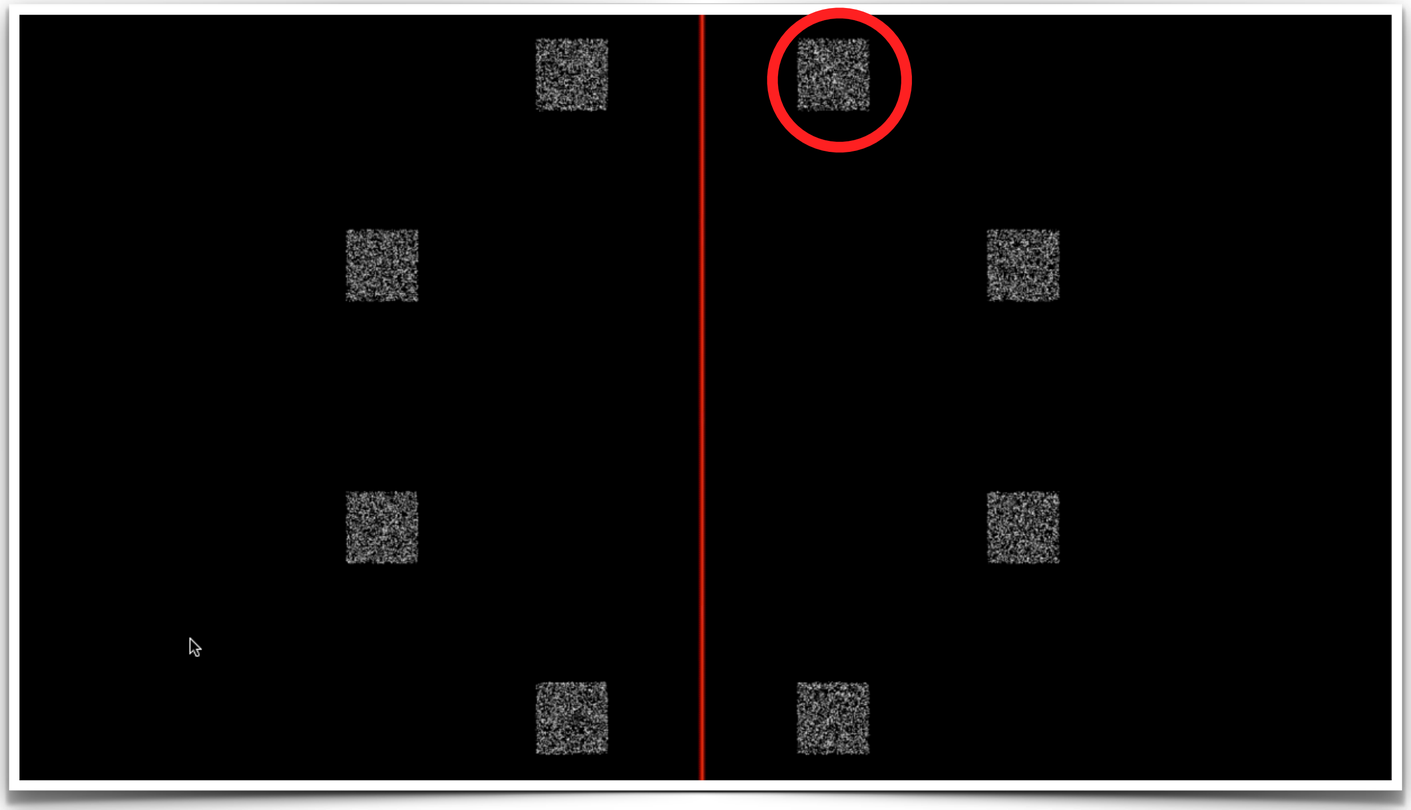
Identify the Odd Movie - 1



Identify the Odd Movie - 1

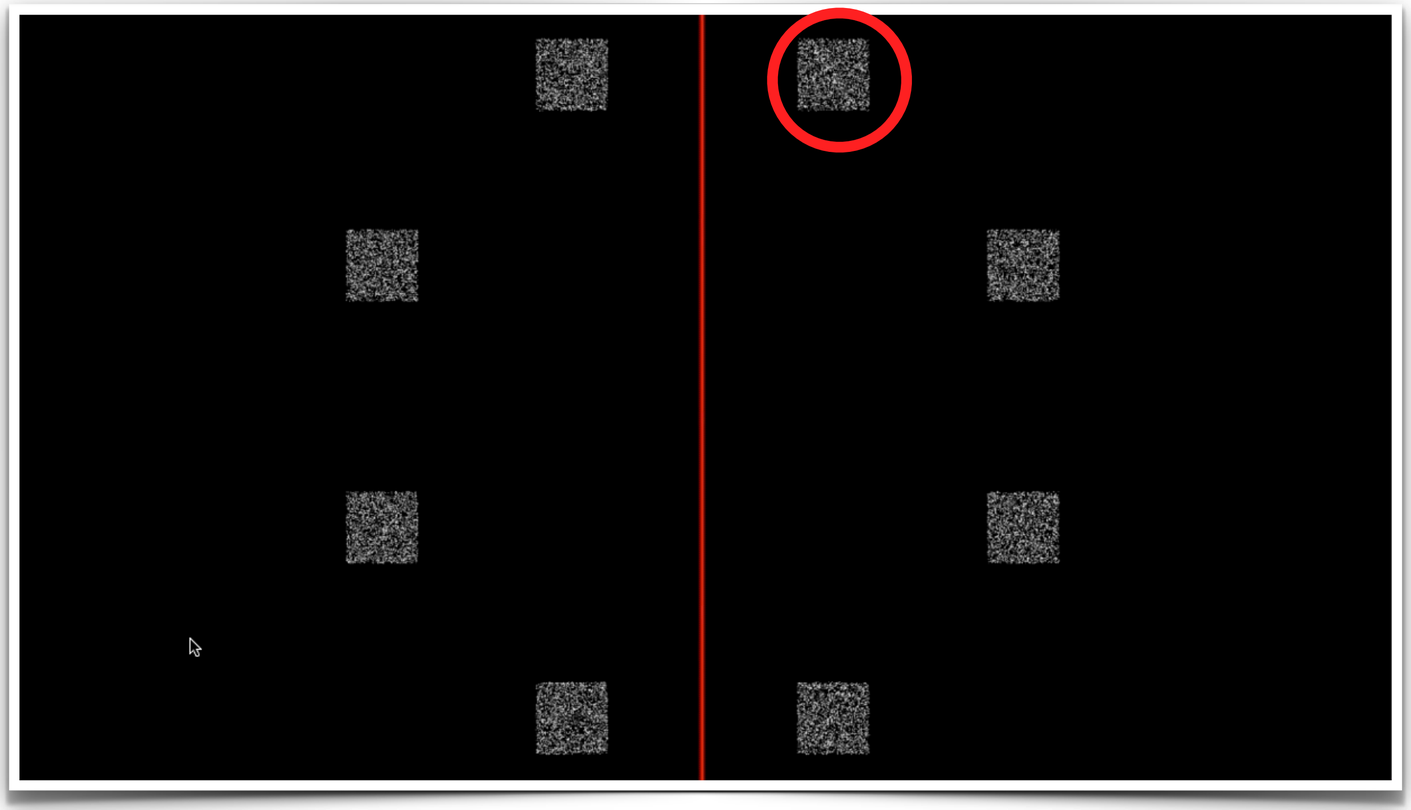


Identify the Odd Movie - 1

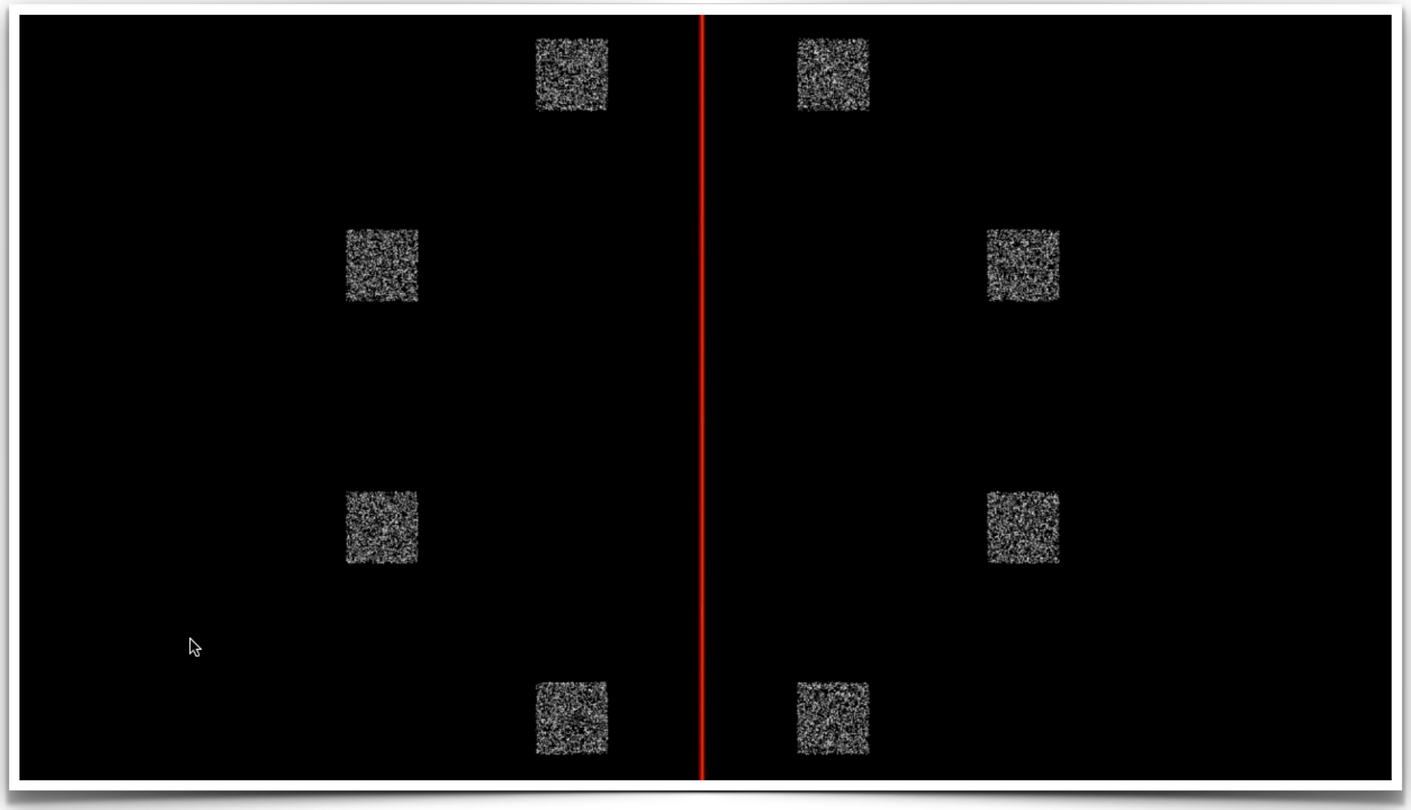


Odd movie

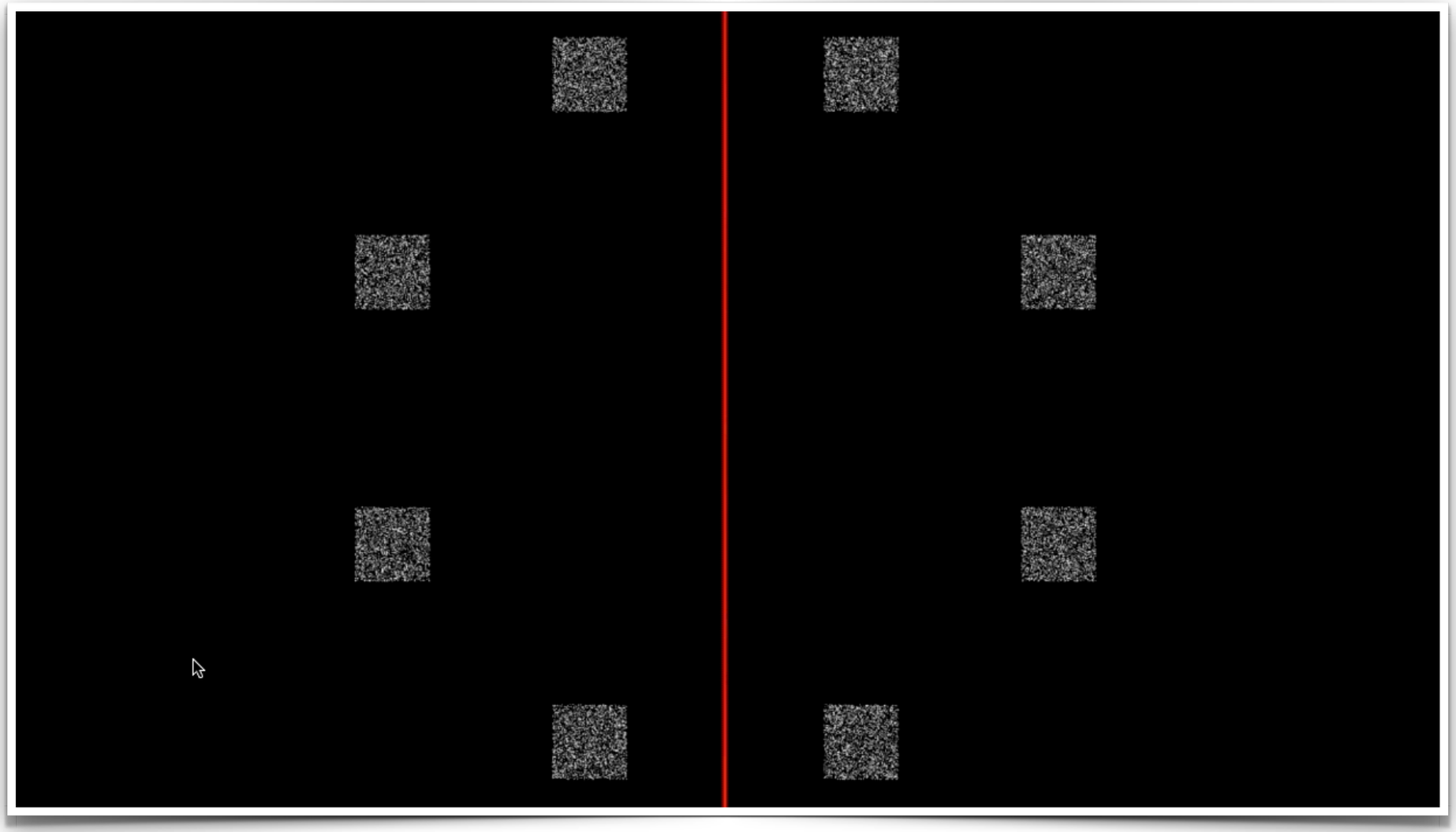
Identify the Odd Movie - 1



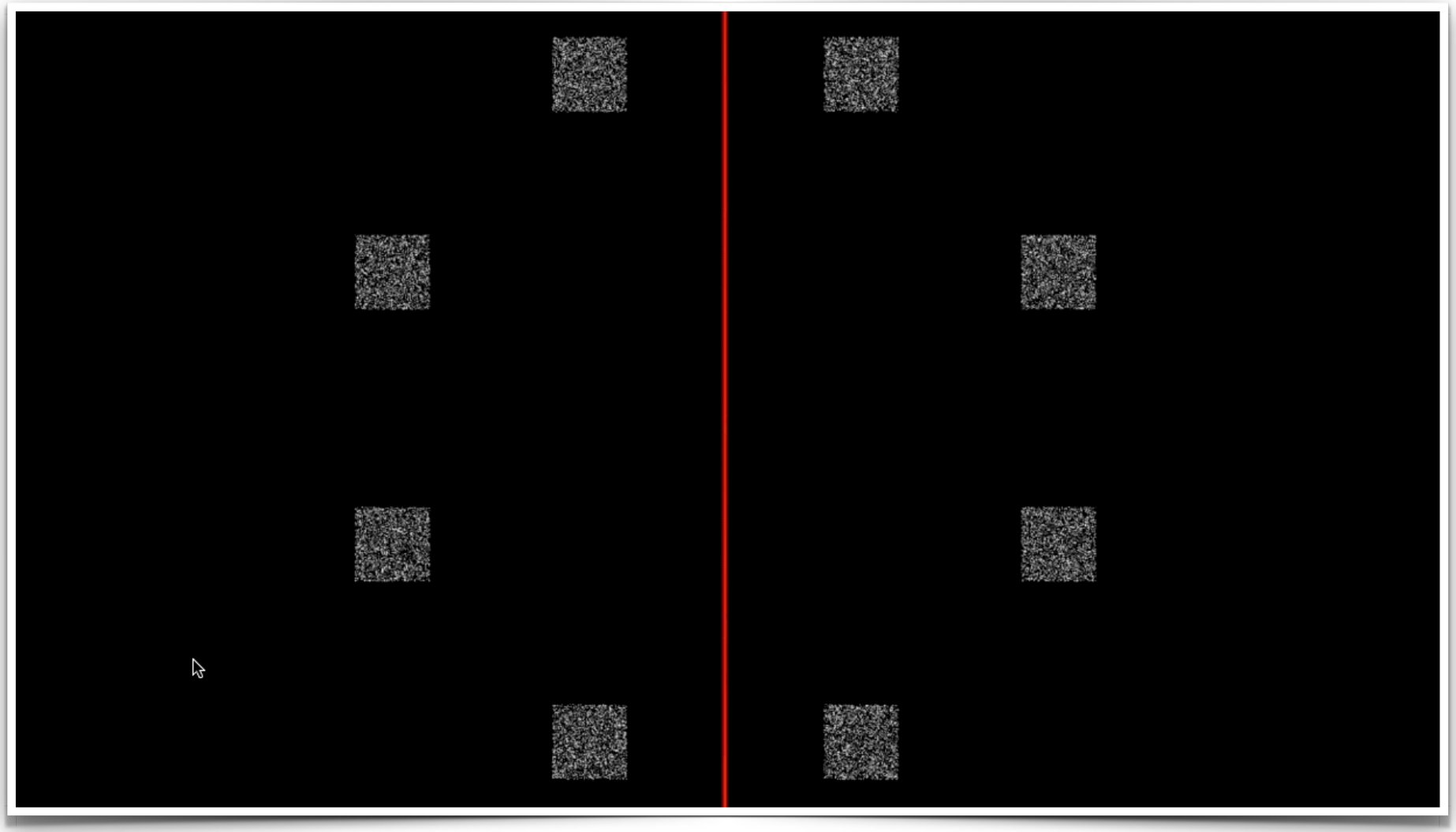
Identify the Odd Movie - 1



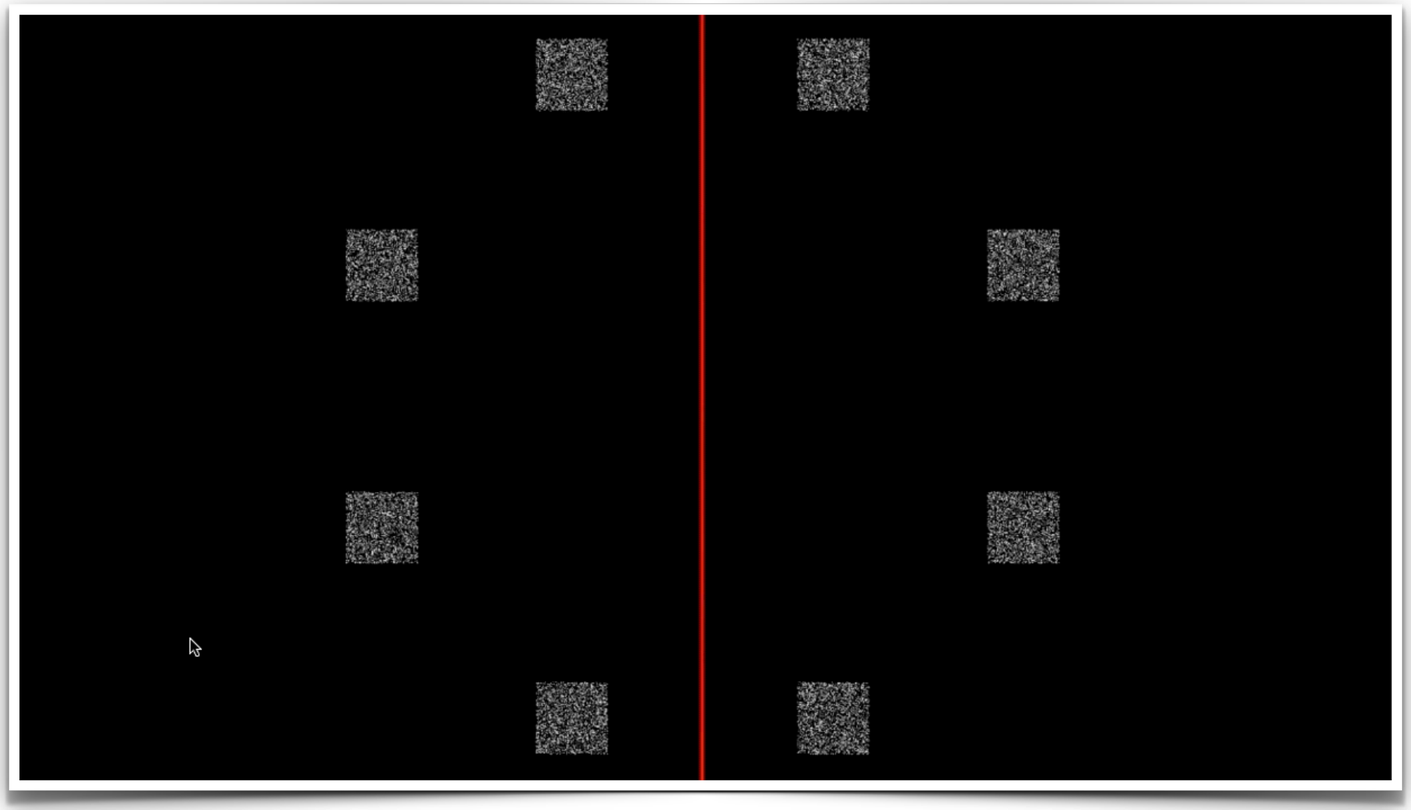
Identify the Odd Movie - 2



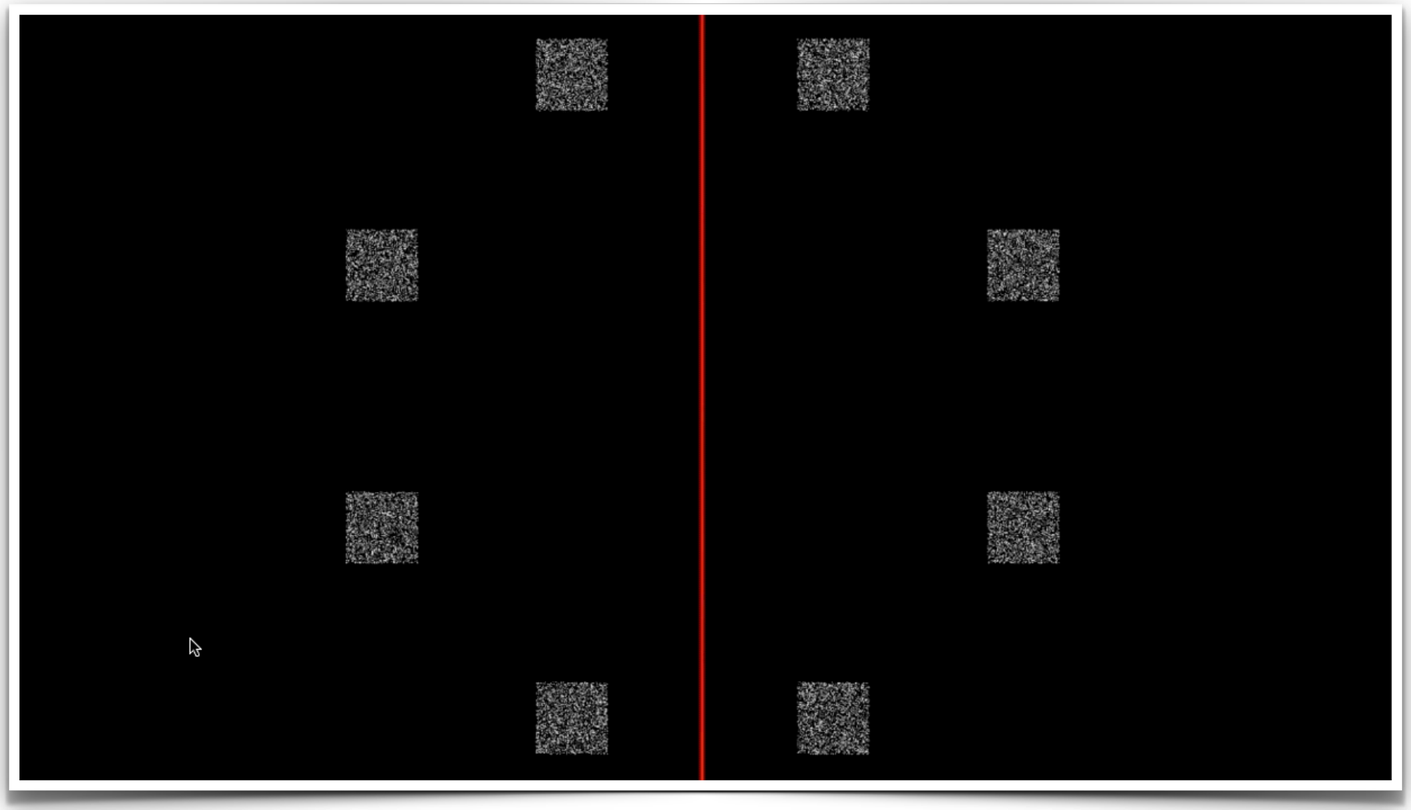
Identify the Odd Movie - 2



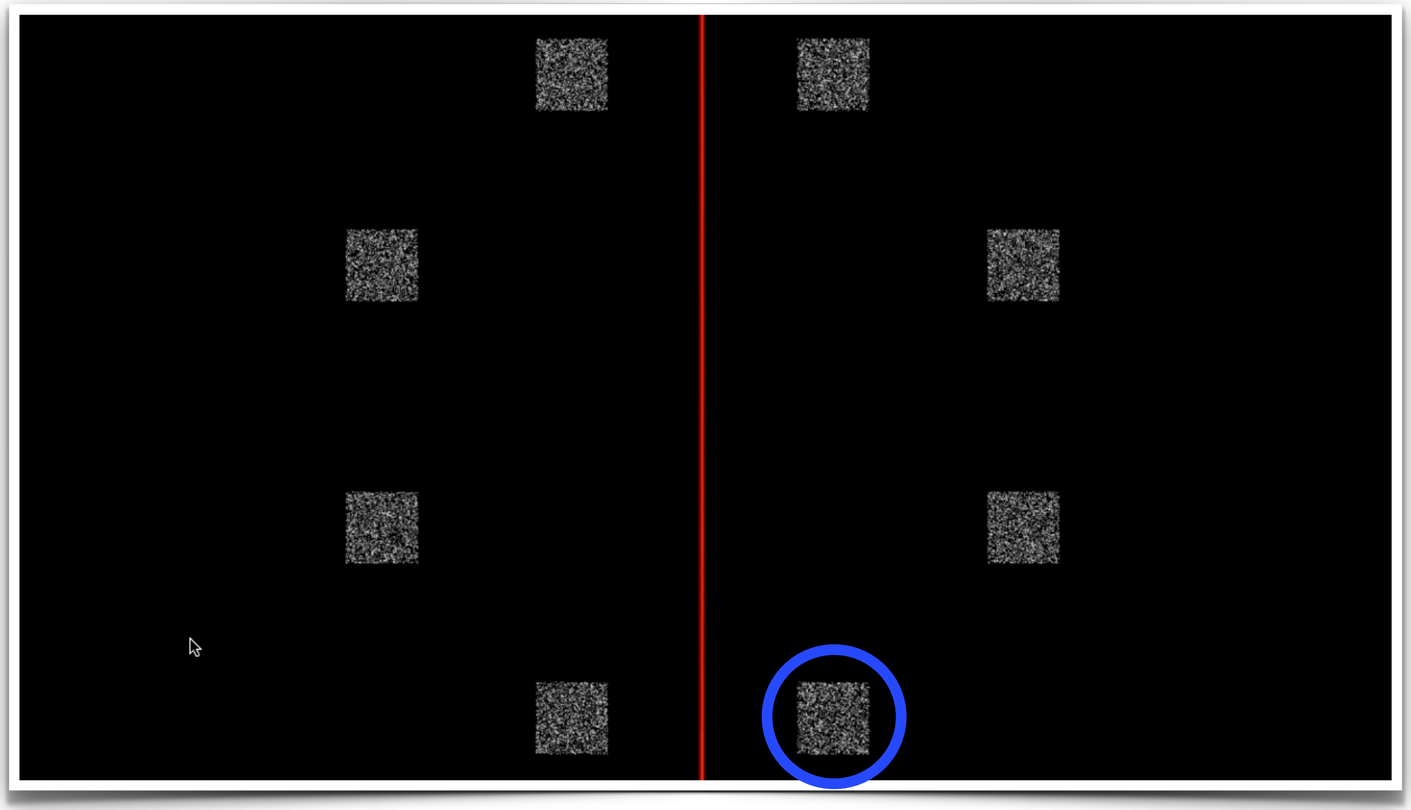
Identify the Odd Movie - 2



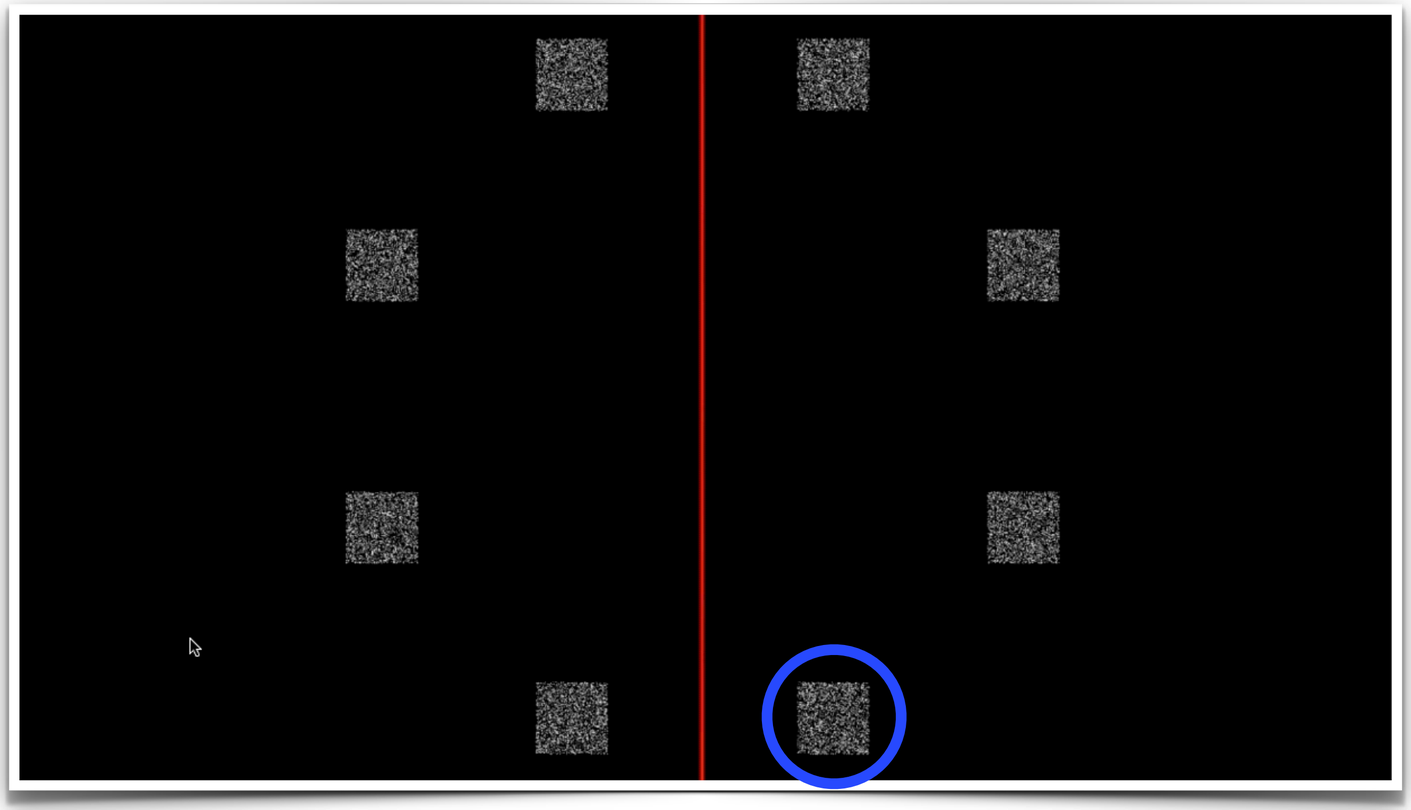
Identify the Odd Movie - 2



Identify the Odd Movie - 2

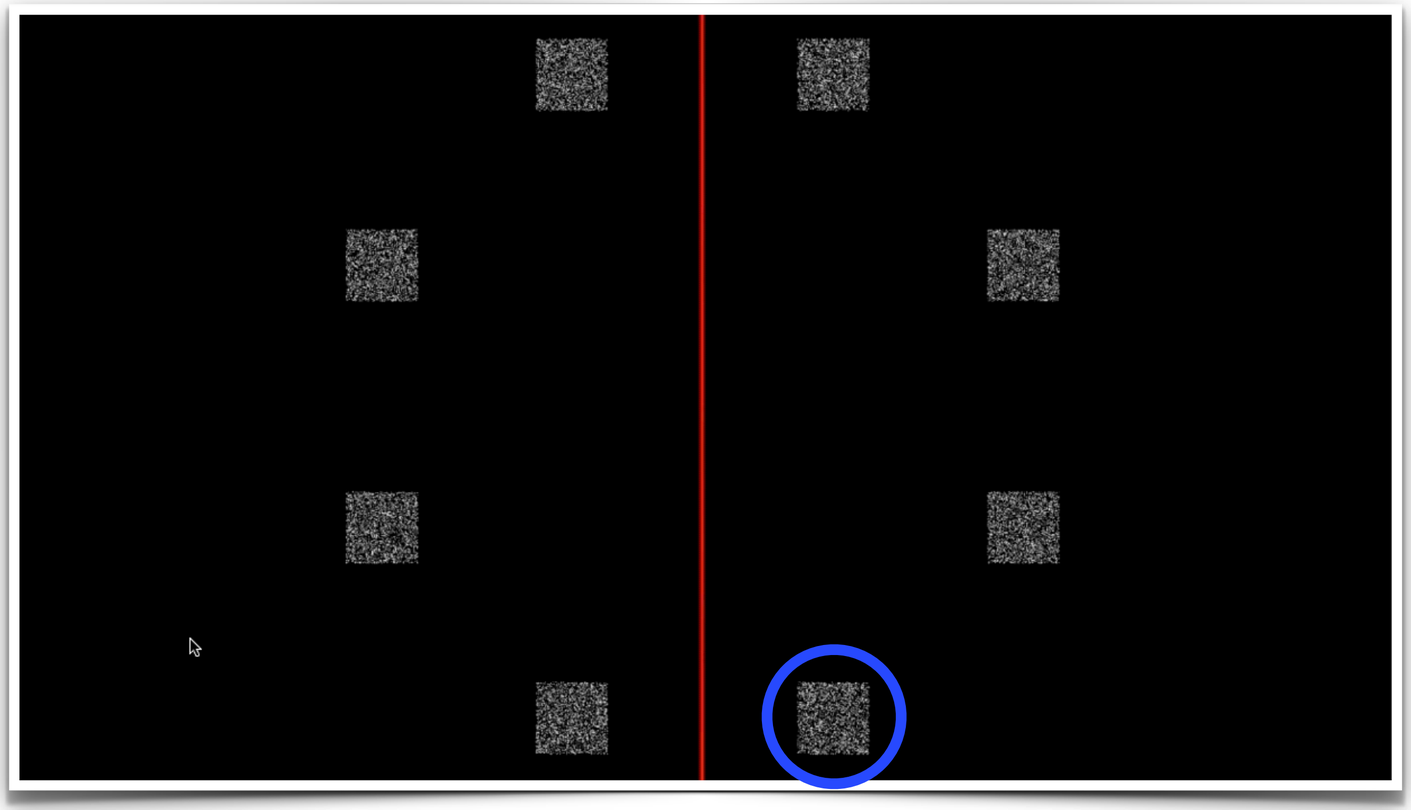


Identify the Odd Movie - 2

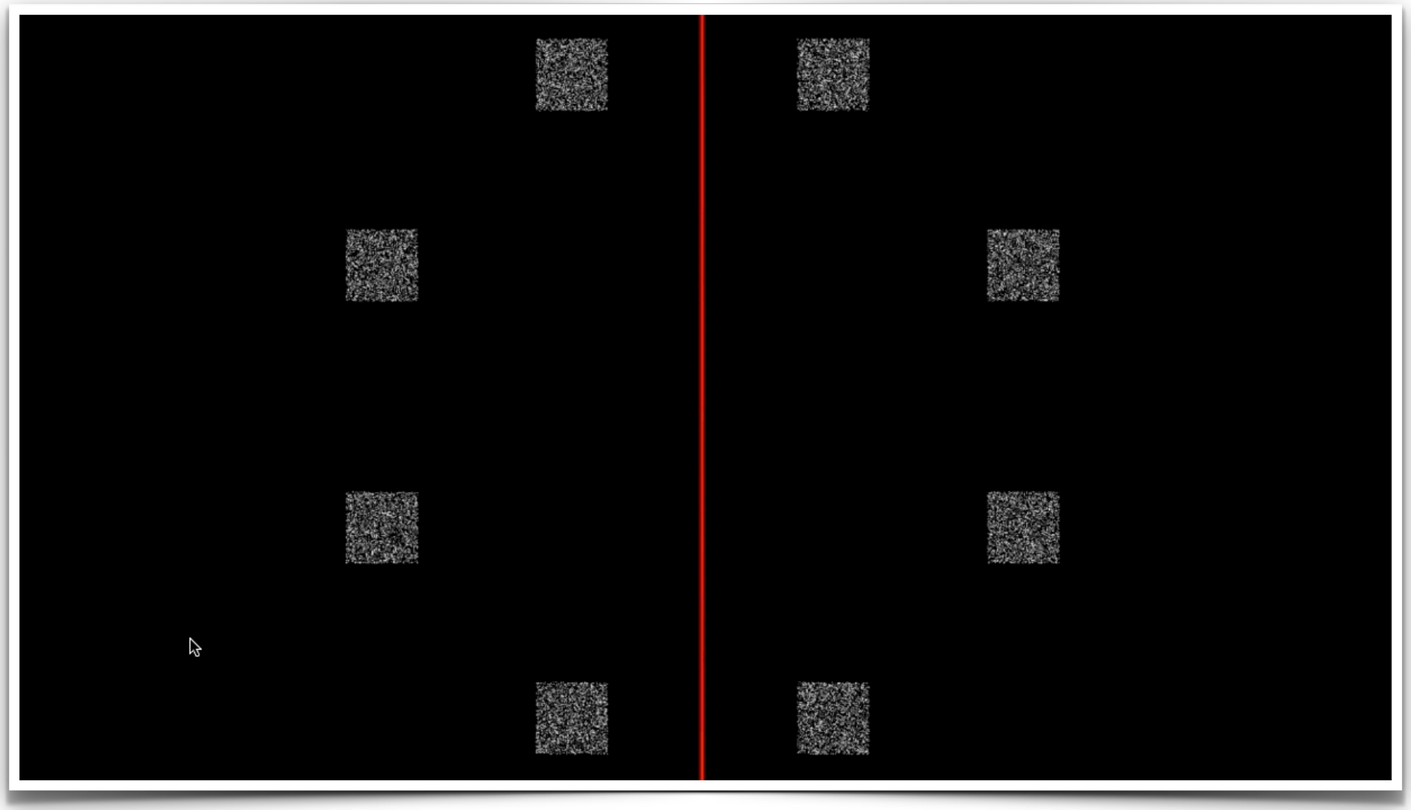


Odd movie

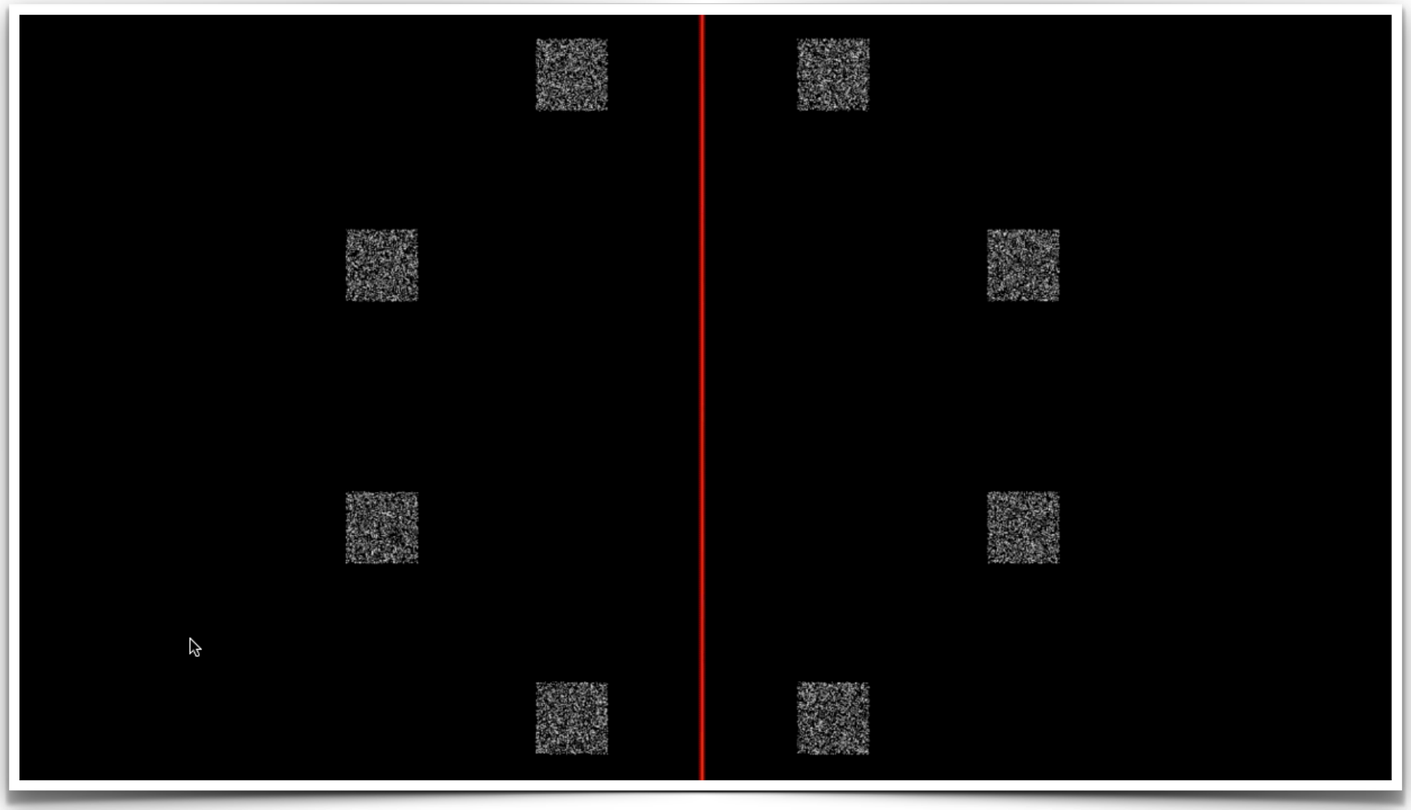
Identify the Odd Movie - 2



Identify the Odd Movie - 2

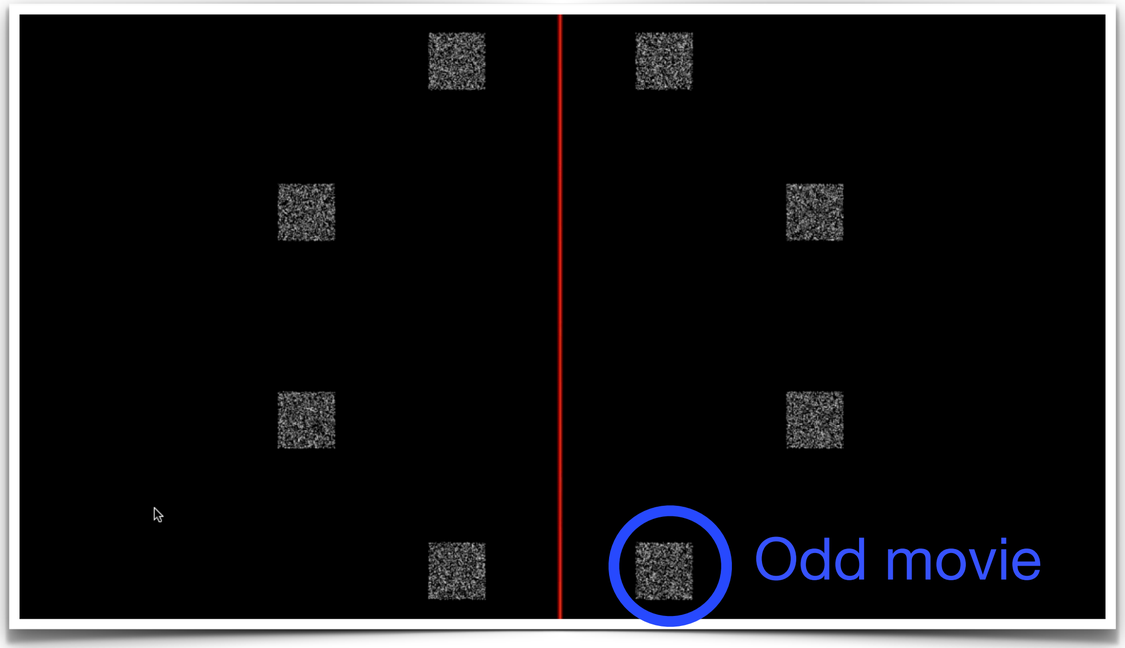


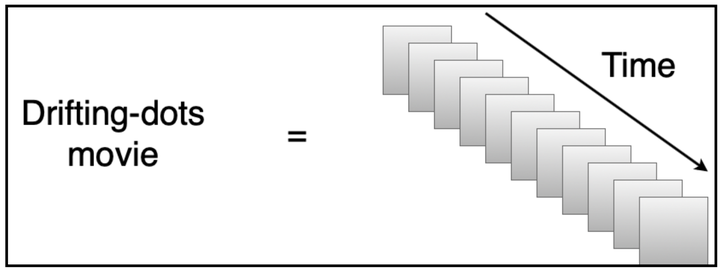
Identify the Odd Movie - 2



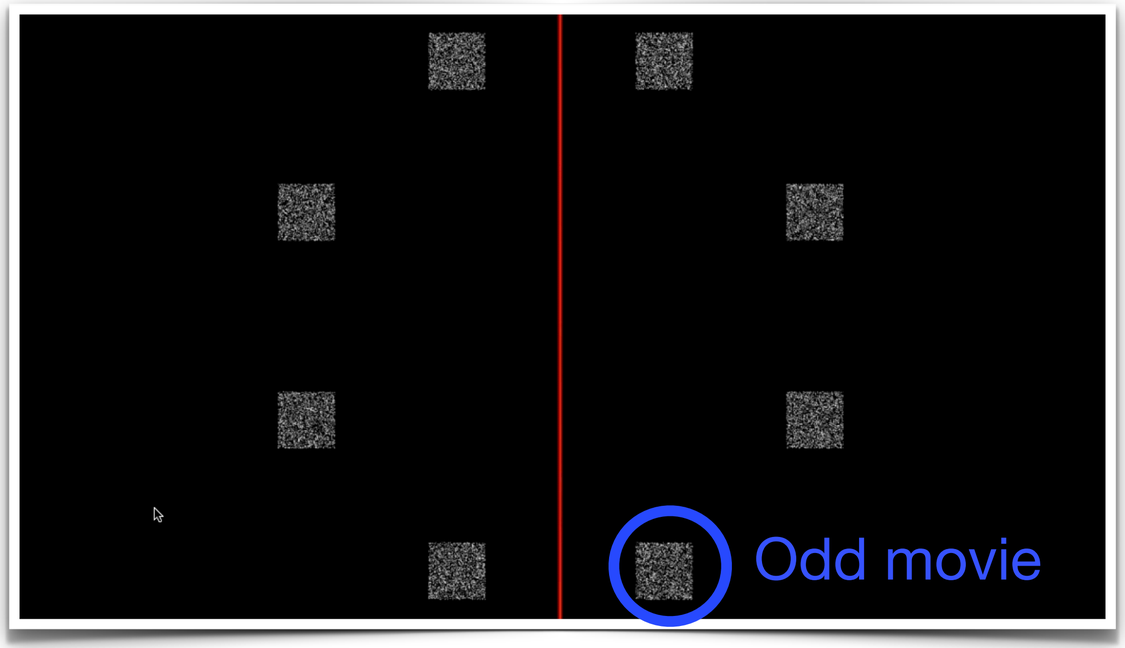
$$\text{Time to identify the odd movie} \propto \frac{1}{\text{hardness of the problem-instance}}$$

Odd Movie Experiment ↦ Multi-Armed Bandits

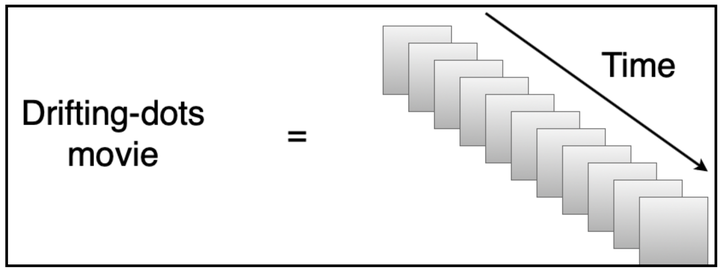




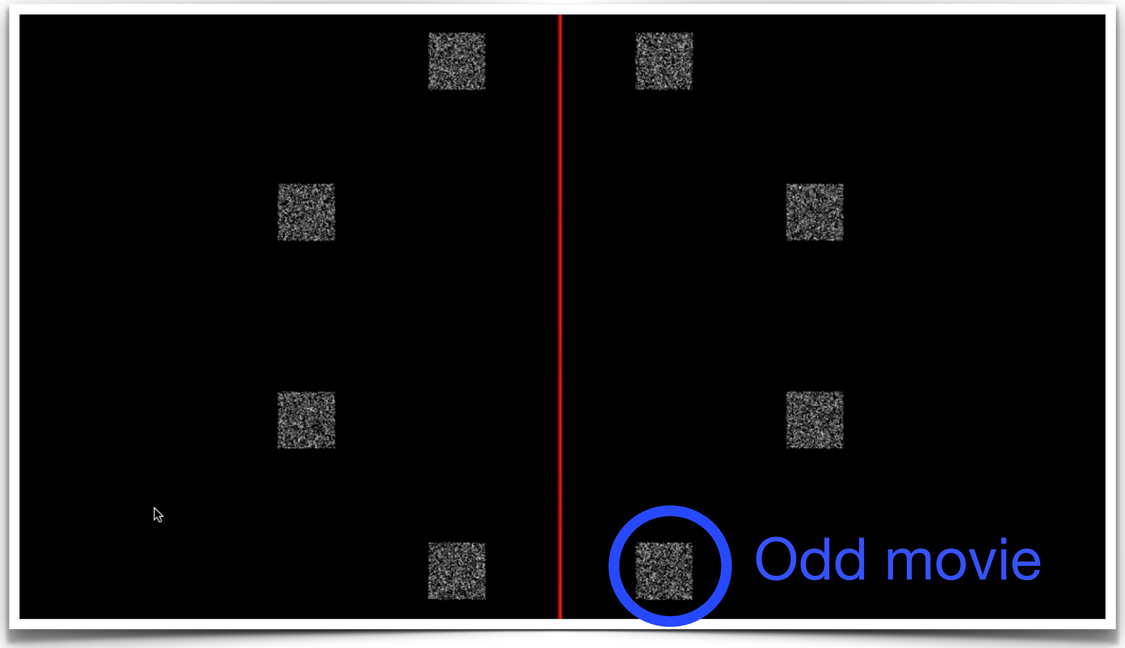
Odd Movie Experiment ↦ Multi-Armed Bandits



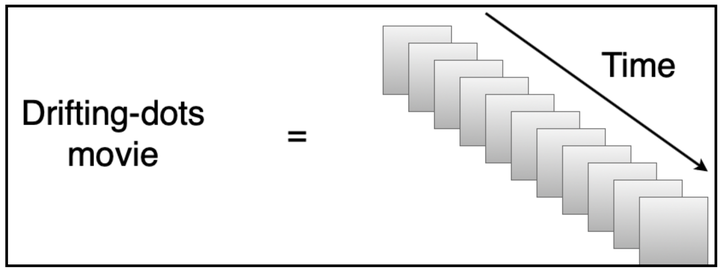
Odd Movie Experiment	Multi-Armed Bandits



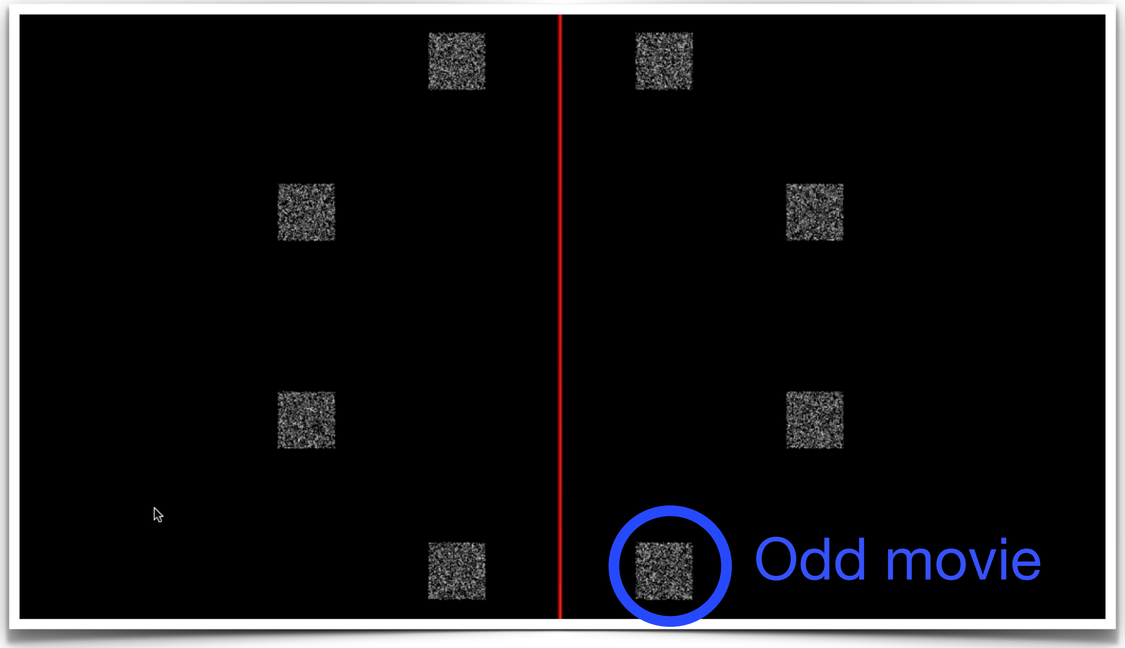
Odd Movie Experiment ↦ Multi-Armed Bandits



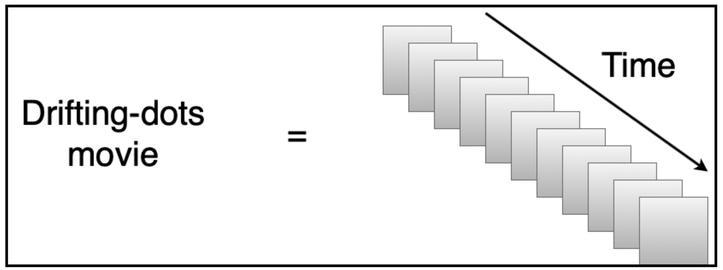
Odd Movie Experiment	Multi-Armed Bandits
Movie	Arm



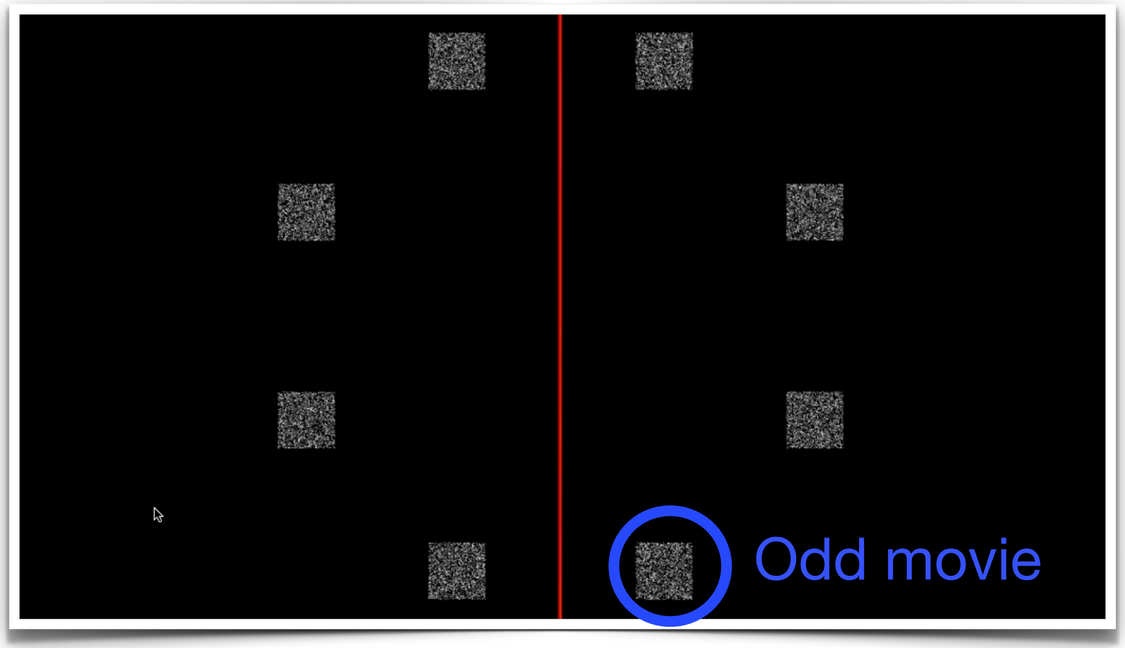
Odd Movie Experiment ↦ Multi-Armed Bandits



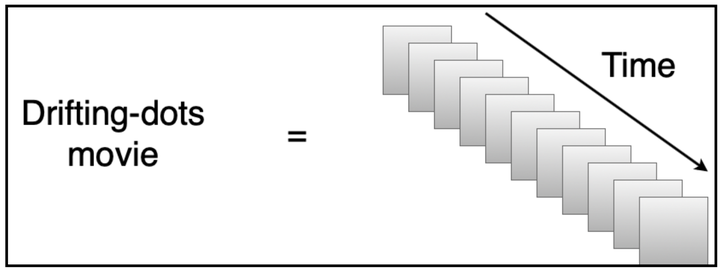
Odd Movie Experiment	Multi-Armed Bandits
Movie	Arm
Frame	Observation



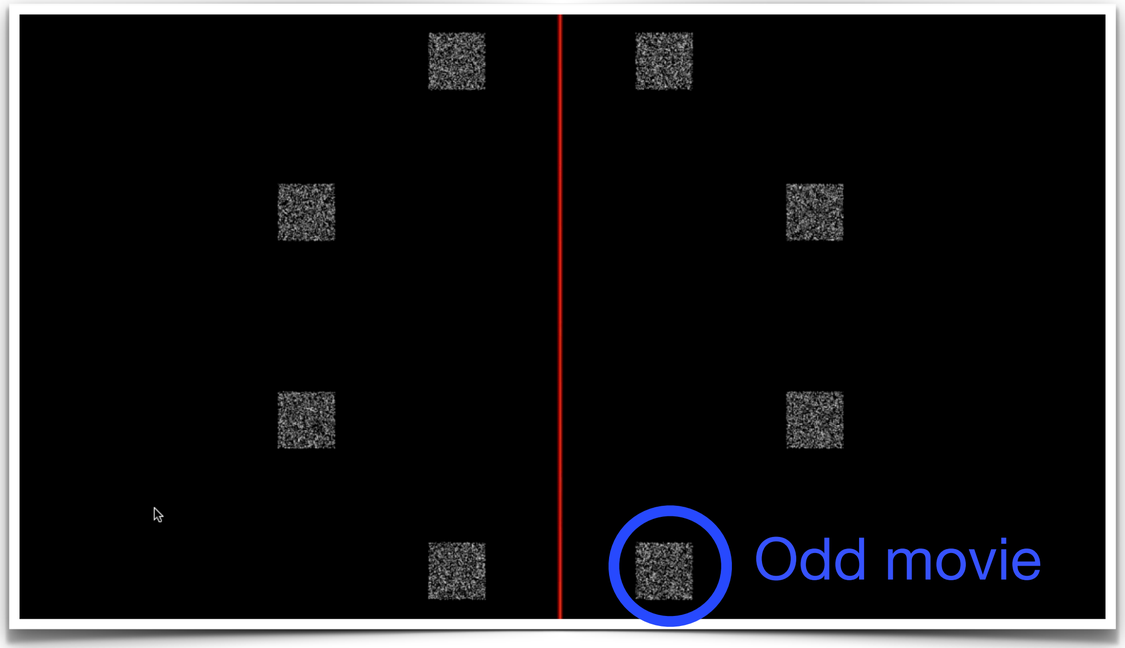
Odd Movie Experiment ↦ Multi-Armed Bandits



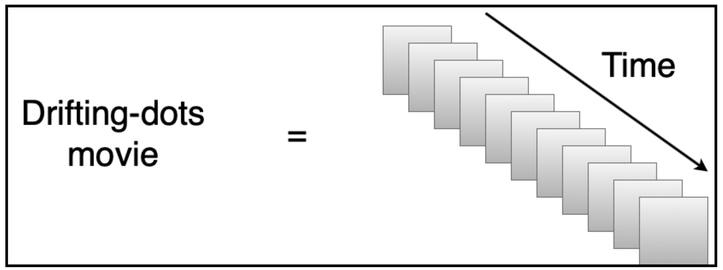
Odd Movie Experiment	Multi-Armed Bandits
Movie	Arm
Frame	Observation
Successive frames related	Observations form a Markov process



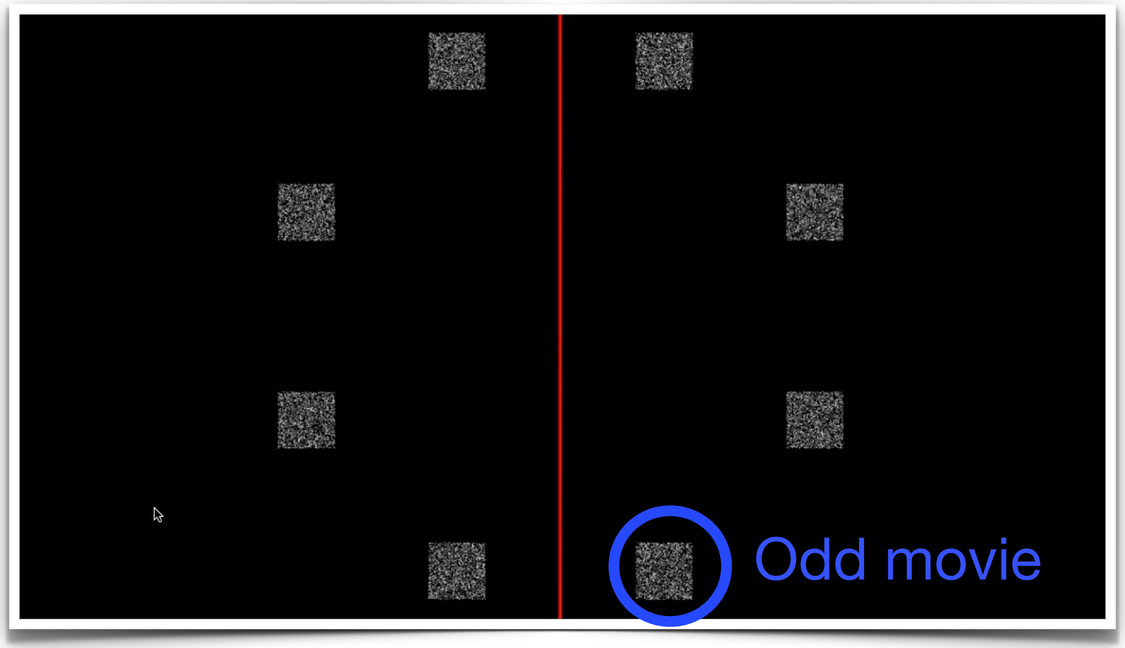
Odd Movie Experiment ↦ Multi-Armed Bandits

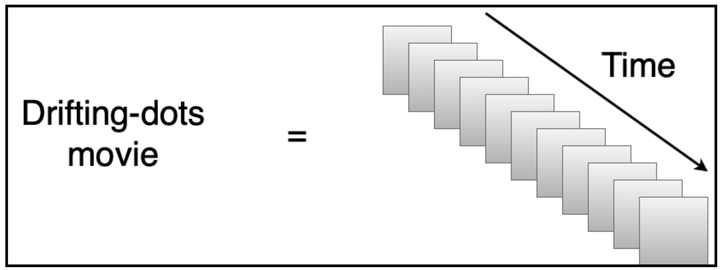


Odd Movie Experiment	Multi-Armed Bandits
Movie	Arm
Frame	Observation
Successive frames related	Observations form a Markov process
One movie is observed at a time	One arm is selected at a time



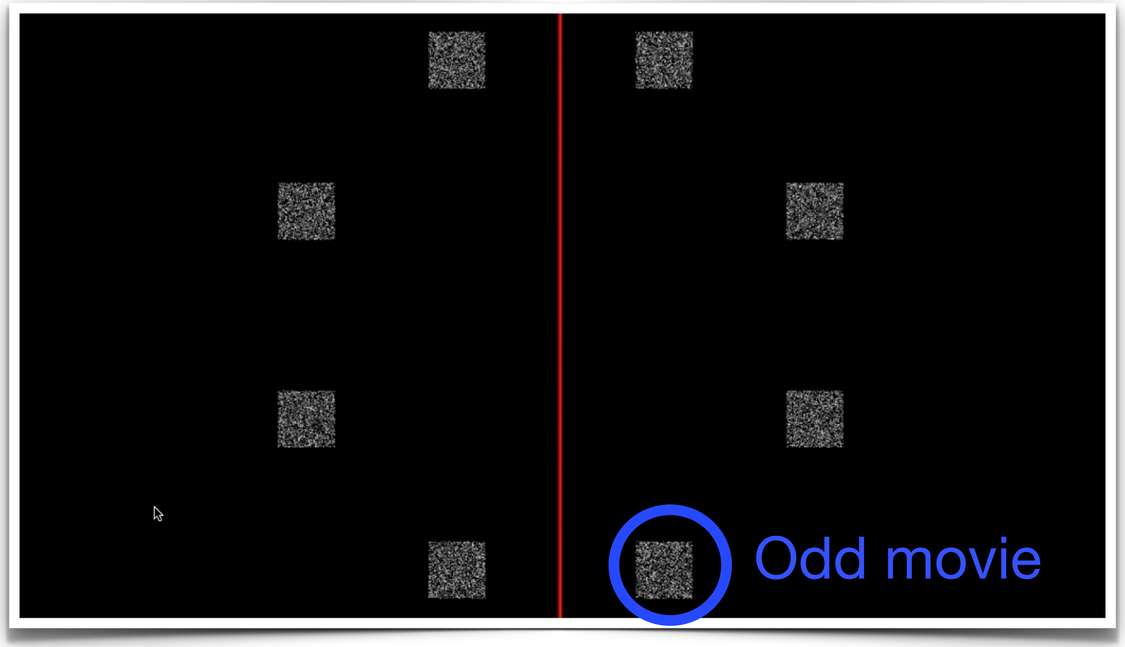
Odd Movie Experiment ↦ Multi-Armed Bandits

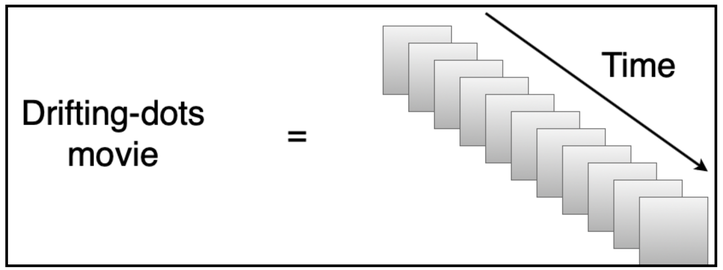




Odd Movie Experiment	Multi-Armed Bandits
Movie	Arm
Frame	Observation
Successive frames related	Observations form a Markov process
One movie is observed at a time	One arm is selected at a time
Unobserved movies continue to play	Unobserved arms continue to evolve (restless arms)

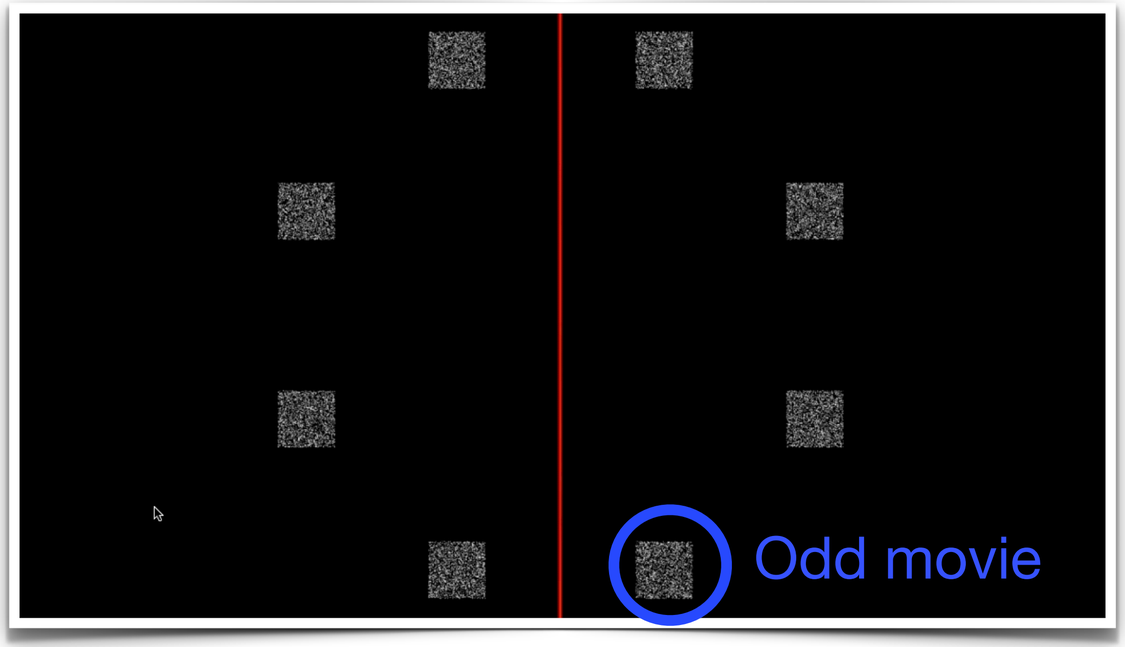
Odd Movie Experiment ↦ Multi-Armed Bandits

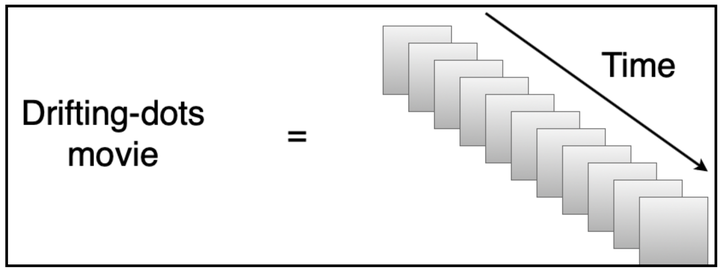




Odd Movie Experiment	Multi-Armed Bandits
Movie	Arm
Frame	Observation
Successive frames related	Observations form a Markov process
One movie is observed at a time	One arm is selected at a time
Unobserved movies continue to play	Unobserved arms continue to evolve (restless arms)
Drift in one of the movies is different	Markov law (TPM) of one of the arms is different

Odd Movie Experiment ↦ Multi-Armed Bandits

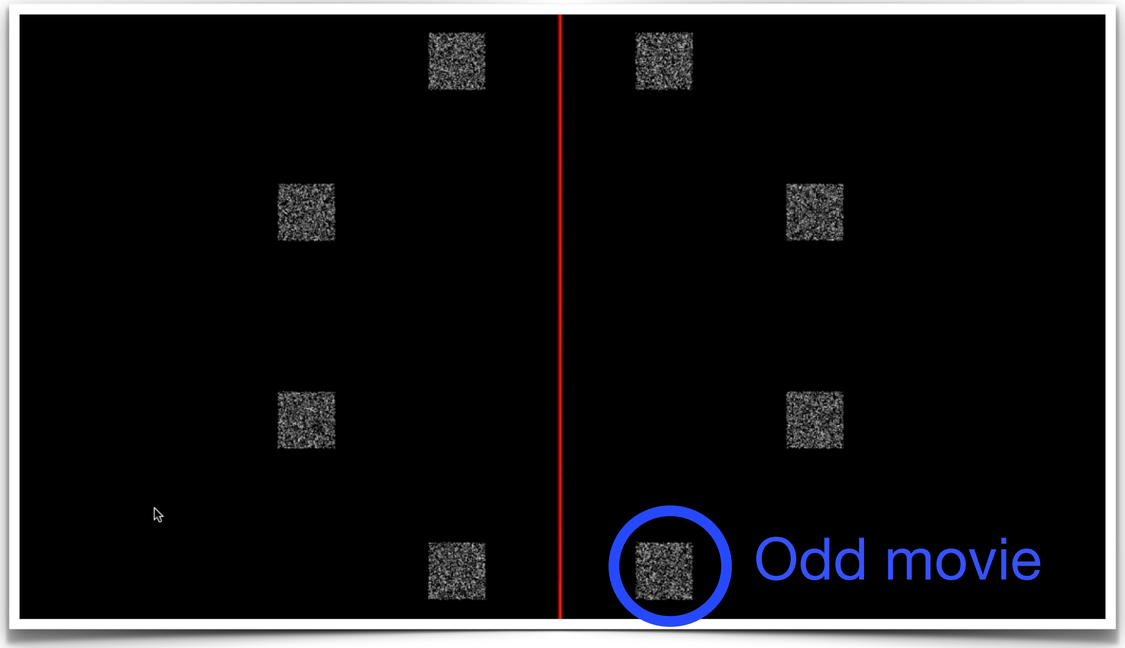


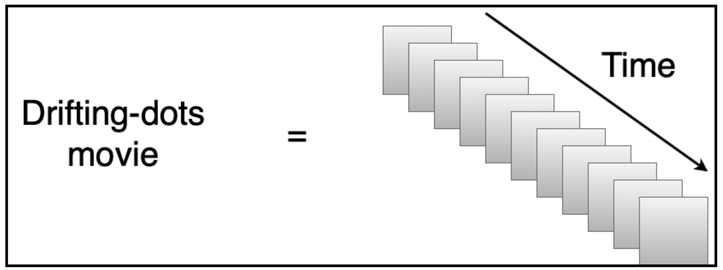


Odd Movie Experiment	Multi-Armed Bandits
Movie	Arm
Frame	Observation
Successive frames related	Observations form a Markov process
One movie is observed at a time	One arm is selected at a time
Unobserved movies continue to play	Unobserved arms continue to evolve (restless arms)
Drift in one of the movies is different	Markov law (TPM) of one of the arms is different

TPM: Transition Probability Matrix

Odd Movie Experiment ↦ Multi-Armed Bandits



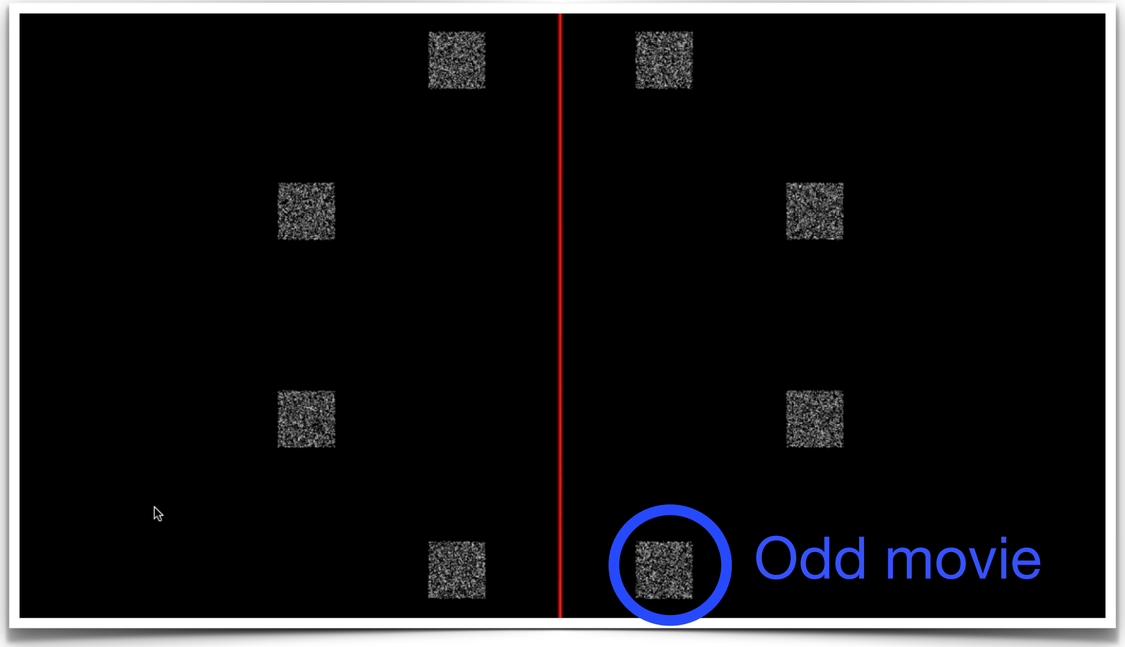


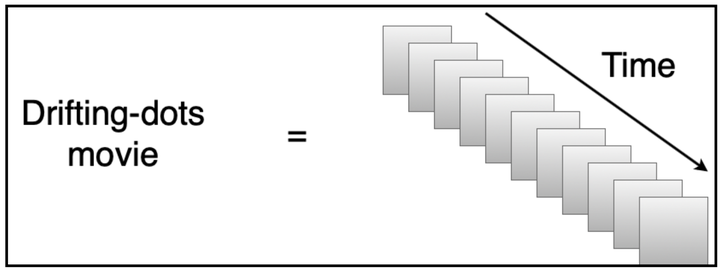
Odd Movie Experiment	Multi-Armed Bandits
Movie	Arm
Frame	Observation
Successive frames related	Observations form a Markov process
One movie is observed at a time	One arm is selected at a time
Unobserved movies continue to play	Unobserved arms continue to evolve (restless arms)
Drift in one of the movies is different	Markov law (TPM) of one of the arms is different

Goal: find the odd restless Markov arm as quickly and accurately as possible without the knowledge of the arm TPMs

TPM: Transition Probability Matrix

Odd Movie Experiment ↦ Multi-Armed Bandits





Odd Movie Experiment	Multi-Armed Bandits
Movie	Arm
Frame	Observation
Successive frames related	Observations form a Markov process
One movie is observed at a time	One arm is selected at a time
Unobserved movies continue to play	Unobserved arms continue to evolve (restless arms)
Drift in one of the movies is different	Markov law (TPM) of one of the arms is different

Goal: find the odd restless Markov arm as quickly and accurately as possible without the knowledge of the arm TPMs

learning

TPM: Transition Probability Matrix

Problem Setup and Objective

Learning to Detect an Odd Restless Markov Arm

Learning to Detect an Odd Restless Markov Arm

Learning to Detect an Odd Restless Markov Arm

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

Learning to Detect an Odd Restless Markov Arm

- A multi-armed bandit with $K \geq 3$ arms
- Each arm is a time homogeneous and ergodic Markov process

Learning to Detect an Odd Restless Markov Arm

- A multi-armed bandit with $K \geq 3$ arms
- Each arm is a time homogeneous and ergodic Markov process
- Markov processes evolve on a common, finite state space

Learning to Detect an Odd Restless Markov Arm

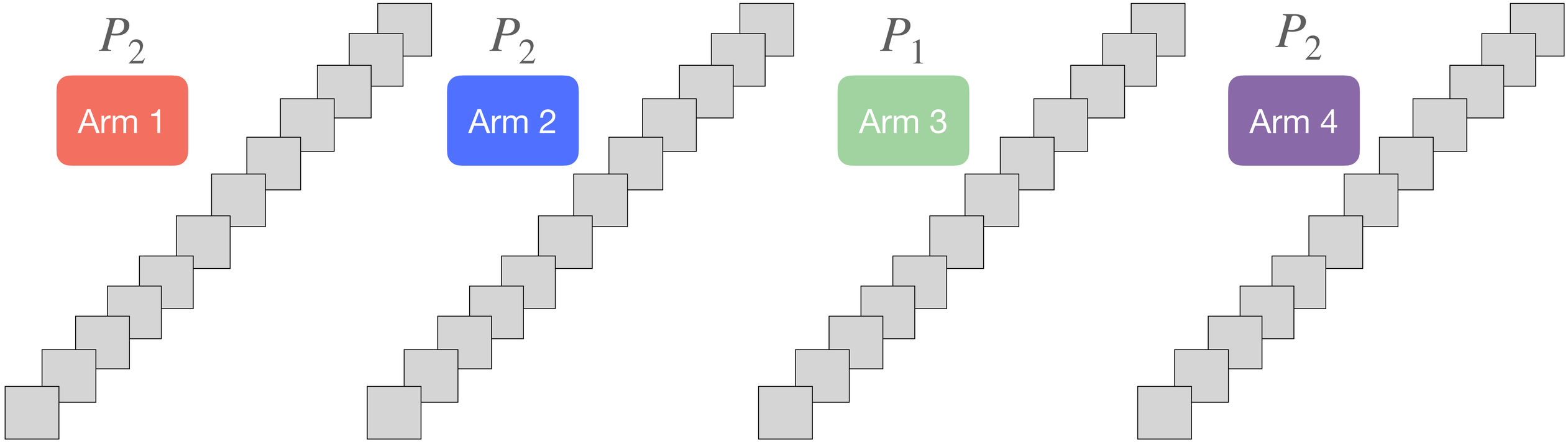
- A multi-armed bandit with $K \geq 3$ arms
- Each arm is a time homogeneous and ergodic Markov process
- Markov processes evolve on a common, finite state space
- The TPM of one of the arms (odd arm) is P_1; TPM of rest of the arms is P_2

Learning to Detect an Odd Restless Markov Arm

- A multi-armed bandit with $K \geq 3$ arms
- Each arm is a time homogeneous and ergodic Markov process
- Markov processes evolve on a common, finite state space
- The TPM of one of the arms (odd arm) is P_1; TPM of rest of the arms is P_2
- Arms are restless

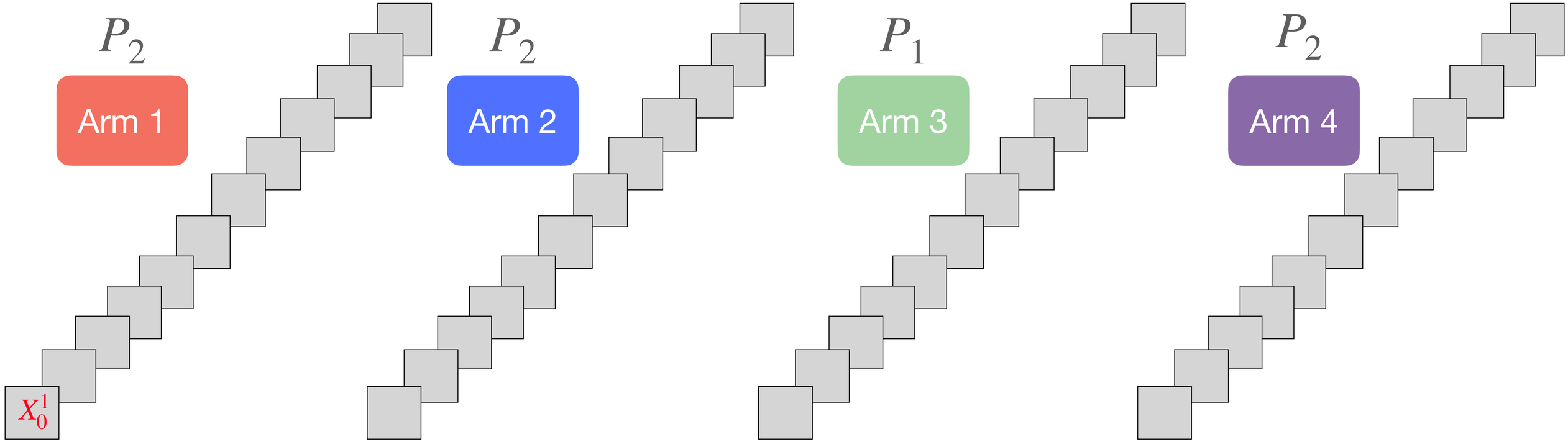
Learning to Detect an Odd Restless Markov Arm

- A multi-armed bandit with $K \geq 3$ arms
- Each arm is a time homogeneous and ergodic Markov process
- Markov processes evolve on a common, finite state space
- The TPM of one of the arms (odd arm) is P_1; TPM of rest of the arms is P_2
- Arms are restless
- TPMs are unknown (learning)



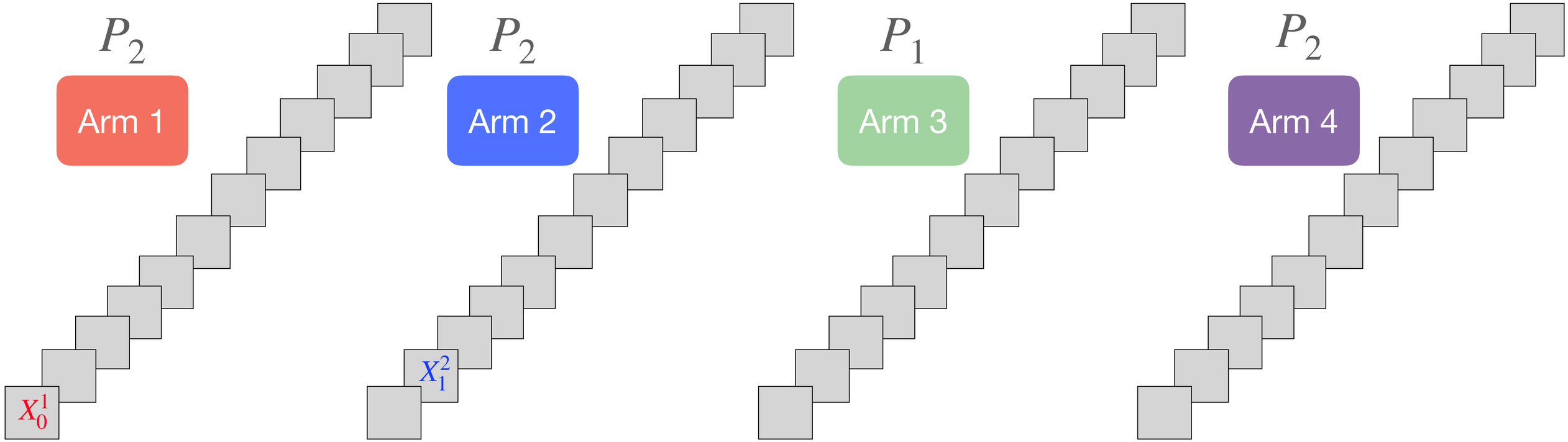
t	A_t	**Arm 1**		**Arm 2**		**Arm 3**		**Arm 4**	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t



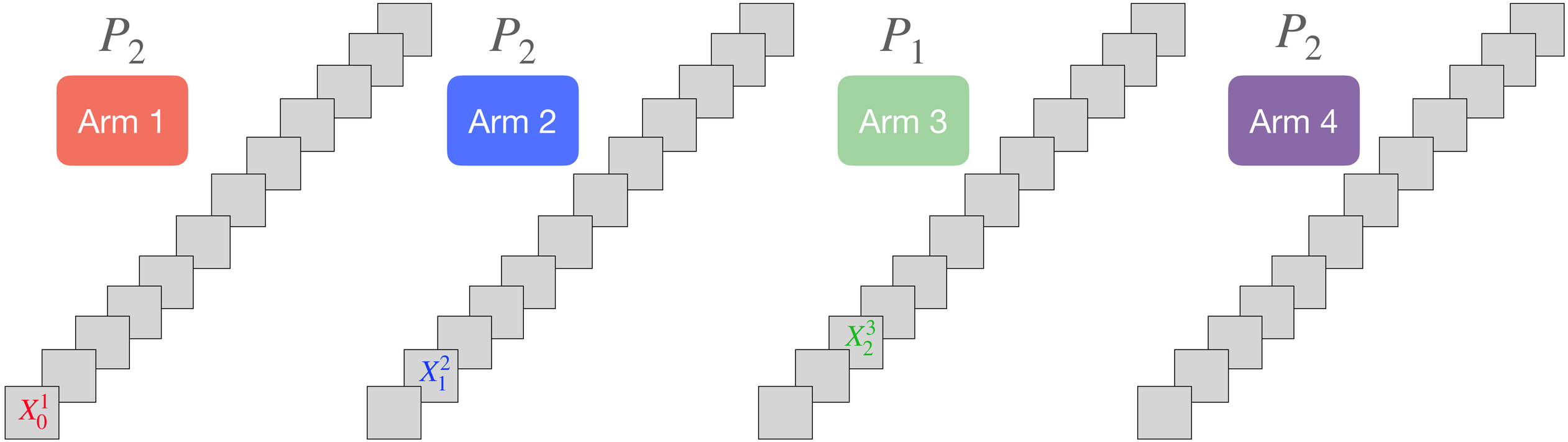
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								

X_t^a : observation from arm a at time t $d_a(t)$: delay of arm a at time t $i_a(t)$: last observed state of arm a at time t



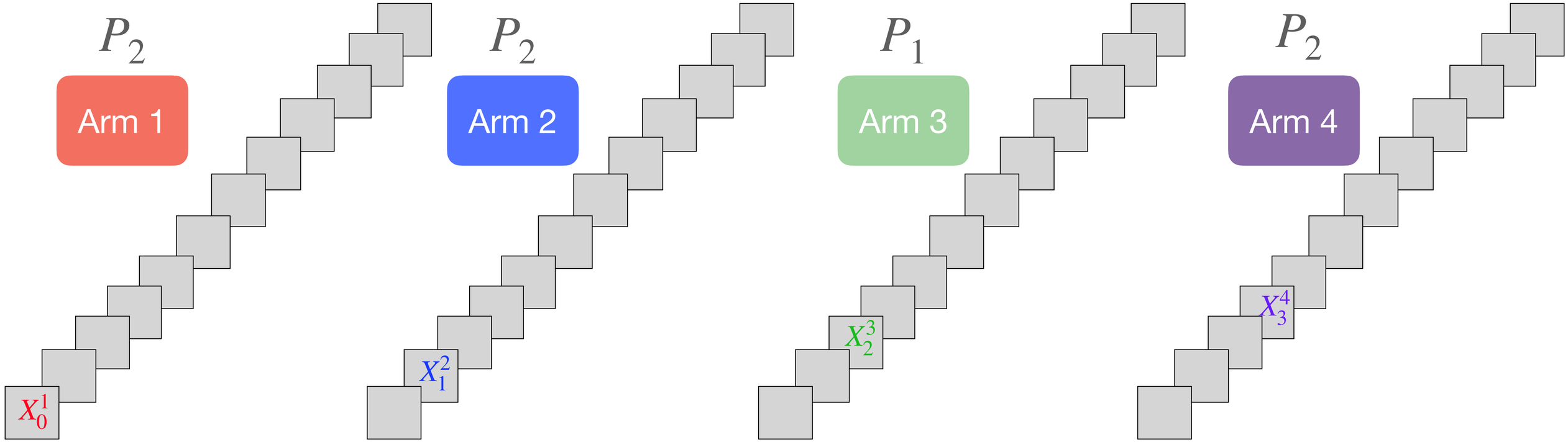
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t



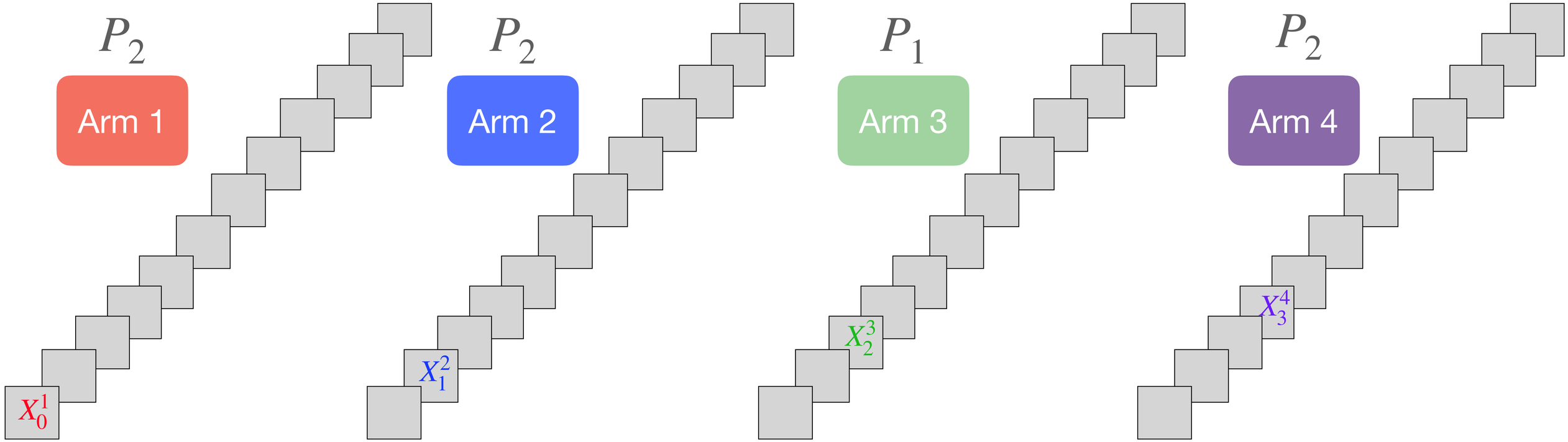
t	A_t	**Arm 1**		**Arm 2**		**Arm 3**		**Arm 4**	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t



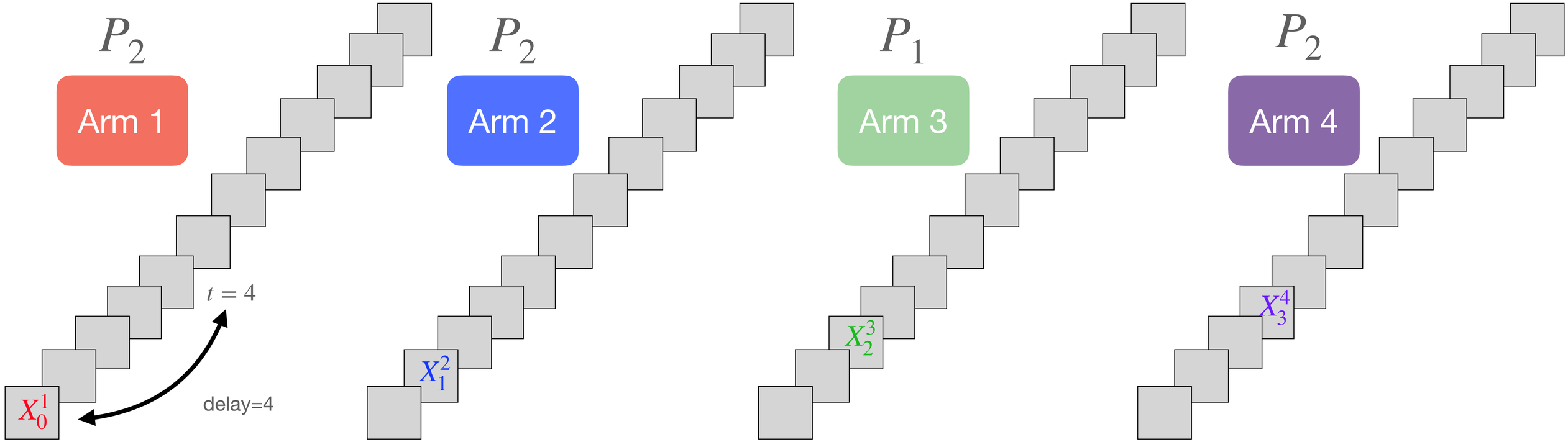
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t



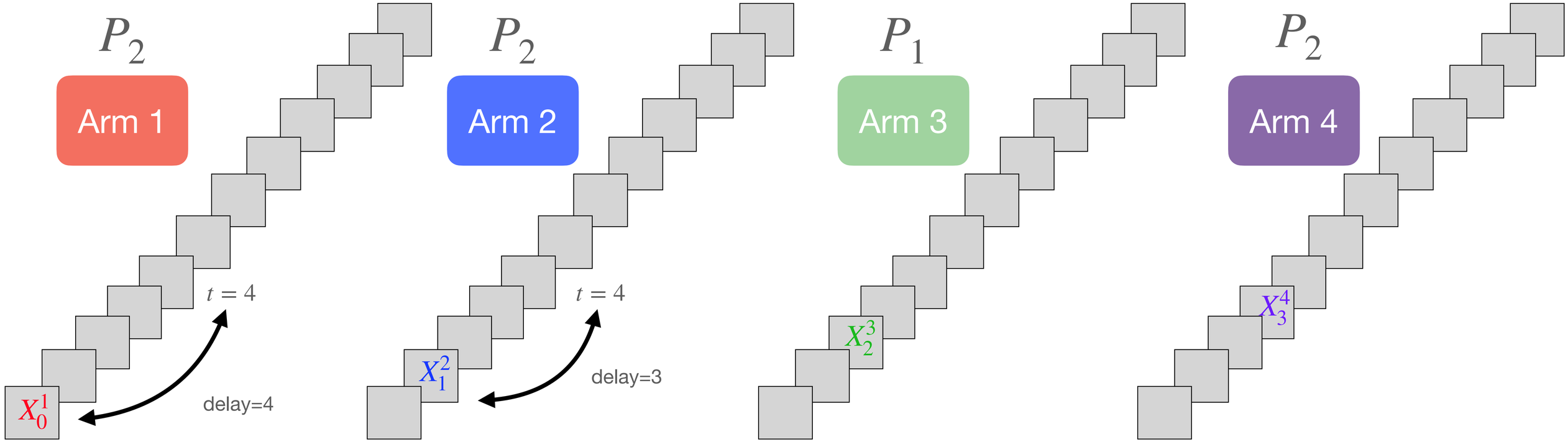
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4		4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t



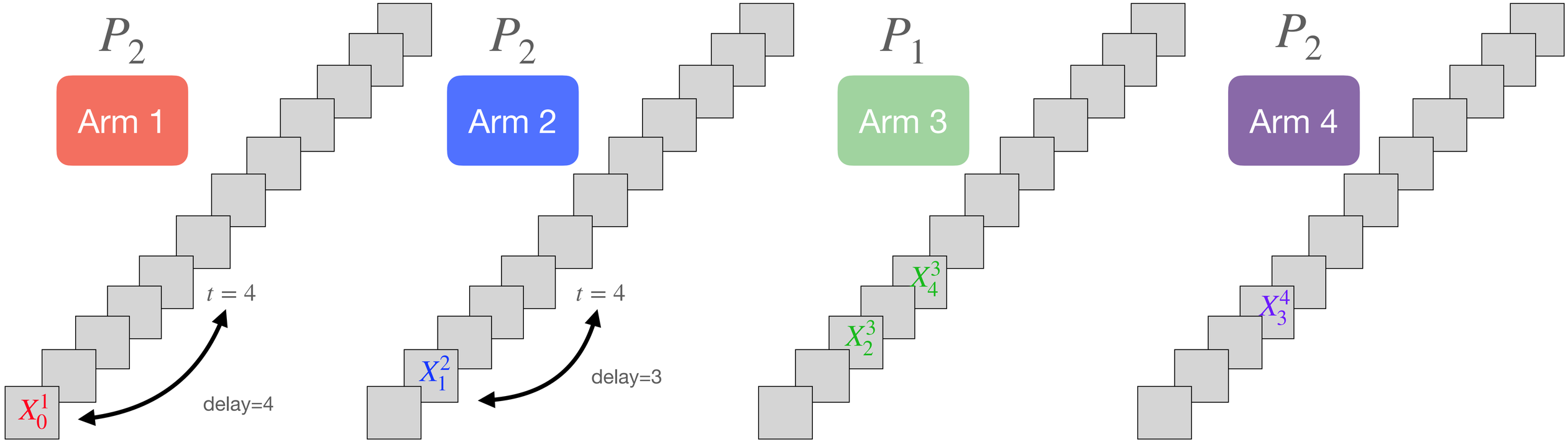
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4		4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t



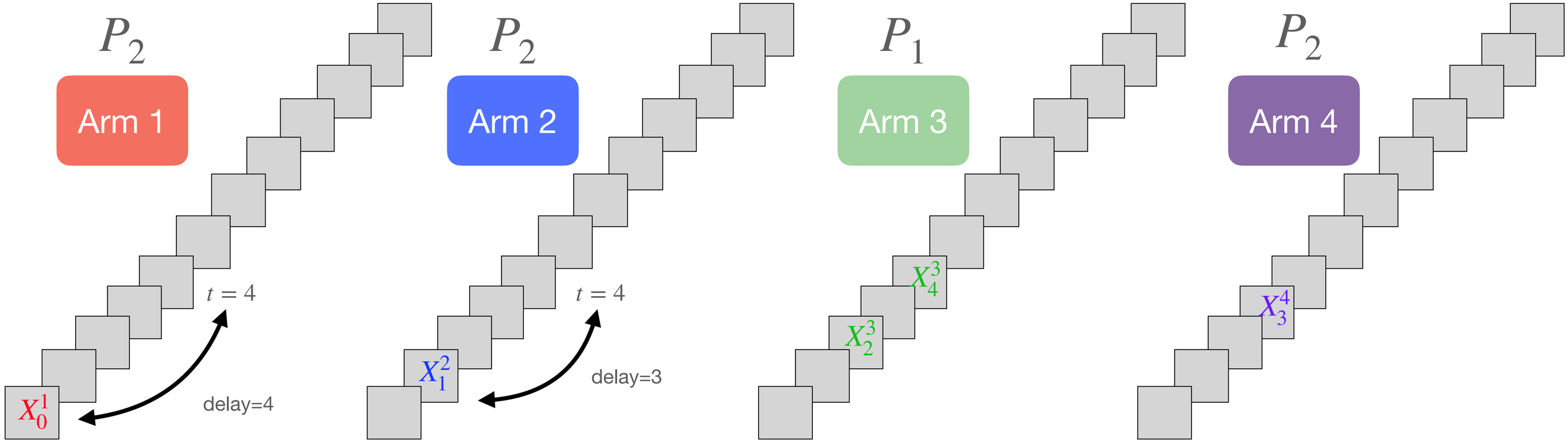
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4		4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t



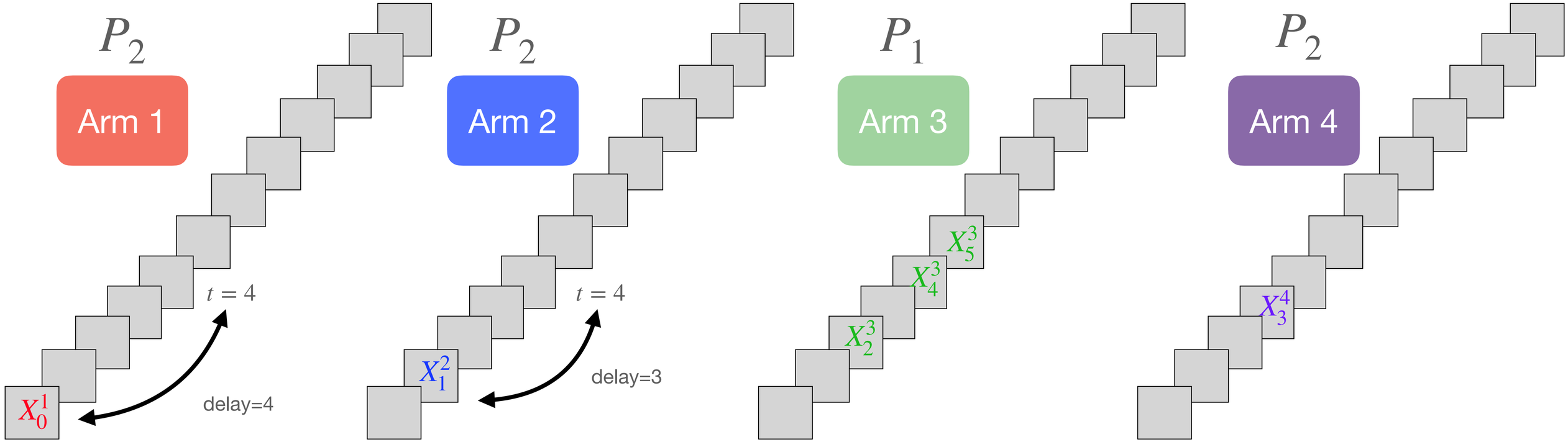
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t



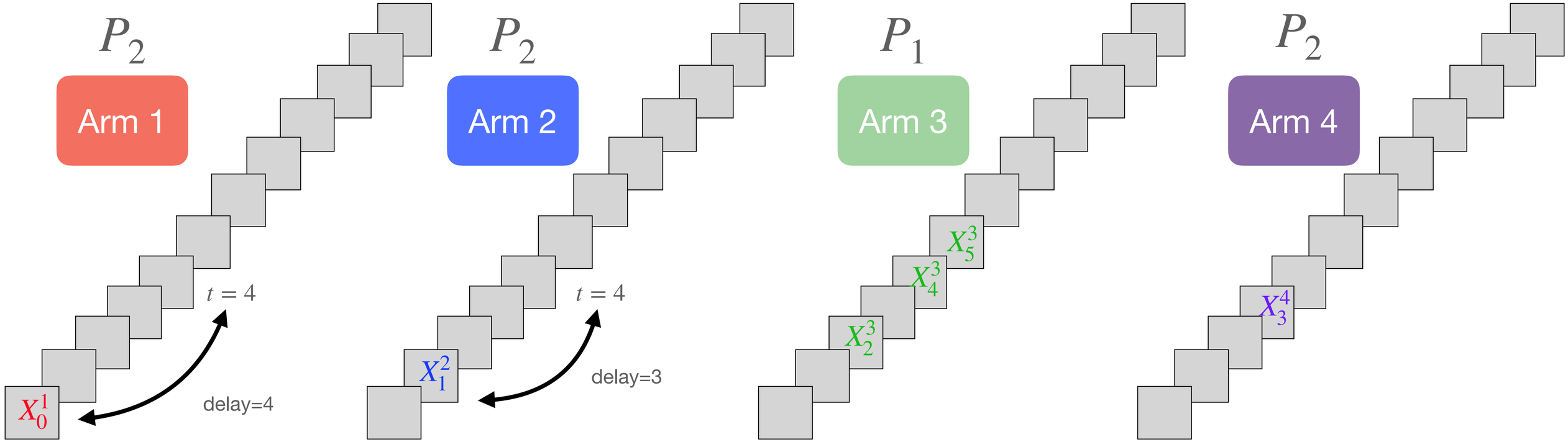
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5		5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t



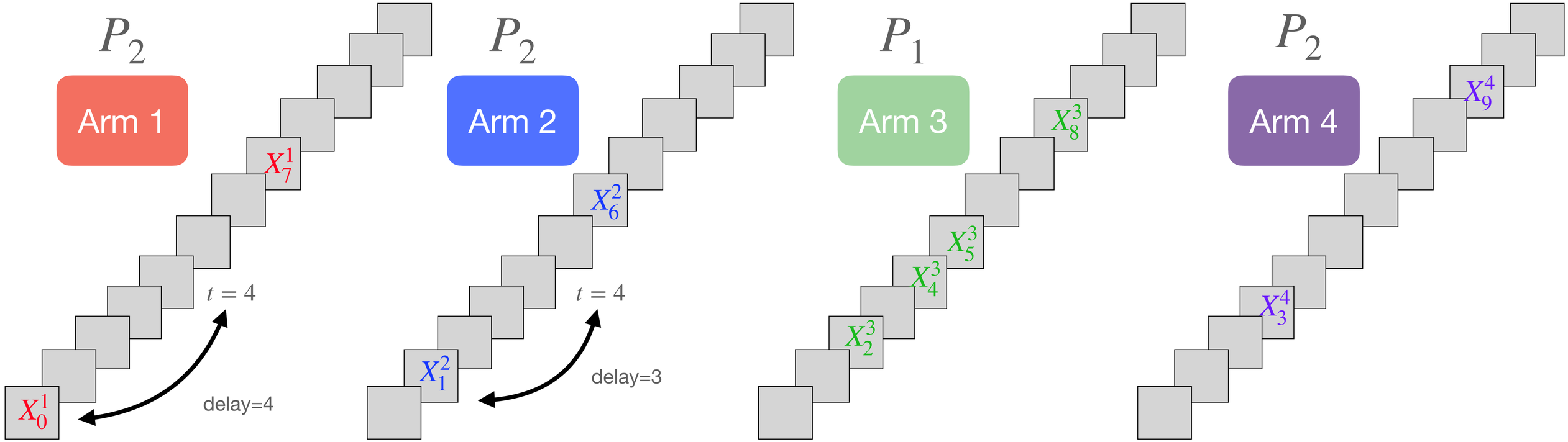
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6		6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t

Objective

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
    - $\tau(\pi)$: stopping time
    - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
    - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

Odd arm

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

TPM of odd arm

Odd arm

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

h: Odd arm
P_1: TPM of odd arm
P_2: TPM of each non-odd arm

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

TPM of odd arm

Odd arm

TPM of each non-odd arm

$$\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$$

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

TPM of odd arm

Odd arm

TPM of each non-odd arm

$$\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$$

Quickly

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

TPM of odd arm

Odd arm

TPM of each non-odd arm

Accurately

$$\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$$

Quickly

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

TPM of odd arm

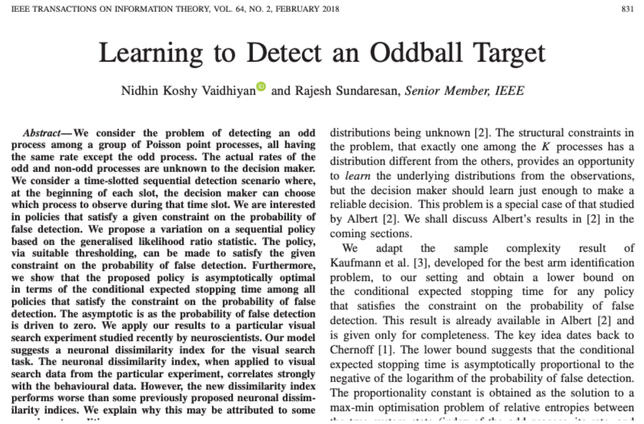
Odd arm

TPM of each non-odd arm

Accurately

$$\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$$

Quickly



IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 64, NO. 2, FEBRUARY 2018 831

Learning to Detect an Oddball Target

Nidhin Koshy Vaidhiyan and Rajesh Sundaresan, *Senior Member, IEEE*

***Abstract*— We consider the problem of detecting an odd process among a group of Poisson point processes, all having the same rate except the odd process. The actual rates of the odd and non-odd processes are unknown to the decision maker. We consider a time-slotted sequential detection scenario where, at the beginning of each slot, the decision maker can choose which process to observe during that time slot. We are interested in policies that satisfy a given constraint on the probability of false detection. We propose a variation on a sequential policy based on the generalised likelihood ratio statistic. The policy, via suitable thresholding, can be made to satisfy the given constraint on the probability of false detection. Furthermore, we show that the proposed policy is asymptotically optimal in terms of the conditional expected stopping time among all policies that satisfy the constraint on the probability of false detection. The asymptotic is as the probability of false detection is driven to zero. We apply our results to a particular visual search experiment studied recently by neuroscientists. Our model suggests a neuronal dissimilarity index for the visual search task. The neuronal dissimilarity index, when applied to visual search data from the particular experiment, correlates strongly with the behavioural data. However, the new dissimilarity index performs worse than some previously proposed neuronal dissimilarity indices. We explain why this may be attributed to some** distributions being unknown [2]. The structural constraints in the problem, that exactly one among the K processes has a distribution different from the others, provides an opportunity to *learn* the underlying distributions from the observations, but the decision maker should learn just enough to make a reliable decision. This problem is a special case of that studied by Albert [2]. We shall discuss Albert's results in [2] in the coming sections.

We adapt the sample complexity result of Kaufmann et al. [3], developed for the best arm identification problem, to our setting and obtain a lower bound on the conditional expected stopping time for any policy that satisfies the constraint on the probability of false detection. This result is already available in Albert [2] and is given only for completeness. The key idea dates back to Chernoff [1]. The lower bound suggests that the conditional expected stopping time is asymptotically proportional to the negative of the logarithm of the probability of false detection. The proportionality constant is obtained as the solution to a max-min optimisation problem of relative entropies between

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

TPM of odd arm

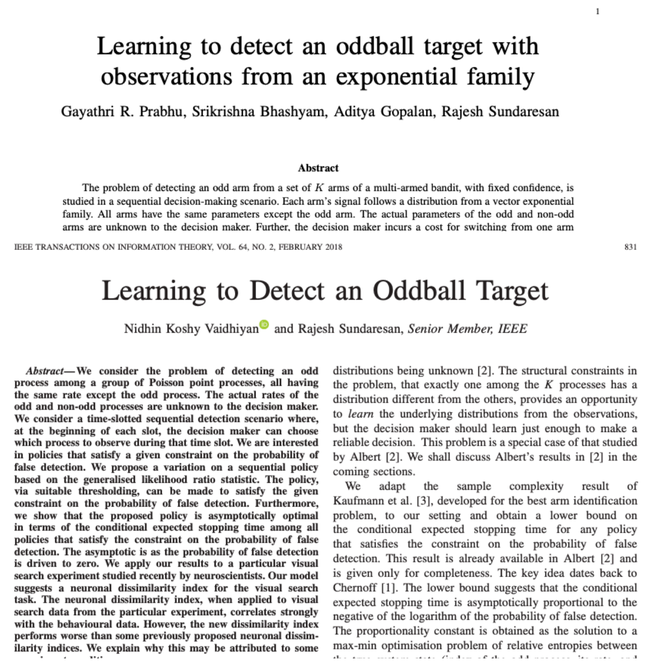
Odd arm

TPM of each non-odd arm

Accurately

$$\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$$

Quickly



1

Learning to detect an oddball target with observations from an exponential family

Gayathri R. Prabhu, Srikrishna Bhashyam, Aditya Gopalan, Rajesh Sundaresan

Abstract

The problem of detecting an odd arm from a set of K arms of a multi-armed bandit, with fixed confidence, is studied in a sequential decision-making scenario. Each arm's signal follows a distribution from a vector exponential family. All arms have the same parameters except the odd arm. The actual parameters of the odd and non-odd arms are unknown to the decision maker. Further, the decision maker incurs a cost for switching from one arm

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 64, NO. 2, FEBRUARY 2018 831

Learning to Detect an Oddball Target

Nidhin Koshy Vaidhiyan and Rajesh Sundaresan, *Senior Member, IEEE*

***Abstract*— We consider the problem of detecting an odd process among a group of Poisson point processes, all having the same rate except the odd process. The actual rates of the odd and non-odd processes are unknown to the decision maker. We consider a time-slotted sequential detection scenario where, at the beginning of each slot, the decision maker can choose which process to observe during that time slot. We are interested in policies that satisfy a given constraint on the probability of false detection. We propose a variation on a sequential policy based on the generalised likelihood ratio statistic. The policy, via suitable thresholding, can be made to satisfy the given constraint on the probability of false detection. Furthermore, we show that the proposed policy is asymptotically optimal in terms of the conditional expected stopping time among all policies that satisfy the constraint on the probability of false detection. The asymptotic is as the probability of false detection is driven to zero. We apply our results to a particular visual search experiment studied recently by neuroscientists. Our model suggests a neuronal dissimilarity index for the visual search task. The neuronal dissimilarity index, when applied to visual search data from the particular experiment, correlates strongly with the behavioural data. However, the new dissimilarity index performs worse than some previously proposed neuronal dissimilarity indices. We explain why this may be attributed to some** distributions being unknown [2]. The structural constraints in the problem, that exactly one among the K processes has a distribution different from the others, provides an opportunity to *learn* the underlying distributions from the observations, but the decision maker should learn just enough to make a reliable decision. This problem is a special case of that studied by Albert [2]. We shall discuss Albert's results in [2] in the coming sections.

We adapt the sample complexity result of Kaufmann et al. [3], developed for the best arm identification problem, to our setting and obtain a lower bound on the conditional expected stopping time for any policy that satisfies the constraint on the probability of false detection. This result is already available in Albert [2] and is given only for completeness. The key idea dates back to Chernoff [1]. The lower bound suggests that the conditional expected stopping time is asymptotically proportional to the negative of the logarithm of the probability of false detection. The proportionality constant is obtained as the solution to a max-min optimisation problem of relative entropies between

Objective

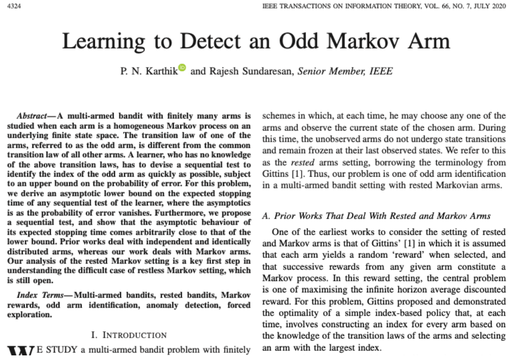
- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

h: Odd arm; P_1: TPM of odd arm; P_2: TPM of each non-odd arm

Accurately

$$\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$$

Quickly



4324 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 66, NO. 7, JULY 2020

Learning to Detect an Odd Markov Arm

P. N. Karthik and Rajesh Sundaresan, *Senior Member, IEEE*

***Abstract*— A multi-armed bandit with finitely many arms is studied when each arm is a homogeneous Markov process on an underlying finite state space. The transition law of one of the arms, referred to as the odd arm, is different from the common transition law of all other arms. A learner, who has no knowledge of the above transition laws, has to devise a sequential test to identify the index of the odd arm as quickly as possible, subject to an upper bound on the probability of error. For this problem, we derive an asymptotic lower bound on the expected stopping time of any sequential test of the learner, where the asymptotics is as the probability of error vanishes. Furthermore, we propose a sequential test, and show that the asymptotic behaviour of its expected stopping time comes arbitrarily close to that of the lower bound. Prior works deal with independent and identically distributed arms, whereas our work deals with Markov arms. Our analysis of the rested Markov setting is a key first step in understanding the difficult case of restless Markov setting, which is still open.**

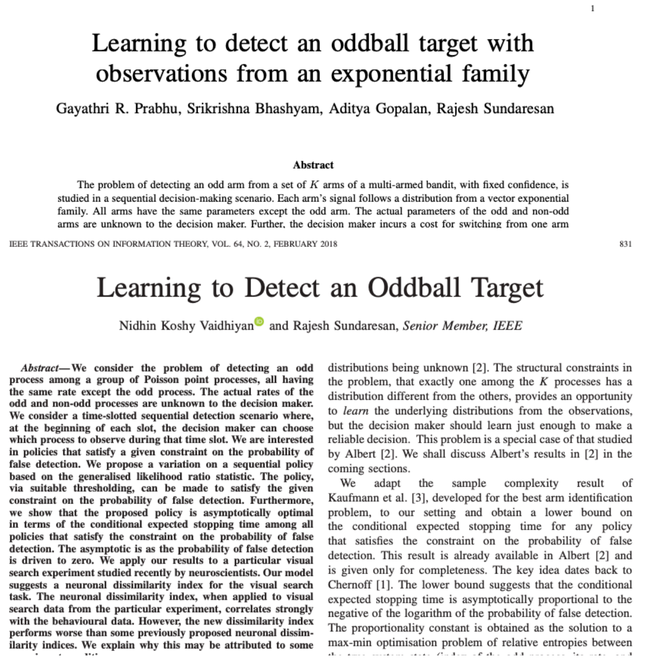
***Index Terms*— Multi-armed bandits, rested bandits, Markov rewards, odd arm identification, anomaly detection, forced exploration.**

I. Introduction

WE STUDY a multi-armed bandit problem with finitely schemes in which, at each time, he may choose any one of the arms and observe the current state of the chosen arm. During this time, the unobserved arms do not undergo state transitions and remain frozen at their last observed states. We refer to this as the *rested* arms setting, borrowing the terminology from Gittins [1]. Thus, our problem is one of odd arm identification in a multi-armed bandit setting with rested Markovian arms.

A. Prior Works That Deal With Rested and Markov Arms

One of the earliest works to consider the setting of rested and Markov arms is that of Gittins' [1] in which it is assumed that each arm yields a random 'reward' when selected, and that successive rewards from any given arm constitute a Markov process. In this reward setting, the central problem is one of maximising the infinite horizon average discounted reward. For this problem, Gittins proposed and demonstrated the optimality of a simple index-based policy that, at each time, involves constructing an index for every arm based on the knowledge of the transition laws of the arms and selecting an arm with the largest index.



1

Learning to detect an oddball target with observations from an exponential family

Gayathri R. Prabhu, Srikrishna Bhashyam, Aditya Gopalan, Rajesh Sundaresan

Abstract

The problem of detecting an odd arm from a set of K arms of a multi-armed bandit, with fixed confidence, is studied in a sequential decision-making scenario. Each arm's signal follows a distribution from a vector exponential family. All arms have the same parameters except the odd arm. The actual parameters of the odd and non-odd arms are unknown to the decision maker. Further, the decision maker incurs a cost for switching from one arm

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 64, NO. 2, FEBRUARY 2018 831

Learning to Detect an Oddball Target

Nidhin Koshy Vaidhiyan and Rajesh Sundaresan, *Senior Member, IEEE*

***Abstract*— We consider the problem of detecting an odd process among a group of Poisson point processes, all having the same rate except the odd process. The actual rates of the odd and non-odd processes are unknown to the decision maker. We consider a time-slotted sequential detection scenario where, at the beginning of each slot, the decision maker can choose which process to observe during that time slot. We are interested in policies that satisfy a given constraint on the probability of false detection. We propose a variation on a sequential policy based on the generalised likelihood ratio statistic. The policy, via suitable thresholding, can be made to satisfy the given constraint on the probability of false detection. Furthermore, we show that the proposed policy is asymptotically optimal in terms of the conditional expected stopping time among all policies that satisfy the constraint on the probability of false detection. The asymptotic is as the probability of false detection is driven to zero. We apply our results to a particular visual search experiment studied recently by neuroscientists. Our model suggests a neuronal dissimilarity index for the visual search task. The neuronal dissimilarity index, when applied to visual search data from the particular experiment, correlates strongly with the behavioural data. However, the new dissimilarity index performs worse than some previously proposed neuronal dissimilarity indices. We explain why this may be attributed to some**

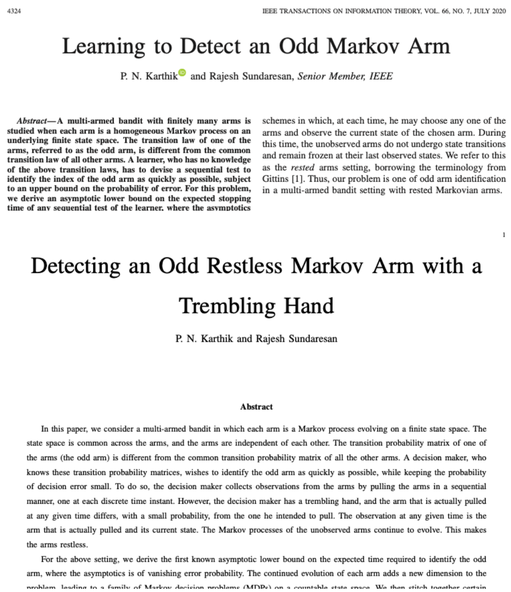
distributions being unknown [2]. The structural constraints in the problem, that exactly one among the K processes has a distribution different from the others, provides an opportunity to *learn* the underlying distributions from the observations, but the decision maker should learn just enough to make a reliable decision. This problem is a special case of that studied by Albert [2]. We shall discuss Albert's results in [2] in the coming sections.

We adapt the sample complexity result of Kaufmann et al. [3], developed for the best arm identification problem, to our setting and obtain a lower bound on the conditional expected stopping time for any policy that satisfies the constraint on the probability of false detection. This result is already available in Albert [2] and is given only for completeness. The key idea dates back to Chernoff [1]. The lower bound suggests that the conditional expected stopping time is asymptotically proportional to the negative of the logarithm of the probability of false detection. The proportionality constant is obtained as the solution to a max-min optimisation problem of relative entropies between

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

h: Odd arm; P_1: TPM of odd arm; P_2: TPM of each non-odd arm



4324 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 66, NO. 7, JULY 2020

Learning to Detect an Odd Markov Arm

P. N. Karthik and Rajesh Sundaresan, *Senior Member, IEEE*

***Abstract*— A multi-armed bandit with finitely many arms is studied when each arm is a homogeneous Markov process on an underlying finite state space. The transition law of one of the arms, referred to as the odd arm, is different from the common transition law of all other arms. A learner, who has no knowledge of the above transition laws, has to devise a sequential test to identify the index of the odd arm as quickly as possible, subject to an upper bound on the probability of error. For this problem, we derive an asymptotic lower bound on the expected stopping time of any sequential test of the learner. where the asymptotics** schemes in which, at each time, he may choose any one of the arms and observe the current state of the chosen arm. During this time, the unobserved arms do not undergo state transitions and remain frozen at their last observed states. We refer to this as the *rested* arms setting, borrowing the terminology from Gittins [1]. Thus, our problem is one of odd arm identification in a multi-armed bandit setting with rested Markovian arms.

1

Detecting an Odd Restless Markov Arm with a Trembling Hand

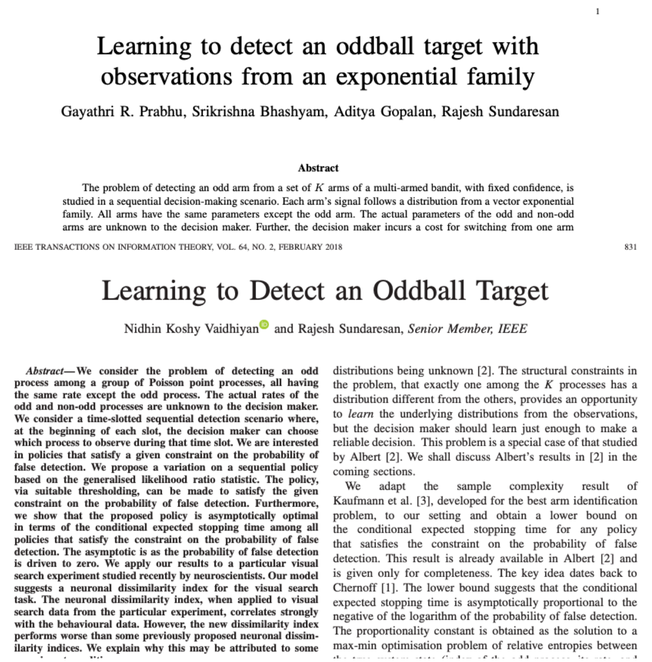
P. N. Karthik and Rajesh Sundaresan

Abstract

In this paper, we consider a multi-armed bandit in which each arm is a Markov process evolving on a finite state space. The state space is common across the arms, and the arms are independent of each other. The transition probability matrix of one of the arms (the odd arm) is different from the common transition probability matrix of all the other arms. A decision maker, who knows these transition probability matrices, wishes to identify the odd arm as quickly as possible, while keeping the probability of decision error small. To do so, the decision maker collects observations from the arms by pulling the arms in a sequential manner, one at each discrete time instant. However, the decision maker has a trembling hand, and the arm that is actually pulled at any given time differs, with a small probability, from the one he intended to pull. The observation at any given time is the arm that is actually pulled and its current state. The Markov processes of the unobserved arms continue to evolve. This makes the arms restless.

For the above setting, we derive the first known asymptotic lower bound on the expected time required to identify the odd arm, where the asymptotics is of vanishing error probability. The continued evolution of each arm adds a new dimension to the problem, leading to a family of Markov decision problems (MDPs) on a countable state space. We then stitch together certain

Accurately $\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$ Quickly



1

Learning to detect an oddball target with observations from an exponential family

Gayathri R. Prabhu, Srikrishna Bhashyam, Aditya Gopalan, Rajesh Sundaresan

Abstract

The problem of detecting an odd arm from a set of K arms of a multi-armed bandit, with fixed confidence, is studied in a sequential decision-making scenario. Each arm's signal follows a distribution from a vector exponential family. All arms have the same parameters except the odd arm. The actual parameters of the odd and non-odd arms are unknown to the decision maker. Further, the decision maker incurs a cost for switching from one arm

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 64, NO. 2, FEBRUARY 2018 831

Learning to Detect an Oddball Target

Nidhin Koshy Vaidhiyan and Rajesh Sundaresan, *Senior Member, IEEE*

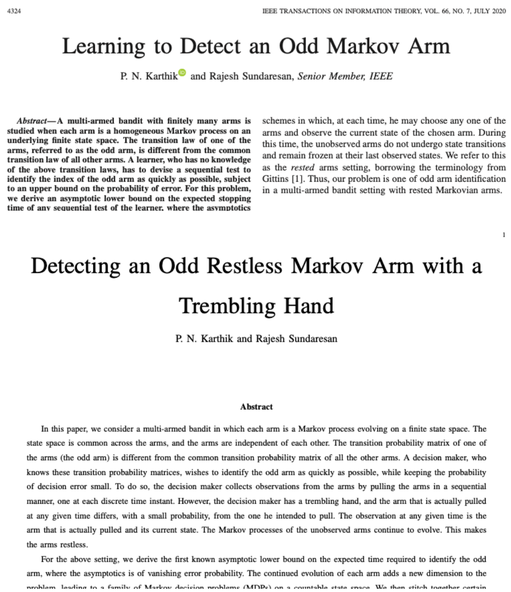
***Abstract*— We consider the problem of detecting an odd process among a group of Poisson point processes, all having the same rate except the odd process. The actual rates of the odd and non-odd processes are unknown to the decision maker. We consider a time-slotted sequential detection scenario where, at the beginning of each slot, the decision maker can choose which process to observe during that time slot. We are interested in policies that satisfy a given constraint on the probability of false detection. We propose a variation on a sequential policy based on the generalised likelihood ratio statistic. The policy, via suitable thresholding, can be made to satisfy the given constraint on the probability of false detection. Furthermore, we show that the proposed policy is asymptotically optimal in terms of the conditional expected stopping time among all policies that satisfy the constraint on the probability of false detection. The asymptotic is as the probability of false detection is driven to zero. We apply our results to a particular visual search experiment studied recently by neuroscientists. Our model suggests a neuronal dissimilarity index for the visual search task. The neuronal dissimilarity index, when applied to visual search data from the particular experiment, correlates strongly with the behavioural data. However, the new dissimilarity index performs worse than some previously proposed neuronal dissimilarity indices. We explain why this may be attributed to some** distributions being unknown [2]. The structural constraints in the problem, that exactly one among the K processes has a distribution different from the others, provides an opportunity to *learn* the underlying distributions from the observations, but the decision maker should learn just enough to make a reliable decision. This problem is a special case of that studied by Albert [2]. We shall discuss Albert's results in [2] in the coming sections.

We adapt the sample complexity result of Kaufmann et al. [3], developed for the best arm identification problem, to our setting and obtain a lower bound on the conditional expected stopping time for any policy that satisfies the constraint on the probability of false detection. This result is already available in Albert [2] and is given only for completeness. The key idea dates back to Chernoff [1]. The lower bound suggests that the conditional expected stopping time is asymptotically proportional to the negative of the logarithm of the probability of false detection. The proportionality constant is obtained as the solution to a max-min optimisation problem of relative entropies between

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

h: Odd arm; P_1: TPM of odd arm; P_2: TPM of each non-odd arm



4324 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 66, NO. 7, JULY 2020

Learning to Detect an Odd Markov Arm

P. N. Karthik and Rajesh Sundaresan, *Senior Member, IEEE*

***Abstract*— A multi-armed bandit with finitely many arms is studied when each arm is a homogeneous Markov process on an underlying finite state space. The transition law of one of the arms, referred to as the odd arm, is different from the common transition law of all other arms. A learner, who has no knowledge of the above transition laws, has to devise a sequential test to identify the index of the odd arm as quickly as possible, subject to an upper bound on the probability of error. For this problem, we derive an asymptotic lower bound on the expected stopping time of any sequential test of the learner. where the asymptotics** schemes in which, at each time, he may choose any one of the arms and observe the current state of the chosen arm. During this time, the unobserved arms do not undergo state transitions and remain frozen at their last observed states. We refer to this as the *rested* arms setting, borrowing the terminology from Gittins [1]. Thus, our problem is one of odd arm identification in a multi-armed bandit setting with rested Markovian arms.

1

Detecting an Odd Restless Markov Arm with a Trembling Hand

P. N. Karthik and Rajesh Sundaresan

Abstract

In this paper, we consider a multi-armed bandit in which each arm is a Markov process evolving on a finite state space. The state space is common across the arms, and the arms are independent of each other. The transition probability matrix of one of the arms (the odd arm) is different from the common transition probability matrix of all the other arms. A decision maker, who knows these transition probability matrices, wishes to identify the odd arm as quickly as possible, while keeping the probability of decision error small. To do so, the decision maker collects observations from the arms by pulling the arms in a sequential manner, one at each discrete time instant. However, the decision maker has a trembling hand, and the arm that is actually pulled at any given time differs, with a small probability, from the one he intended to pull. The observation at any given time is the arm that is actually pulled and its current state. The Markov processes of the unobserved arms continue to evolve. This makes the arms restless.

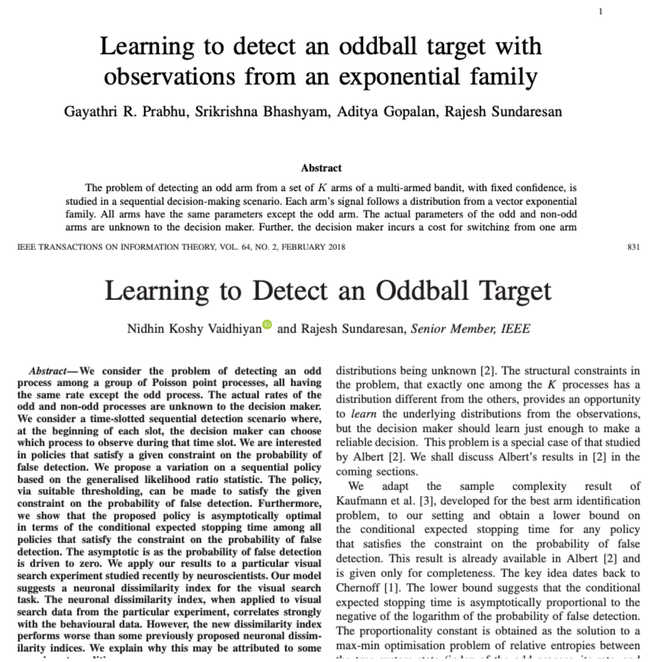
For the above setting, we derive the first known asymptotic lower bound on the expected time required to identify the odd arm, where the asymptotics is of vanishing error probability. The continued evolution of each arm adds a new dimension to the problem, leading to a family of Markov decision problems (MDPs) on a countable state space. We then stitch together certain

Accurately

$$\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$$

Quickly

$$\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C] \approx O\left(\log \frac{1}{\epsilon}\right)$$



1

Learning to detect an oddball target with observations from an exponential family

Gayathri R. Prabhu, Srikrishna Bhashyam, Aditya Gopalan, Rajesh Sundaresan

Abstract

The problem of detecting an odd arm from a set of K arms of a multi-armed bandit, with fixed confidence, is studied in a sequential decision-making scenario. Each arm's signal follows a distribution from a vector exponential family. All arms have the same parameters except the odd arm. The actual parameters of the odd and non-odd arms are unknown to the decision maker. Further, the decision maker incurs a cost for switching from one arm

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 64, NO. 2, FEBRUARY 2018 831

Learning to Detect an Oddball Target

Nidhin Koshy Vaidhiyan and Rajesh Sundaresan, *Senior Member, IEEE*

***Abstract*— We consider the problem of detecting an odd process among a group of Poisson point processes, all having the same rate except the odd process. The actual rates of the odd and non-odd processes are unknown to the decision maker. We consider a time-slotted sequential detection scenario where, at the beginning of each slot, the decision maker can choose which process to observe during that time slot. We are interested in policies that satisfy a given constraint on the probability of false detection. We propose a variation on a sequential policy based on the generalised likelihood ratio statistic. The policy, via suitable thresholding, can be made to satisfy the given constraint on the probability of false detection. Furthermore, we show that the proposed policy is asymptotically optimal in terms of the conditional expected stopping time among all policies that satisfy the constraint on the probability of false detection. The asymptotic is as the probability of false detection is driven to zero. We apply our results to a particular visual search experiment studied recently by neuroscientists. Our model suggests a neuronal dissimilarity index for the visual search task. The neuronal dissimilarity index, when applied to visual search data from the particular experiment, correlates strongly with the behavioural data. However, the new dissimilarity index performs worse than some previously proposed neuronal dissimilarity indices. We explain why this may be attributed to some** distributions being unknown [2]. The structural constraints in the problem, that exactly one among the K processes has a distribution different from the others, provides an opportunity to *learn* the underlying distributions from the observations, but the decision maker should learn just enough to make a reliable decision. This problem is a special case of that studied by Albert [2]. We shall discuss Albert's results in [2] in the coming sections.

We adapt the sample complexity result of Kaufmann et al. [3], developed for the best arm identification problem, to our setting and obtain a lower bound on the conditional expected stopping time for any policy that satisfies the constraint on the probability of false detection. This result is already available in Albert [2] and is given only for completeness. The key idea dates back to Chernoff [1]. The lower bound suggests that the conditional expected stopping time is asymptotically proportional to the negative of the logarithm of the probability of false detection. The proportionality constant is obtained as the solution to a max-min optimisation problem of relative entropies between

Objective

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

h: Odd arm; P_1: TPM of odd arm; P_2: TPM of each non-odd arm

4324 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 66, NO. 7, JULY 2020

Learning to Detect an Odd Markov Arm

P. N. Karthik and Rajesh Sundaresan, *Senior Member, IEEE*

***Abstract*— A multi-armed bandit with finitely many arms is studied when each arm is a homogeneous Markov process on an underlying finite state space. The transition law of one of the arms, referred to as the odd arm, is different from the common transition law of all other arms. A learner, who has no knowledge of the above transition laws, has to devise a sequential test to identify the index of the odd arm as quickly as possible, subject to an upper bound on the probability of error. For this problem, we derive an asymptotic lower bound on the expected stopping time of any sequential test of the learner. where the asymptotics** schemes in which, at each time, he may choose any one of the arms and observe the current state of the chosen arm. During this time, the unobserved arms do not undergo state transitions and remain frozen at their last observed states. We refer to this as the *rested* arms setting, borrowing the terminology from Gittins [1]. Thus, our problem is one of odd arm identification in a multi-armed bandit setting with rested Markovian arms.

1

Detecting an Odd Restless Markov Arm with a Trembling Hand

P. N. Karthik and Rajesh Sundaresan

Abstract

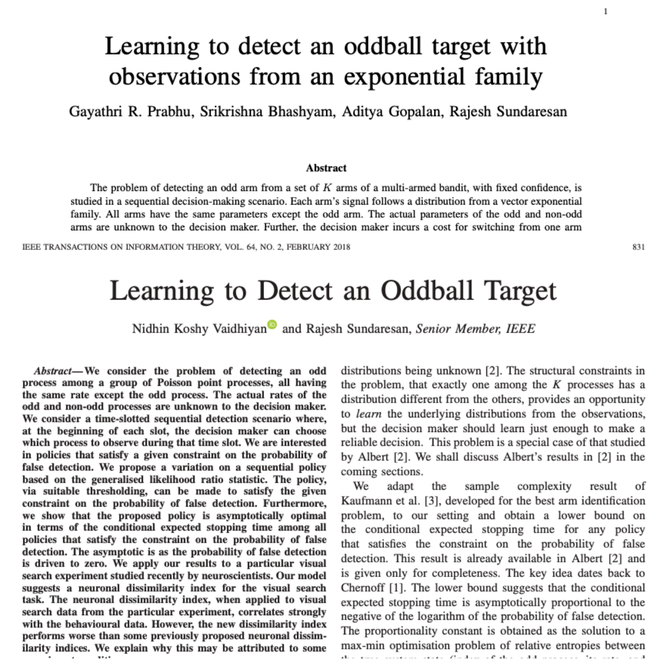
In this paper, we consider a multi-armed bandit in which each arm is a Markov process evolving on a finite state space. The state space is common across the arms, and the arms are independent of each other. The transition probability matrix of one of the arms (the odd arm) is different from the common transition probability matrix of all the other arms. A decision maker, who knows these transition probability matrices, wishes to identify the odd arm as quickly as possible, while keeping the probability of decision error small. To do so, the decision maker collects observations from the arms by pulling the arms in a sequential manner, one at each discrete time instant. However, the decision maker has a trembling hand, and the arm that is actually pulled at any given time differs, with a small probability, from the one he intended to pull. The observation at any given time is the arm that is actually pulled and its current state. The Markov processes of the unobserved arms continue to evolve. This makes the arms restless.

For the above setting, we derive the first known asymptotic lower bound on the expected time required to identify the odd arm, where the asymptotics is of vanishing error probability. The continued evolution of each arm adds a new dimension to the problem, leading to a family of Markov decision problems (MDPs) on a countable state space. We then stitch together certain

Accurately $\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$ Quickly

$$\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C] \approx O\left(\log \frac{1}{\epsilon}\right)$$

Characterise $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$



1

Learning to detect an oddball target with observations from an exponential family

Gayathri R. Prabhu, Srikrishna Bhashyam, Aditya Gopalan, Rajesh Sundaresan

Abstract

The problem of detecting an odd arm from a set of K arms of a multi-armed bandit, with fixed confidence, is studied in a sequential decision-making scenario. Each arm's signal follows a distribution from a vector exponential family. All arms have the same parameters except the odd arm. The actual parameters of the odd and non-odd arms are unknown to the decision maker. Further, the decision maker incurs a cost for switching from one arm

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 64, NO. 2, FEBRUARY 2018 831

Learning to Detect an Oddball Target

Nidhin Koshy Vaidhiyan and Rajesh Sundaresan, *Senior Member, IEEE*

***Abstract*— We consider the problem of detecting an odd process among a group of Poisson point processes, all having the same rate except the odd process. The actual rates of the odd and non-odd processes are unknown to the decision maker. We consider a time-slotted sequential detection scenario where, at the beginning of each slot, the decision maker can choose which process to observe during that time slot. We are interested in policies that satisfy a given constraint on the probability of false detection. We propose a variation on a sequential policy based on the generalised likelihood ratio statistic. The policy, via suitable thresholding, can be made to satisfy the given constraint on the probability of false detection. Furthermore, we show that the proposed policy is asymptotically optimal in terms of the conditional expected stopping time among all policies that satisfy the constraint on the probability of false detection. The asymptotic is as the probability of false detection is driven to zero. We apply our results to a particular visual search experiment studied recently by neuroscientists. Our model suggests a neuronal dissimilarity index for the visual search task. The neuronal dissimilarity index, when applied to visual search data from the particular experiment, correlates strongly with the behavioural data. However, the new dissimilarity index performs worse than some previously proposed neuronal dissimilarity indices. We explain why this may be attributed to some** distributions being unknown [2]. The structural constraints in the problem, that exactly one among the K processes has a distribution different from the others, provides an opportunity to *learn* the underlying distributions from the observations, but the decision maker should learn just enough to make a reliable decision. This problem is a special case of that studied by Albert [2]. We shall discuss Albert's results in [2] in the coming sections.

We adapt the sample complexity result of Kaufmann et al. [3], developed for the best arm identification problem, to our setting and obtain a lower bound on the conditional expected stopping time for any policy that satisfies the constraint on the probability of false detection. This result is already available in Albert [2] and is given only for completeness. The key idea dates back to Chernoff [1]. The lower bound suggests that the conditional expected stopping time is asymptotically proportional to the negative of the logarithm of the probability of false detection. The proportionality constant is obtained as the solution to a max-min optimisation problem of relative entropies between

Preliminaries

Trembling Hand, Markov Decision Problem, SRS Policy

Trembling Hand

Trembling Hand

- Often in visual search experiments, at each time t, the actual focus location (A_t) differs from the intended focus location (B_t) with small probability

Trembling Hand

- Often in visual search experiments, at each time t, the actual focus location (A_t) differs from the intended focus location (B_t) with small probability
- This can be captured as a trembling hand:

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

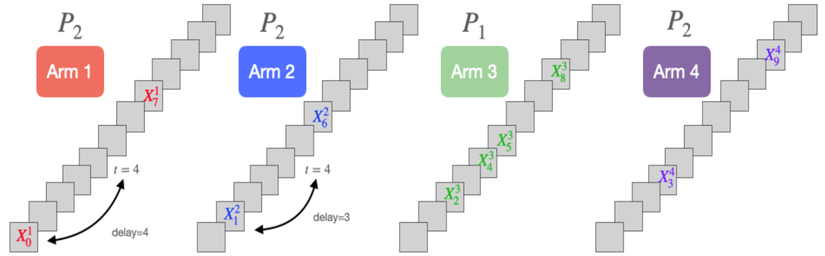
Trembling Hand

- Often in visual search experiments, at each time t, the actual focus location (A_t) differs from the intended focus location (B_t) with small probability
- This can be captured as a trembling hand:

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

- $\eta \in (0,1]$: trembling hand parameter

Delays and Last Observed States

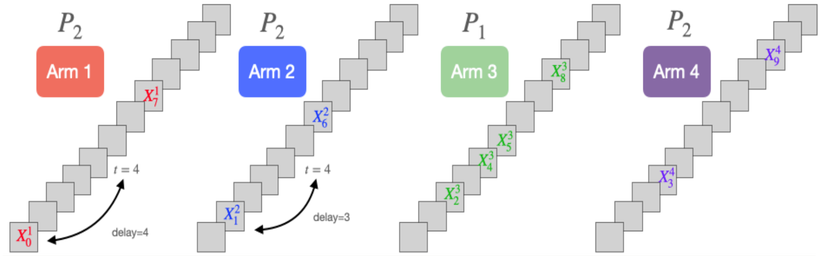


t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_8^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t

Delays and Last Observed States

- $\underline{d}(t) = (d_1(t), \ldots, d_K(t))$ $\quad\quad$ $\underline{i}(t) = (i_1(t), \ldots, i_K(t))$

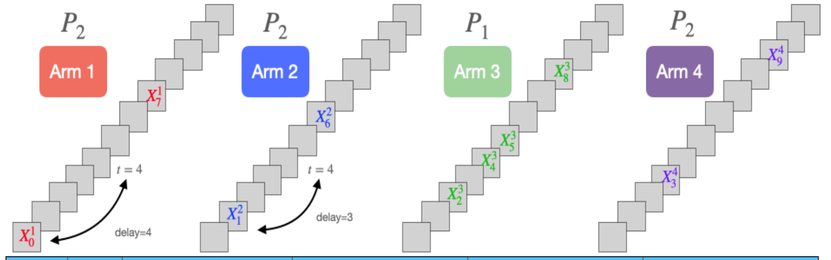


t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_8^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t

Delays and Last Observed States

- $\underline{d}(t) = (d_1(t), \ldots, d_K(t))$ $\qquad\qquad$ $\underline{i}(t) = (i_1(t), \ldots, i_K(t))$
- $(B_0, A_0, X_0^{A_0}, \ldots, B_{t-1}, A_{t-1}, X_{t-1}^{A_{t-1}}) \equiv (B_0, \ldots, B_{t-1}, \{\underline{d}(s), \underline{i}(s) : K \leq s \leq t\})$

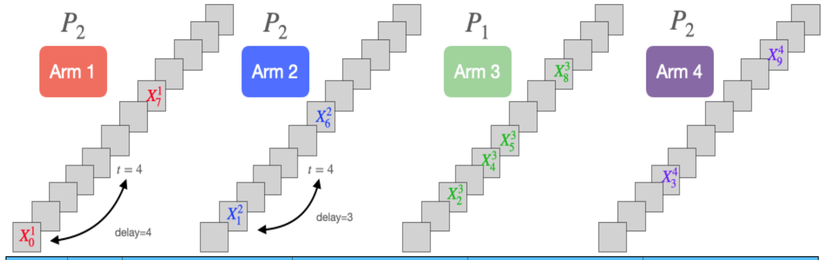


t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_8^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\qquad$ $d_a(t)$: delay of arm a at time t $\qquad$ $i_a(t)$: last observed state of arm a at time t

Delays and Last Observed States

- $\underline{d}(t) = (d_1(t), \ldots, d_K(t))$ $\qquad\qquad$ $\underline{i}(t) = (i_1(t), \ldots, i_K(t))$
- $(B_0, A_0, X_0^{A_0}, \ldots, B_{t-1}, A_{t-1}, X_{t-1}^{A_{t-1}}) \equiv (B_0, \ldots, B_{t-1}, \{\underline{d}(s), \underline{i}(s) : K \le s \le t\})$



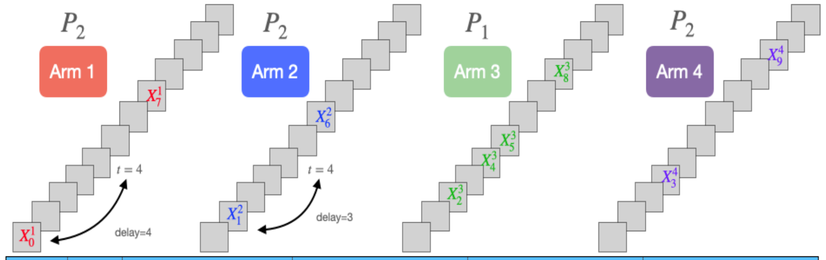
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_8^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

Delays and Last Observed States

- $\underline{d}(t) = (d_1(t), \ldots, d_K(t))$ $\qquad\qquad$ $\underline{i}(t) = (i_1(t), \ldots, i_K(t))$
- $(B_0, A_0, X_0^{A_0}, \ldots, B_{t-1}, A_{t-1}, X_{t-1}^{A_{t-1}}) \equiv (B_0, \ldots, B_{t-1}, \{\underline{d}(s), \underline{i}(s) : K \le s \le t\})$
- $\{(\underline{d}(t), \underline{i}(t)) : t \ge K\}$ is a controlled Markov process with controls $\{B_t : t \ge 0\}$



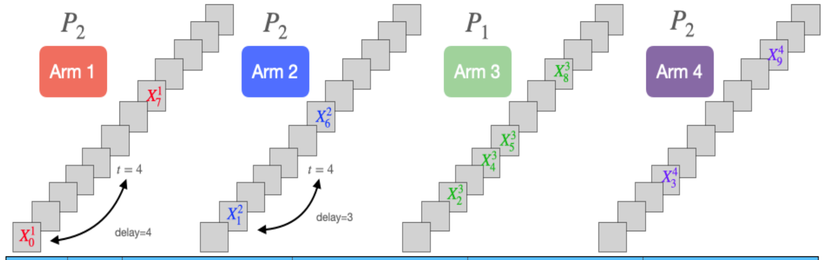
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\qquad$ $d_a(t)$: delay of arm a at time t $\qquad$ $i_a(t)$: last observed state of arm a at time t

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

Delays and Last Observed States

- $\underline{d}(t) = (d_1(t), \ldots, d_K(t))$ $\qquad\qquad$ $\underline{i}(t) = (i_1(t), \ldots, i_K(t))$
- $(B_0, A_0, X_0^{A_0}, \ldots, B_{t-1}, A_{t-1}, X_{t-1}^{A_{t-1}}) \equiv (B_0, \ldots, B_{t-1}, \{\underline{d}(s), \underline{i}(s) : K \le s \le t\})$
- $\{(\underline{d}(t), \underline{i}(t)) : t \ge K\}$ is a controlled Markov process with controls $\{B_t : t \ge 0\}$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t

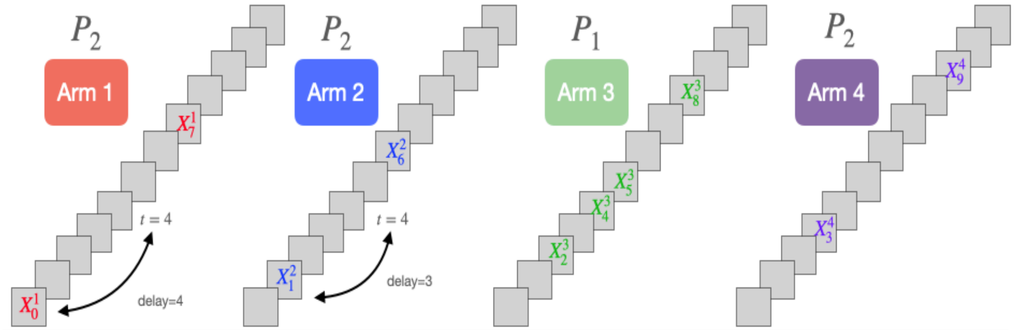
$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

$$P(\underline{d}(t+1), \underline{i}(t+1) \,|\, B_0, \ldots, B_t, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) = P(\underline{d}(t+1), \underline{i}(t+1) \,|\, B_t, (\underline{d}(t), \underline{i}(t)))$$

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) \rightarrow B_t \rightarrow (\underline{d}(t+1), \underline{i}(t+1))$$

State space	
Action space	
State at time t	
Action at time t	



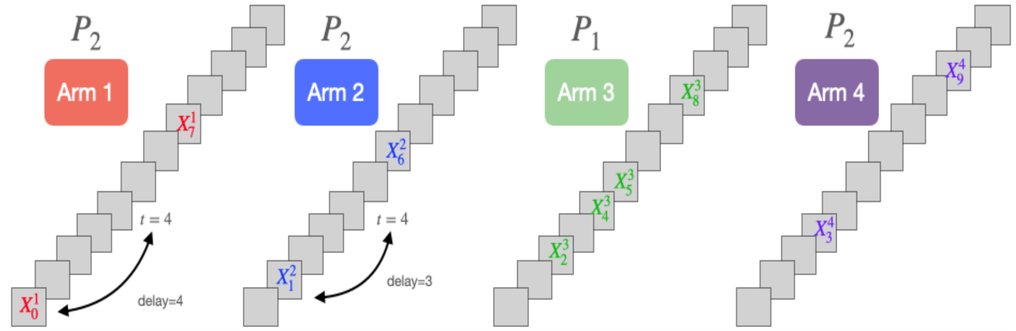
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_1^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_2^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) \rightarrow B_t \rightarrow (\underline{d}(t+1), \underline{i}(t+1))$$

State space	$\mathbb{S} = \{(\underline{d}, \underline{i})\}$
Action space	
State at time t	
Action at time t	



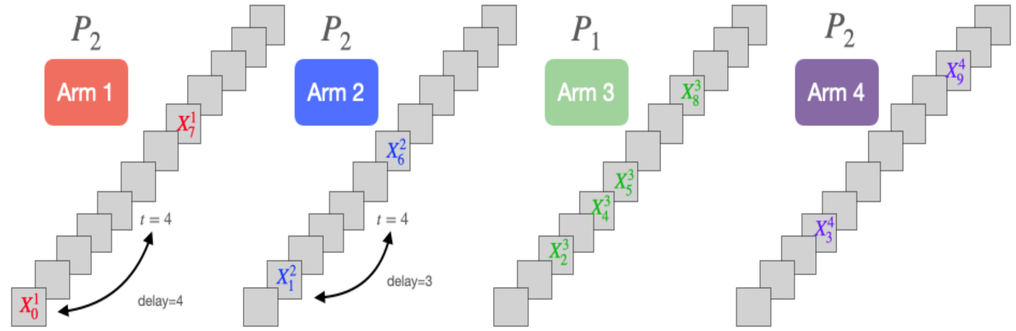
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_1^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) \rightarrow B_t \rightarrow (\underline{d}(t+1), \underline{i}(t+1))$$

State space	$\mathbb{S} = \{(\underline{d}, \underline{i})\}$
Action space	$\{1, \ldots, K\}$
State at time t	
Action at time t	



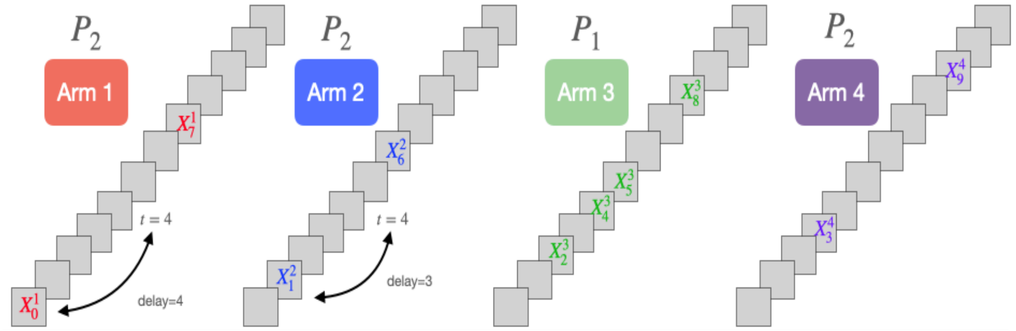
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_1^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) \rightarrow B_t \rightarrow (\underline{d}(t+1), \underline{i}(t+1))$$

State space	$\mathbb{S} = \{(\underline{d}, \underline{i})\}$
Action space	$\{1, \ldots, K\}$
State at time t	$(\underline{d}(t), \underline{i}(t))$
Action at time t	



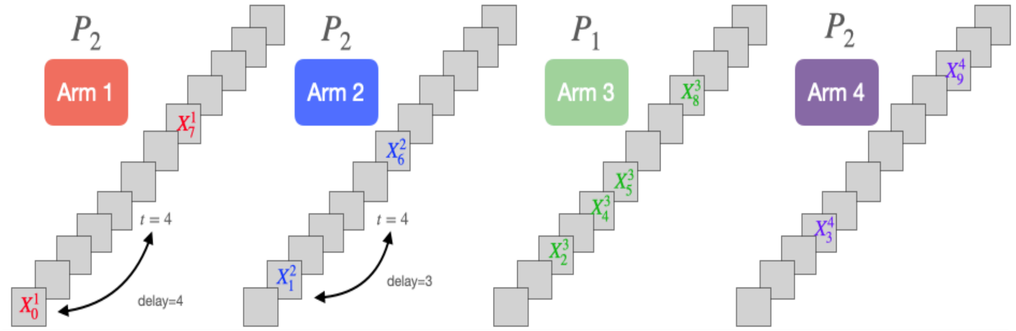
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

State space	$\mathbb{S} = \{(\underline{d}, \underline{i})\}$
Action space	$\{1, \ldots, K\}$
State at time t	$(\underline{d}(t), \underline{i}(t))$
Action at time t	B_t



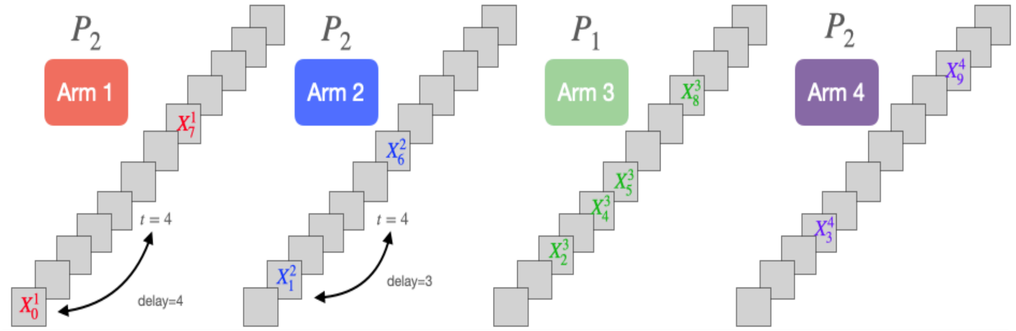
t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_1^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_2^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) \rightarrow B_t \rightarrow (\underline{d}(t+1), \underline{i}(t+1))$$

State space	$\mathbb{S} = \{(\underline{d}, \underline{i})\}$
Action space	$\{1, \ldots, K\}$
State at time t	$(\underline{d}(t), \underline{i}(t))$
Action at time t	B_t



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_1^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_2^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

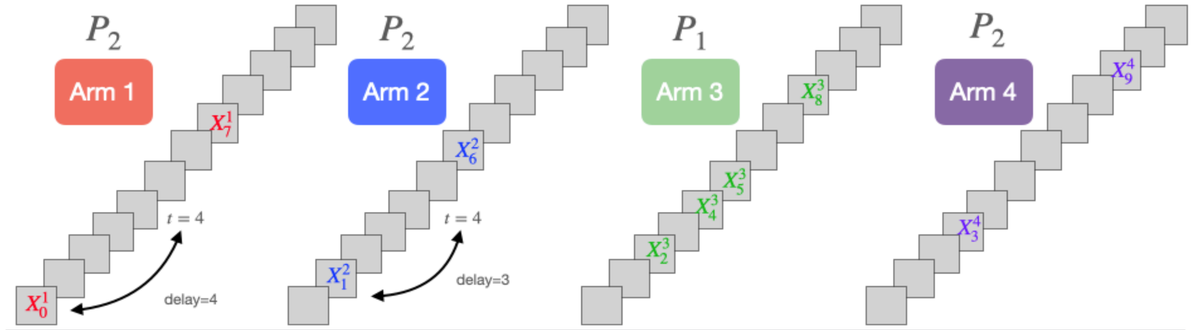
X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

Characterise $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

MDP Transition Probabilities



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

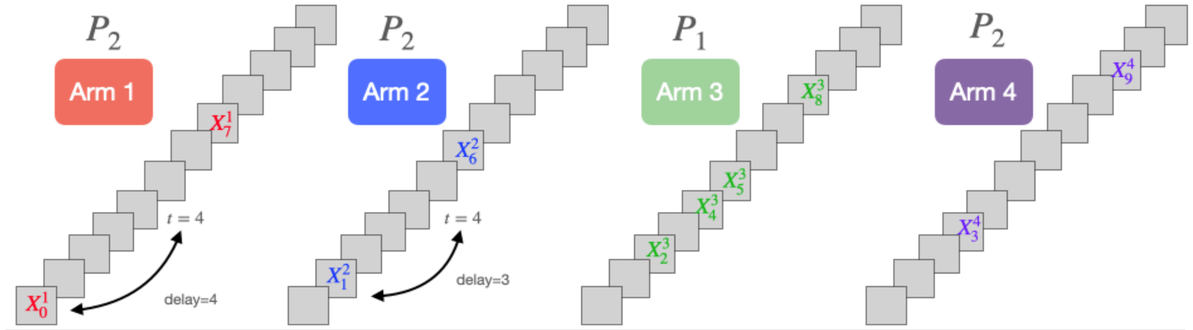
Characterise $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) \rightarrow B_t \rightarrow (\underline{d}(t+1), \underline{i}(t+1))$$

MDP Transition Probabilities

$\underline{d}(t) = \underline{d} = (4,3,2,1)$ $\quad\quad$ $\underline{i}(t) = \underline{i} = (i_1, i_2, i_3, i_4)$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t

Characterise $\quad \lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \dfrac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$

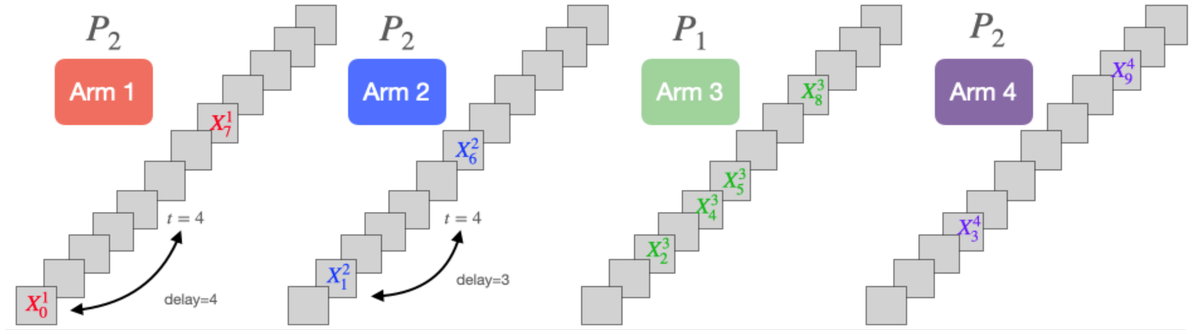
Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) \rightarrow B_t \rightarrow (\underline{d}(t+1), \underline{i}(t+1))$$

MDP Transition Probabilities

$\underline{d}(t) = \underline{d} = (4,3,2,1)$ $\qquad$ $\underline{i}(t) = \underline{i} = (i_1, i_2, i_3, i_4)$

$B_t = b$ $\qquad$ $A_t = 1$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\qquad$ $d_a(t)$: delay of arm a at time t $\qquad$ $i_a(t)$: last observed state of arm a at time t

Characterise $\qquad \lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \dfrac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$

Markov Decision Problem (MDP)

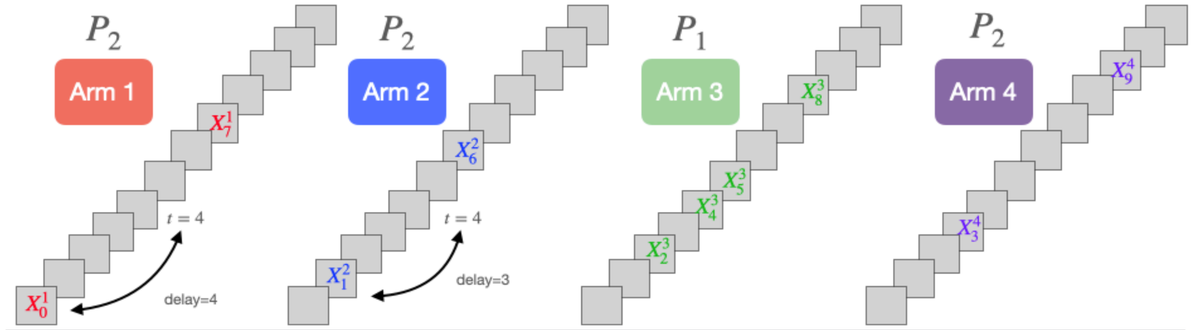
$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

MDP Transition Probabilities

$$\underline{d}(t) = \underline{d} = (4,3,2,1) \qquad \underline{i}(t) = \underline{i} = (i_1, i_2, i_3, i_4)$$

$$B_t = b \qquad A_t = 1$$

$$P(A_t = 1 \,|\, B_t = b) = \frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=1\}}$$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

Characterise $\displaystyle\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \,|\, C]}{\log(1/\epsilon)}$

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

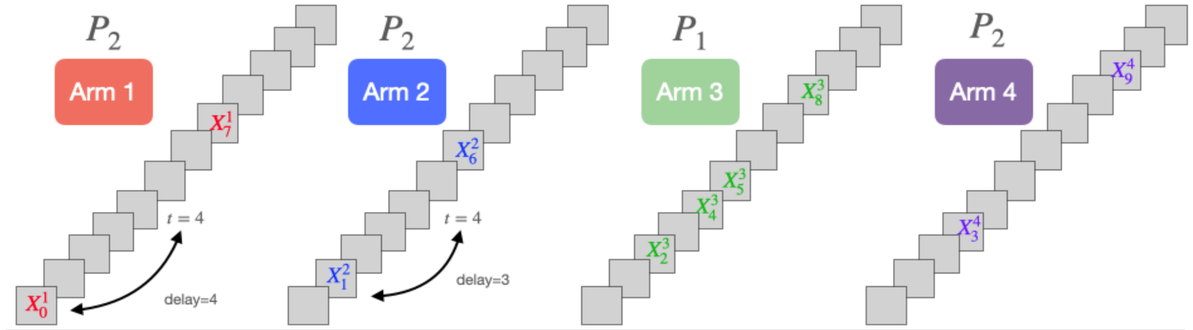
MDP Transition Probabilities

$$\underline{d}(t) = \underline{d} = (4,3,2,1) \qquad \underline{i}(t) = \underline{i} = (i_1, i_2, i_3, i_4)$$

$$B_t = b \qquad A_t = 1$$

$$P(A_t = 1 \,|\, B_t = b) = \frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=1\}}$$

$$\underline{d}(t+1) = \underline{d}' = (1,4,3,2) \qquad \underline{i}(t+1) = \underline{i}' = (X_t^1, i_2, i_3, i_4)$$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t

Characterise $\displaystyle\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \,|\, C]}{\log(1/\epsilon)}$

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

MDP Transition Probabilities

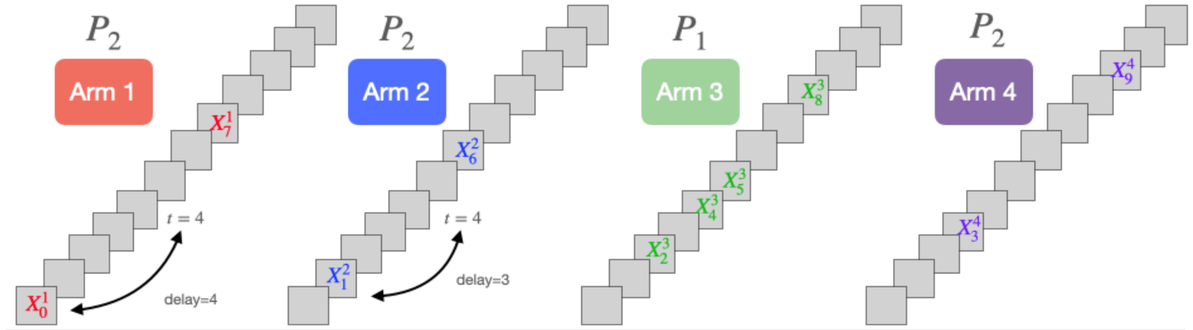
$$\underline{d}(t) = \underline{d} = (4,3,2,1) \qquad \underline{i}(t) = \underline{i} = (i_1, i_2, i_3, i_4)$$

$$B_t = b \qquad A_t = 1$$

$$P(A_t = 1 \,|\, B_t = b) = \frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=1\}}$$

$$\underline{d}(t+1) = \underline{d}' = (1,4,3,2) \qquad \underline{i}(t+1) = \underline{i}' = (X_t^1, i_2, i_3, i_4)$$

$$P(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' \,|\, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, B_t = b) = \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=1\}}\right) (P_2)^4(X_t^1 \,|\, i_1)$$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

Characterise $\quad \lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \,|\, C]}{\log(1/\epsilon)}$

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

MDP Transition Probabilities

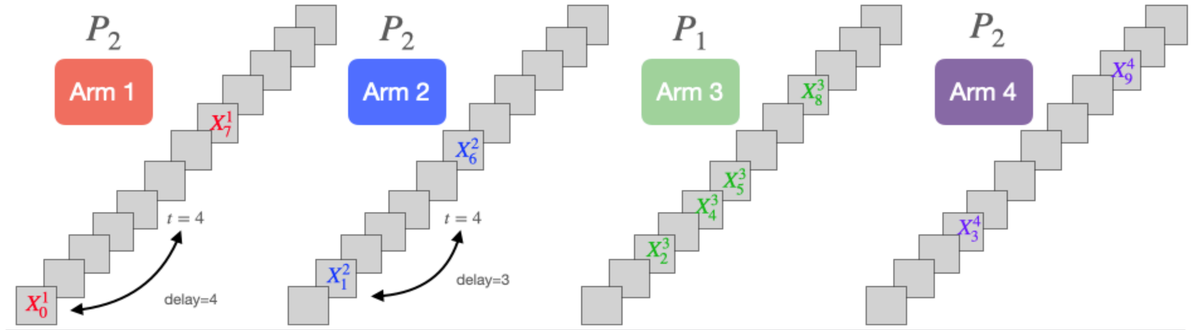
$$\underline{d}(t) = \underline{d} = (4,3,2,1) \qquad \underline{i}(t) = \underline{i} = (i_1, i_2, i_3, i_4)$$

$$B_t = b \qquad A_t = 1$$

$$P(A_t = 1 \,|\, B_t = b) = \frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=1\}}$$

$$\underline{d}(t+1) = \underline{d}' = (1,4,3,2) \qquad \underline{i}(t+1) = \underline{i}' = (X_t^1, i_2, i_3, i_4)$$

$$P(\underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}' \,|\, \underline{d}(t) = \underline{d}, \underline{i}(t) = \underline{i}, B_t = b) = \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b=1\}}\right) (P_2)^4 [X_t^1 \,|\, i_1)$$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t

Characterise $\quad \lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \,|\, C]}{\log(1/\epsilon)}$

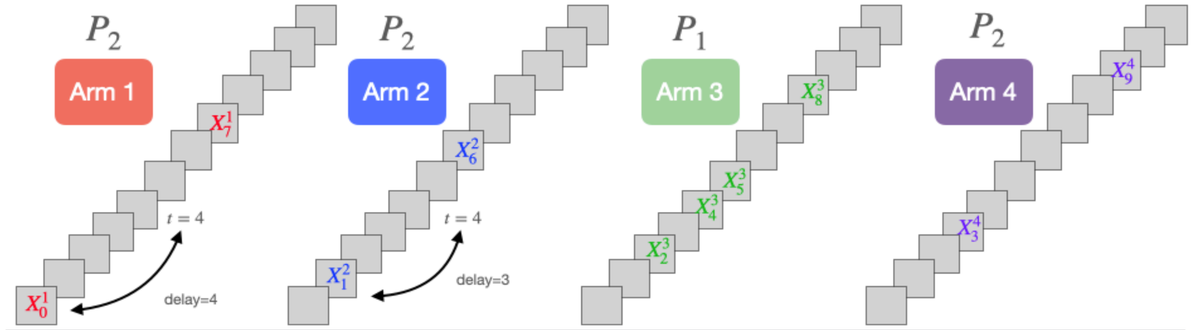
Fourth power of the TPM P_2

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) \rightarrow B_t \rightarrow (\underline{d}(t+1), \underline{i}(t+1))$$

MDP Transition Probabilities

$\underline{d}(t+1) = \underline{d}' = (1,4,3,2)$ $\quad$ $\underline{i}(t+1) = \underline{i}' = (X_t^1, i_2, i_3, i_4)$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad$ $d_a(t)$: delay of arm a at time t $\quad$ $i_a(t)$: last observed state of arm a at time t

Characterise $\quad \lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \dfrac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$

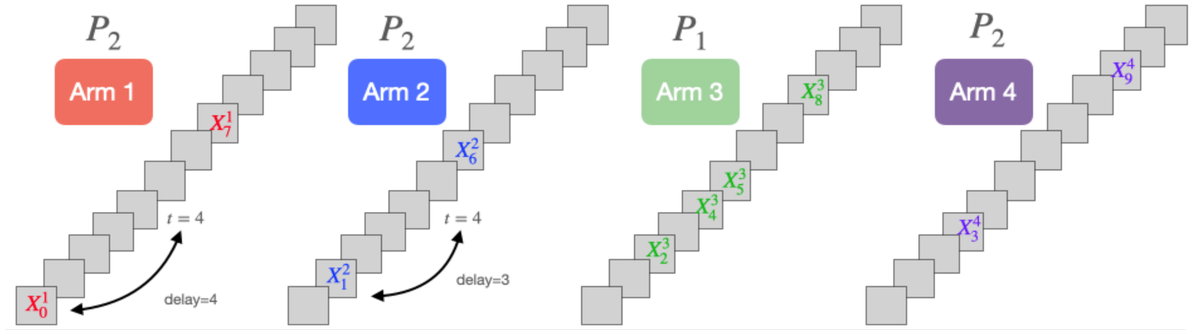
Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) \rightarrow B_t \rightarrow (\underline{d}(t+1), \underline{i}(t+1))$$

MDP Transition Probabilities

$\underline{d}(t+1) = \underline{d}' = (1,4,3,2)$ $\qquad$ $\underline{i}(t+1) = \underline{i}' = (X_t^1, i_2, i_3, i_4)$

$B_{t+1} = b'$ $\qquad$ $A_{t+1} = 3$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\qquad$ $d_a(t)$: delay of arm a at time t $\qquad$ $i_a(t)$: last observed state of arm a at time t

Characterise $\qquad \lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \dfrac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$

Markov Decision Problem (MDP)

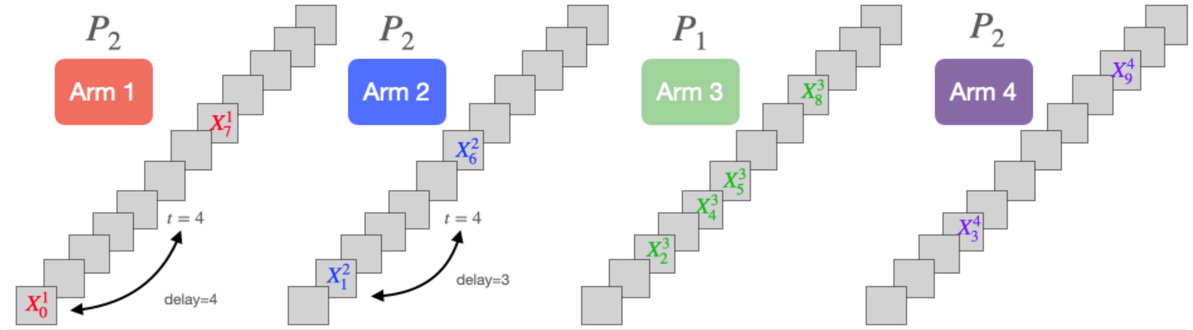
$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

MDP Transition Probabilities

$$\underline{d}(t+1) = \underline{d}' = (1,4,3,2) \qquad \underline{i}(t+1) = \underline{i}' = (X_t^1, i_2, i_3, i_4)$$

$$B_{t+1} = b' \qquad A_{t+1} = 3$$

$$P(A_{t+1} = 3 \,|\, B_{t+1} = b') = \frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b'=3\}}$$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

Characterise $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \dfrac{E[\tau(\pi) \,|\, C]}{\log(1/\epsilon)}$

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

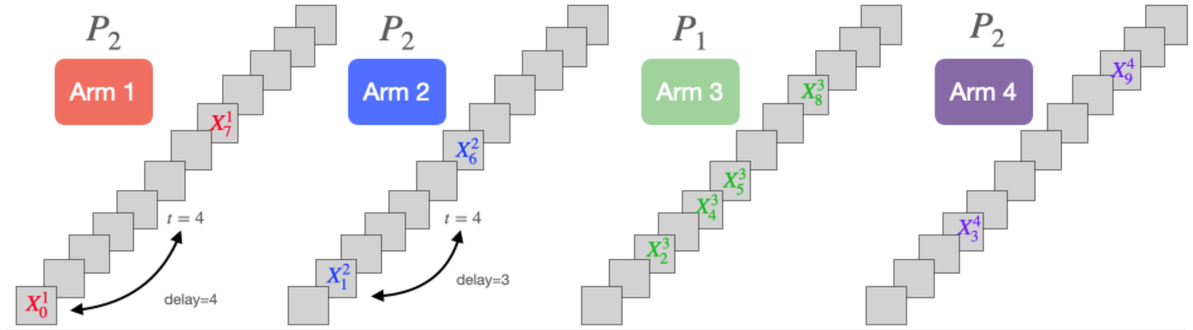
MDP Transition Probabilities

$$\underline{d}(t+1) = \underline{d}' = (1,4,3,2) \qquad \underline{i}(t+1) = \underline{i}' = (X_t^1, i_2, i_3, i_4)$$

$$B_{t+1} = b' \qquad A_{t+1} = 3$$

$$P(A_{t+1} = 3 \,|\, B_{t+1} = b') = \frac{\eta}{K} + (1-\eta)\, \mathbb{I}_{\{b'=3\}}$$

$$\underline{d}(t+2) = \underline{d}'' = (2,5,1,3) \qquad \underline{i}(t+2) = \underline{i}'' = (X_t^1, i_2, X_{t+1}^3, i_4)$$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

Characterise $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \,|\, C]}{\log(1/\epsilon)}$

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

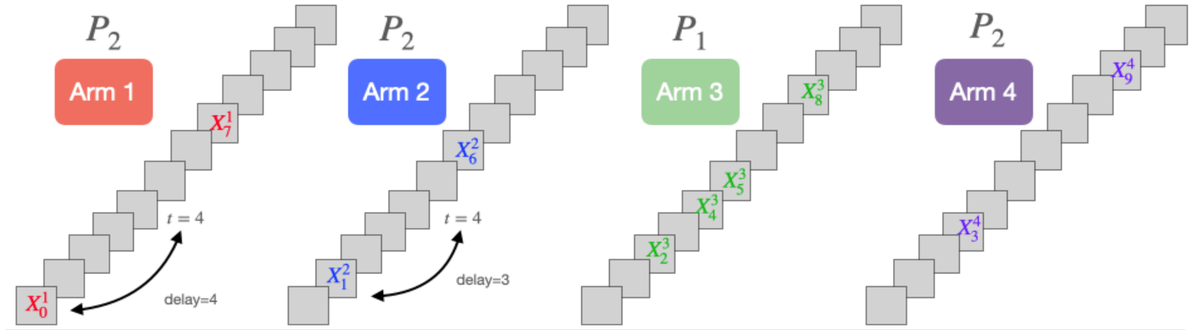
MDP Transition Probabilities

$$\underline{d}(t+1) = \underline{d}' = (1,4,3,2) \qquad \underline{i}(t+1) = \underline{i}' = (X_t^1, i_2, i_3, i_4)$$

$$B_{t+1} = b' \qquad A_{t+1} = 3$$

$$P(A_{t+1} = 3 \,|\, B_{t+1} = b') = \frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b'=3\}}$$

$$\underline{d}(t+2) = \underline{d}'' = (2,5,1,3) \qquad \underline{i}(t+2) = \underline{i}'' = (X_t^1, i_2, X_{t+1}^3, i_4)$$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

$$P(\underline{d}(t+2) = \underline{d}'', \underline{i}(t+2) = \underline{i}'' \,|\, \underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}', B_{t+1} = b') = \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b'=3\}}\right) (P_1)^3 (X_{t+1}^3 \,|\, i_3)$$

Characterise $\quad \lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \,|\, C]}{\log(1/\epsilon)}$

Markov Decision Problem (MDP)

$$(B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \le s \le t\}) \to B_t \to (\underline{d}(t+1), \underline{i}(t+1))$$

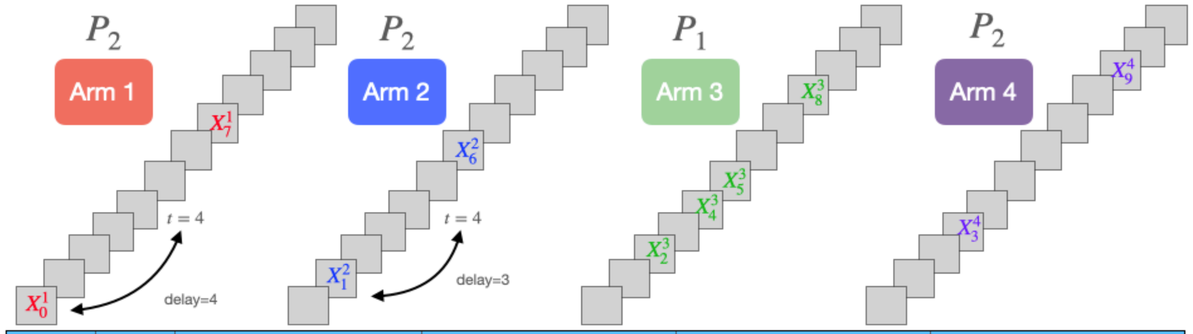
MDP Transition Probabilities

$$\underline{d}(t+1) = \underline{d}' = (1,4,3,2) \qquad \underline{i}(t+1) = \underline{i}' = (X_t^1, i_2, i_3, i_4)$$

$$B_{t+1} = b' \qquad A_{t+1} = 3$$

$$P(A_{t+1} = 3 \,|\, B_{t+1} = b') = \frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b'=3\}}$$

$$\underline{d}(t+2) = \underline{d}'' = (2,5,1,3) \qquad \underline{i}(t+2) = \underline{i}'' = (X_t^1, i_2, X_{t+1}^3, i_4)$$



t	A_t	Arm 1		Arm 2		Arm 3		Arm 4	
		$d_1(t)$	$i_1(t)$	$d_2(t)$	$i_2(t)$	$d_3(t)$	$i_3(t)$	$d_4(t)$	$i_4(t)$
0	1								
1	2								
2	3								
3	4								
4	3	4	X_0^1	3	X_1^2	2	X_2^3	1	X_3^4
5	3	5	X_0^1	4	X_1^2	1	X_4^3	2	X_3^4
6	2	6	X_0^1	5	X_1^2	1	X_5^3	3	X_3^4
7	1	7	X_0^1	1	X_6^2	2	X_5^3	4	X_3^4
8	3	1	X_7^1	2	X_6^2	3	X_5^3	5	X_3^4
9		2	X_7^1	3	X_6^2	1	X_8^3	6	X_3^4

X_t^a : observation from arm a at time t $\quad d_a(t)$: delay of arm a at time t $\quad i_a(t)$: last observed state of arm a at time t

$$P(\underline{d}(t+2) = \underline{d}'', \underline{i}(t+2) = \underline{i}'' \,|\, \underline{d}(t+1) = \underline{d}', \underline{i}(t+1) = \underline{i}', B_{t+1} = b') = \left(\frac{\eta}{K} + (1-\eta)\,\mathbb{I}_{\{b'=3\}}\right) (P_1)^3 (X_{t+1}^3 \,|\, i_3)$$

Characterise $\quad \lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \dfrac{E[\tau(\pi) \,|\, C]}{\log(1/\epsilon)}$

Third power of the TPM P_1

MDP Transition Probabilities

MDP Transition Probabilities

- The MDP transition probabilities are parameterised by the arms configuration

MDP Transition Probabilities

- The MDP transition probabilities are parameterised by the arms configuration

- The value of the true parameter (underlying arms configuration) is unknown and must be learnt (identification / identifiability)

MDP Transition Probabilities

- The MDP transition probabilities are parameterised by the arms configuration
- The value of the true parameter (underlying arms configuration) is unknown and must be learnt (identification / identifiability)
- The set of all possible parameters is uncountably infinite

SRS Policy

SRS Policy

- π is a stationary randomised strategy (SRS policy in short) if $\exists\ \lambda$ such that

SRS Policy

- π is a stationary randomised strategy (SRS policy in short) if $\exists\ \lambda$ such that

$$P(B_t \mid B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) = \lambda(B_t \mid \underline{d}(t), \underline{i}(t))$$

SRS Policy

- π is a stationary randomised strategy (SRS policy in short) if $\exists\ \lambda$ such that

$$P(B_t \mid B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) = \lambda(B_t \mid \underline{d}(t), \underline{i}(t))$$

- Such an SRS policy will be denoted as π^{λ}

SRS Policy

- π is a stationary randomised strategy (SRS policy in short) if $\exists\ \lambda$ such that

$$P(B_t \mid B_0, \ldots, B_{t-1}, \{(\underline{d}(s), \underline{i}(s)) : K \leq s \leq t\}) = \lambda(B_t \mid \underline{d}(t), \underline{i}(t))$$

- Such an SRS policy will be denoted as π^{λ}
- Π_{SRS} : set of all SRS policies

Converse: Lower Bound

Converse: Lower Bound

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

h: Odd arm; P_1: TPM of odd arm; P_2: TPM of each non-odd arm

Characterise $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$

Accurately $\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$ Quickly

Converse: Lower Bound

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

TPM of odd arm

Odd arm

TPM of each non-odd arm

Characterise $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$

Accurately $\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$ Quickly

Lower Bound : $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$

Converse: Lower Bound

- For a policy π and arms configuration $C = (h, P_1, P_2)$:
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

TPM of odd arm

Odd arm

TPM of each non-odd arm

Characterise $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$

Accurately $\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$ Quickly

Lower Bound : $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$

Captures the hardness of the problem-instance

Converse: Lower Bound

- For a policy π and arms configuration $C = (h, P_1, P_2)$: (P_1: TPM of odd arm; h: Odd arm; P_2: TPM of each non-odd arm)
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

Characterise $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$

Accurately $\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$ Quickly

Lower Bound : $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$

$$R^*(h, P_1, P_2) = \sup_{\pi^\lambda \in \Pi_{SRS}} f(h, P_1, P_2, \lambda)$$

Captures the hardness of the problem-instance

Converse: Lower Bound

- For a policy π and arms configuration $C = (h, P_1, P_2)$: (P_1: TPM of odd arm; h: Odd arm; P_2: TPM of each non-odd arm)
  - $\tau(\pi)$: stopping time
  - $P_{\text{error}}(\pi \mid C)$: error probability π of under the configuration C
  - For $\epsilon \in (0,1)$, let $\Pi(\epsilon) = \{\pi : P_{\text{error}}(\pi \mid C) \leq \epsilon\}$

Characterise $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)}$

Accurately $\inf_{\pi \in \Pi(\epsilon)} E[\tau(\pi) \mid C]$ Quickly

Lower Bound : $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$

$$R^*(h, P_1, P_2) = \sup_{\pi^\lambda \in \Pi_{SRS}} f(h, P_1, P_2, \lambda)$$

Captures the hardness of the problem-instance

Contains terms of the form $D(P_1^d(\cdot \mid i) \| P_2^d(\cdot \mid i))$

Achievability

δ-Optimal Solutions, Two Key Assumptions, Policy, Performance

δ-Optimal Solutions

$$R^*(h, P_1, P_2) = \sup_{\pi^\lambda \in \Pi_{SRS}} f(h, P_1, P_2, \lambda)$$

δ-Optimal Solutions

$$R^*(h, P_1, P_2) = \sup_{\pi^\lambda \in \Pi_{SRS}} f(h, P_1, P_2, \lambda)$$

- Computability of the sup is an issue.
  Q-learning may be needed.

δ-Optimal Solutions

$$R^*(h, P_1, P_2) = \sup_{\pi^\lambda \in \Pi_{SRS}} f(h, P_1, P_2, \lambda)$$

- Computability of the sup is an issue.
  Q-learning may be needed.

- It is not clear if the supremum is attained

δ-Optimal Solutions

$$R^*(h, P_1, P_2) = \sup_{\pi^\lambda \in \Pi_{SRS}} f(h, P_1, P_2, \lambda)$$

- Computability of the sup is an issue.
  Q-learning may be needed.

- It is not clear if the supremum is attained

- For $\delta > 0$, under $C = (h, P_1, P_2)$, let $\lambda_{h,P_1,P_2,\delta}$ be such that

δ-Optimal Solutions

$$R^*(h, P_1, P_2) = \sup_{\pi^\lambda \in \Pi_{SRS}} f(h, P_1, P_2, \lambda)$$

- Computability of the sup is an issue.
  Q-learning may be needed.

- It is not clear if the supremum is attained

- For $\delta > 0$, under $C = (h, P_1, P_2)$, let $\lambda_{h,P_1,P_2,\delta}$ be such that

$$f(h, P_1, P_2, \lambda_{h,P_1,P_2,\delta}) \geq \frac{R^*(h, P_1, P_2)}{1 + \delta}$$

δ-Optimal Solutions

$$R^*(h, P_1, P_2) = \sup_{\pi^\lambda \in \Pi_{SRS}} f(h, P_1, P_2, \lambda)$$

- Computability of the sup is an issue.
  Q-learning may be needed.

- It is not clear if the supremum is attained

- For $\delta > 0$, under $C = (h, P_1, P_2)$, let $\lambda_{h,P_1,P_2,\delta}$ be such that

$$f(h, P_1, P_2, \lambda_{h,P_1,P_2,\delta}) \geq \frac{R^*(h, P_1, P_2)}{1+\delta}$$

δ-optimal solution for $C = (h, P_1, P_2)$

Structure on the δ-Optimal Solutions

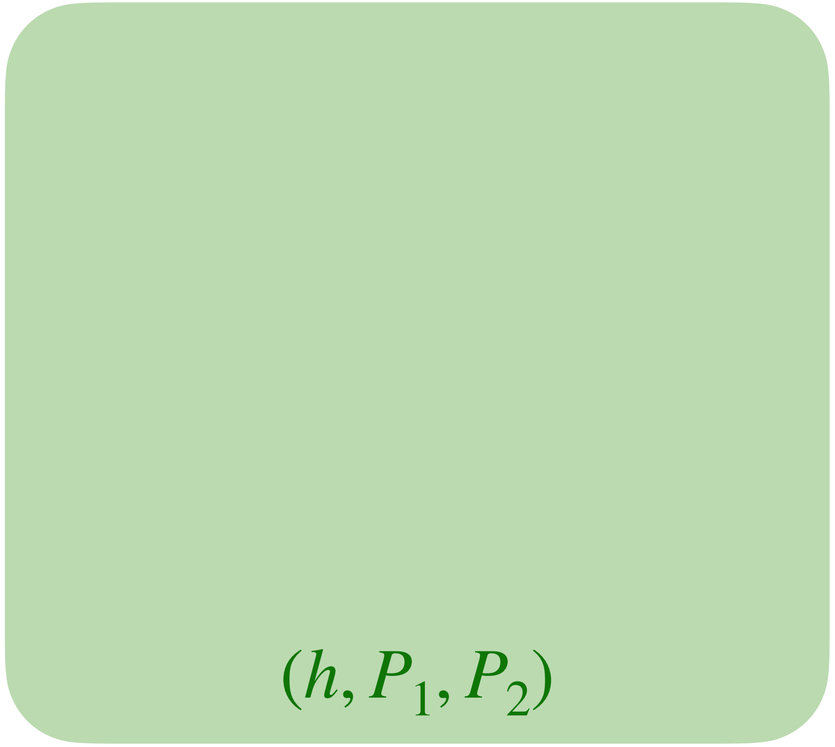
Fix $\delta > 0$

Set of all possible arms configurations (parameters)

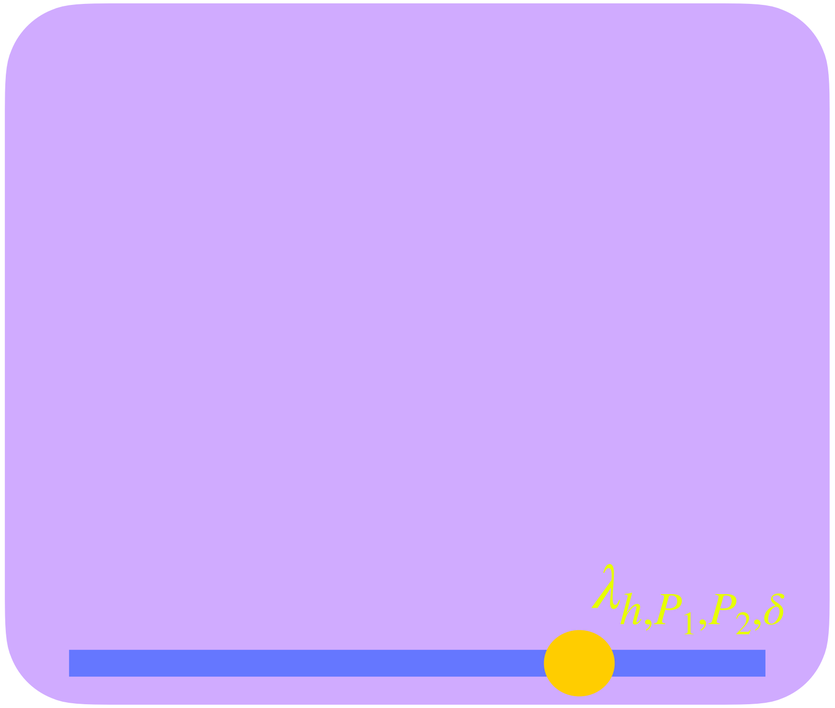
$\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$

Structure on the δ-Optimal Solutions

Fix $\delta > 0$



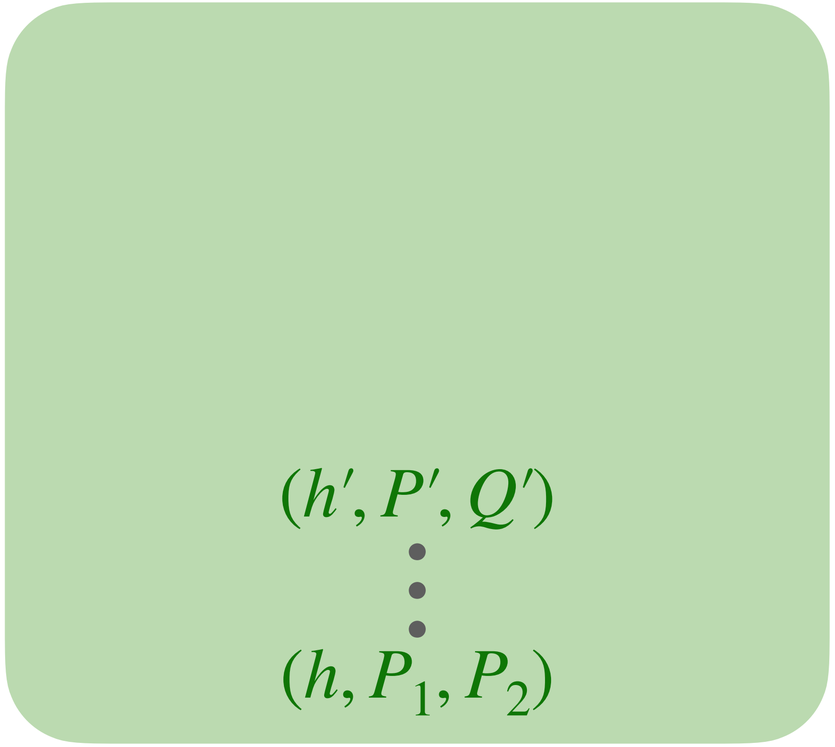
Set of all possible arms configurations (parameters)



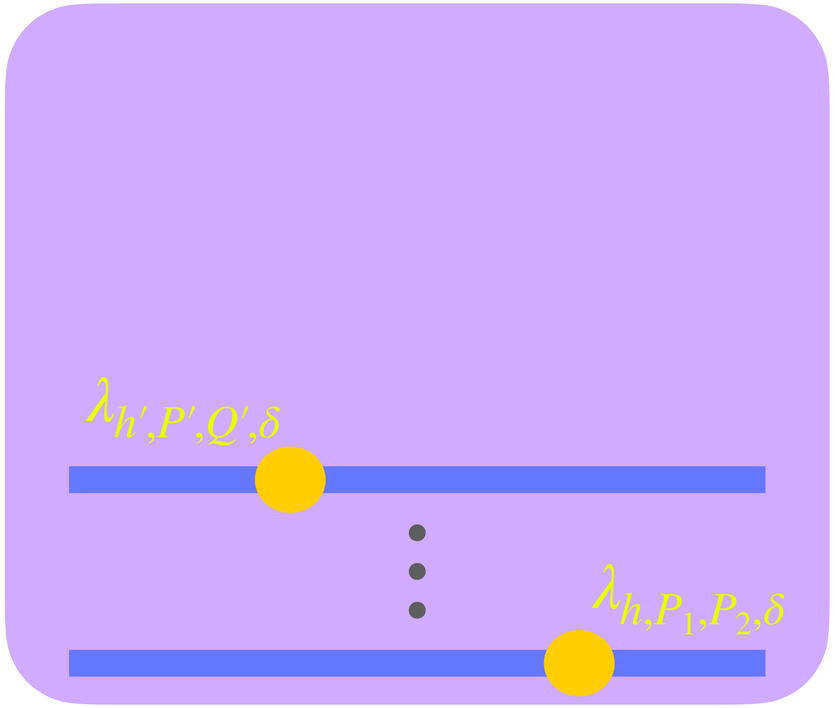
$\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$

Structure on the δ-Optimal Solutions

Fix $\delta > 0$



Set of all possible arms configurations (parameters)



$\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$

Structure on the δ-Optimal Solutions

Fix $\delta > 0$

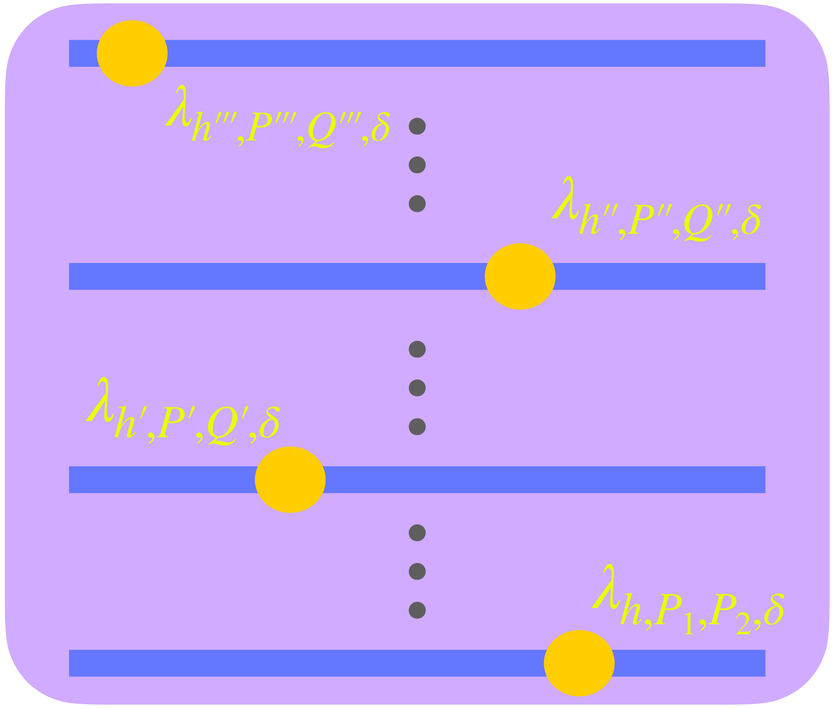
(h''', P''', Q''')

(h'', P'', Q'')

(h', P', Q')

(h, P_1, P_2)

Set of all possible arms configurations (parameters)



$\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$

Structure on the δ-Optimal Solutions

Fix $\delta > 0$

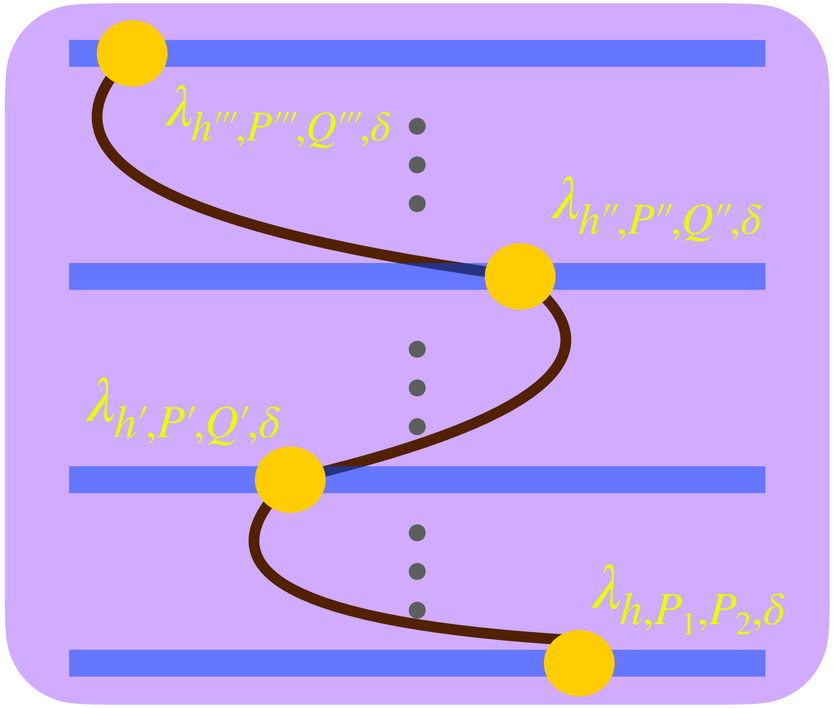
(h''', P''', Q''')

(h'', P'', Q'')

(h', P', Q')

(h, P_1, P_2)

Set of all possible arms configurations (parameters)



$\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$

Structure on the δ-Optimal Solutions

Fix $\delta > 0$

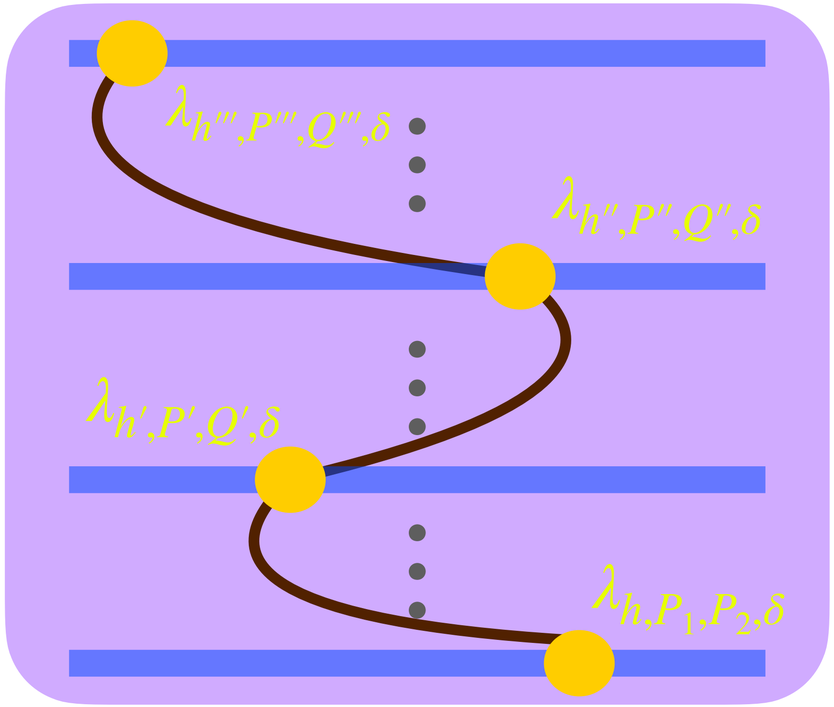
(h''', P''', Q''')

(h'', P'', Q'')

(h', P', Q')

(h, P_1, P_2)

Set of all possible arms configurations (parameters)



$\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$

Convergence of parameters $\Longrightarrow$ convergence of δ-optimal solutions

Two Key Assumptions

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

Two Key Assumptions

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

$$\mathscr{P}(\bar{\varepsilon}^*) = \{P : \ P \text{ is ergodic, } P^d(j \mid i) > 0 \implies P^d(j \mid i) \geq \bar{\varepsilon}^* \text{ for all } d \geq 1,\ i, j\}$$

Two Key Assumptions

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

$$\mathscr{P}(\bar{\varepsilon}^*) = \{P : \ P \text{ is ergodic, } P^d(j \mid i) > 0 \implies P^d(j \mid i) \geq \bar{\varepsilon}^* \text{ for all } d \geq 1,\ i, j\}$$

There exists $\bar{\varepsilon}^* \in (0,1)$ such that for any (h, P, Q), the TPMs $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$

Two Key Assumptions

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

$$\mathscr{P}(\bar{\varepsilon}^*) = \{P : \ P \text{ is ergodic, } P^d(j \mid i) > 0 \implies P^d(j \mid i) \geq \bar{\varepsilon}^* \text{ for all } d \geq 1,\ i, j\}$$

There exists $\bar{\varepsilon}^* \in (0,1)$ such that for any (h, P, Q), the TPMs $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$

$P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$ are harder to distinguish than otherwise

Two Key Assumptions

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

$$\mathscr{P}(\bar{\varepsilon}^*) = \{P : \ P \text{ is ergodic, } P^d(j \mid i) > 0 \implies P^d(j \mid i) \geq \bar{\varepsilon}^* \text{ for all } d \geq 1,\ i, j\}$$

There exists $\bar{\varepsilon}^* \in (0,1)$ such that for any (h, P, Q), the TPMs $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$

$P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$ are harder to distinguish than otherwise

Arbitrary P, Q

$0 \leq D(P^d(\cdot \mid i) \| Q^d(\cdot \mid i)) \leq \infty$

Two Key Assumptions

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h,P,Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

$$\mathscr{P}(\bar{\varepsilon}^*) = \{P : \ P \text{ is ergodic, } P^d(j\,|\,i) > 0 \implies P^d(j\,|\,i) \geq \bar{\varepsilon}^* \text{ for all } d \geq 1,\ i,j\}$$

There exists $\bar{\varepsilon}^* \in (0,1)$ such that for any (h,P,Q), the TPMs $P,Q \in \mathscr{P}(\bar{\varepsilon}^*)$

$P,Q \in \mathscr{P}(\bar{\varepsilon}^*)$ are harder to distinguish than otherwise

Arbitrary P,Q

$$0 \leq D(P^d(\cdot\,|\,i) \| Q^d(\cdot\,|\,i)) \leq \infty$$

$P,Q \in \mathscr{P}(\bar{\varepsilon}^*)$

$$0 \leq D(P^d(\cdot\,|\,i) \| Q^d(\cdot\,|\,i)) \leq \log \frac{1}{\bar{\varepsilon}^*}$$

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

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

- For $n = 0, \ldots, K-1$, sample each of the K arms once

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

- For $n = 0, \ldots, K-1$, sample each of the K arms once
- For all $n \geq K$, repeat the following steps until stoppage:

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

- For $n = 0, \ldots, K-1$, sample each of the K arms once
- For all $n \geq K$, repeat the following steps until stoppage:
  - Compute ML estimates $(\hat{P}_1(n), \hat{P}_2(n))$ of the TPMs

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

- For $n = 0, \ldots, K-1$, sample each of the K arms once
- For all $n \geq K$, repeat the following steps until stoppage:
  - Compute ML estimates $(\hat{P}_1(n), \hat{P}_2(n))$ of the TPMs
  - Let

$$\hat{h}(n) \in \arg\max_h \underbrace{\min_{h' \neq h} \log \frac{\text{avg. likelihood up to time } n \text{ when } h \text{ is the odd arm}}{\text{max. likelihood up to time } n \text{ when } h' \text{ is the odd arm}}}_{M_h(n)}$$

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

- For $n = 0, \ldots, K-1$, sample each of the K arms once
- For all $n \geq K$, repeat the following steps until stoppage:
  - Compute ML estimates $(\hat{P}_1(n), \hat{P}_2(n))$ of the TPMs
  - Let

$$\hat{h}(n) \in \arg\max_{h} \underbrace{\min_{h' \neq h} \log \frac{\text{avg. likelihood up to time } n \text{ when } h \text{ is the odd arm}}{\text{max. likelihood up to time } n \text{ when } h' \text{ is the odd arm}}}_{M_h(n)}$$

  - If $M_{\hat{h}(n)}(n) \geq \log((K-1)L)$, stop and declare $\hat{h}(n)$ as the odd arm

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

- For $n = 0, \ldots, K-1$, sample each of the K arms once
- For all $n \geq K$, repeat the following steps until stoppage:
  - Compute ML estimates $(\hat{P}_1(n), \hat{P}_2(n))$ of the TPMs
  - Let

$$\hat{h}(n) \in \arg\max_{h} \underbrace{\min_{h' \neq h} \log \frac{\text{avg. likelihood up to time } n \text{ when } h \text{ is the odd arm}}{\text{max. likelihood up to time } n \text{ when } h' \text{ is the odd arm}}}_{M_h(n)}$$

  - If $M_{\hat{h}(n)}(n) \geq \log((K-1)L)$, stop and declare $\hat{h}(n)$ as the odd arm
  - Else, sample the next arm according to $\lambda_{\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n), \delta}(\,\cdot \,|\, \underline{d}(n), \underline{i}(n))$

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

- For $n = 0, \ldots, K-1$, sample each of the K arms once
- For all $n \geq K$, repeat the following steps until stoppage:
  - Compute ML estimates $(\hat{P}_1(n), \hat{P}_2(n))$ of the TPMs
  - Let

$$\hat{h}(n) \in \arg\max_{h} \underbrace{\min_{h' \neq h} \log \frac{\text{avg. likelihood up to time } n \text{ when } h \text{ is the odd arm}}{\text{max. likelihood up to time } n \text{ when } h' \text{ is the odd arm}}}_{M_h(n)}$$

  - If $M_{\hat{h}(n)}(n) \geq \log((K-1)L)$, stop and declare $\hat{h}(n)$ as the odd arm
  - Else, sample the next arm according to $\lambda_{\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n), \delta}(\cdot \,|\, \underline{d}(n), \underline{i}(n))$

Principle of certainty equivalence

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

- For $n = 0, \ldots, K-1$, sample each of the K arms once
- For all $n \geq K$, repeat the following steps until stoppage:
  - Compute ML estimates $(\hat{P}_1(n), \hat{P}_2(n))$ of the TPMs
  - Let

$$\hat{h}(n) \in \arg\max_{h} \underbrace{\min_{h' \neq h} \log \frac{\text{avg. likelihood up to time } n \text{ when } h \text{ is the odd arm}}{\text{max. likelihood up to time } n \text{ when } h' \text{ is the odd arm}}}_{M_h(n)}$$

  - If $M_{\hat{h}(n)}(n) \geq \log((K-1)L)$, stop and declare $\hat{h}(n)$ as the odd arm
  - Else, sample the next arm according to $\lambda_{\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n), \delta}(\cdot \,|\, \underline{d}(n), \underline{i}(n))$
  - Update $n \leftarrow n+1$

Principle of certainty equivalence

Performance of $\pi^\star(L, \delta)$

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

There exists $\bar{\varepsilon}^* \in (0,1)$ such that for any $C = (h, P, Q)$, the TPMs $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$

Performance of $\pi^{\star}(L, \delta)$

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

There exists $\bar{\varepsilon}^* \in (0,1)$ such that for any $C = (h, P, Q)$, the TPMs $P, Q \in \mathcal{P}(\bar{\varepsilon}^*)$

- $(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n)) \to (h, P_1, P_2)$ (identification/identifiability)

Performance of $\pi^\star(L, \delta)$

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

There exists $\bar{\varepsilon}^* \in (0,1)$ such that for any $C = (h, P, Q)$, the TPMs $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$

- $(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n)) \to (h, P_1, P_2)$ (identification/identifiability)
- If $L = 1/\epsilon$, then $\pi^\star(L, \delta) \in \Pi(\epsilon)$ for all $\delta > 0$

Performance of $\pi^\star(L, \delta)$

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

There exists $\bar{\varepsilon}^* \in (0,1)$ such that for any $C = (h, P, Q)$, the TPMs $P, Q \in \mathscr{P}(\bar{\varepsilon}^*)$

- $(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n)) \to (h, P_1, P_2)$ (identification/identifiability)
- If $L = 1/\epsilon$, then $\pi^\star(L, \delta) \in \Pi(\epsilon)$ for all $\delta > 0$
- Under $C = (h, P_1, P_2)$, we have

Performance of $\pi^\star(L, \delta)$

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

There exists $\bar{\varepsilon}^* \in (0,1)$ such that for any $C = (h, P, Q)$, the TPMs $P, Q \in \mathcal{P}(\bar{\varepsilon}^*)$

- $(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n)) \to (h, P_1, P_2)$ (identification/identifiability)
- If $L = 1/\epsilon$, then $\pi^\star(L, \delta) \in \Pi(\epsilon)$ for all $\delta > 0$
- Under $C = (h, P_1, P_2)$, we have

$$\limsup_{L\to\infty} \frac{E[\tau(\pi^\star(L,\delta)) \,|\, C]}{\log L} \leq \frac{(1+\delta)^2}{R^*(h, P_1, P_2)}$$

Performance of $\pi^\star(L, \delta)$

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

There exists $\bar{\varepsilon}^* \in (0,1)$ such that for any $C = (h, P, Q)$, the TPMs $P, Q \in \mathcal{P}(\bar{\varepsilon}^*)$

- $(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n)) \to (h, P_1, P_2)$ (identification/identifiability)
- If $L = 1/\epsilon$, then $\pi^\star(L, \delta) \in \Pi(\epsilon)$ for all $\delta > 0$
- Under $C = (h, P_1, P_2)$, we have

$$\limsup_{L\to\infty} \frac{E[\tau(\pi^\star(L,\delta)) \,|\, C]}{\log L} \le \frac{(1+\delta)^2}{R^*(h, P_1, P_2)}$$

$$\lim_{\delta\downarrow 0} \limsup_{L\to\infty} \frac{E[\tau(\pi^\star(L,\delta)) \,|\, C]}{\log L} \le \frac{1}{R^*(h, P_1, P_2)}$$

Performance of $\pi^\star(L, \delta)$

For each $\delta > 0$, there exists a selection $\{\lambda_{h,P,Q,\delta}\}_{h,P,Q}$ such that $(h, P, Q) \mapsto \lambda_{h,P,Q,\delta}$ is continuous

There exists $\bar{\varepsilon}^* \in (0,1)$ such that for any $C = (h, P, Q)$, the TPMs $P, Q \in \mathcal{P}(\bar{\varepsilon}^*)$

- $(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n)) \to (h, P_1, P_2)$ (identification/identifiability)
- If $L = 1/\epsilon$, then $\pi^\star(L, \delta) \in \Pi(\epsilon)$ for all $\delta > 0$
- Under $C = (h, P_1, P_2)$, we have

Lower Bound : $$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \,|\, C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$$

$$\limsup_{L \to \infty} \frac{E[\tau(\pi^\star(L, \delta)) \,|\, C]}{\log L} \leq \frac{(1+\delta)^2}{R^*(h, P_1, P_2)}$$

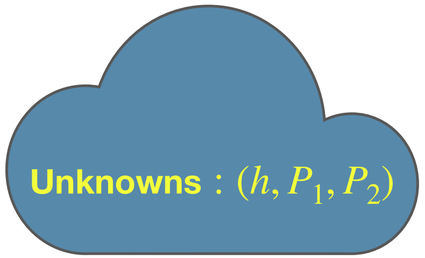
$$\lim_{\delta \downarrow 0} \limsup_{L \to \infty} \frac{E[\tau(\pi^\star(L, \delta)) \,|\, C]}{\log L} \leq \frac{1}{R^*(h, P_1, P_2)}$$

Achievability: A Summary of the Key Ideas

Lower Bound : $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$

Goal: find the odd restless Markov arm as quickly and accurately as possible without the knowledge of the arm TPMs

learning



Uncountably many possibilities

Achievability: A Summary of the Key Ideas

Lower Bound : $$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$$

Goal: find the odd restless Markov arm as quickly and accurately as possible without the knowledge of the arm TPMs

learning

Unknowns : (h, P_1, P_2)

Uncountably many possibilities

Estimate the unknowns

At time n :

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

Achievability: A Summary of the Key Ideas

Lower Bound : $$\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$$

Goal: find the odd restless Markov arm as quickly and accurately as possible without the knowledge of the arm TPMs

learning

Unknowns : (h, P_1, P_2)

Uncountably many possibilities

Estimate the unknowns

At time n :

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

Next arm is selected assuming

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

is the true arms configuration

Achievability: A Summary of the Key Ideas

Lower Bound : $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$

Goal: find the odd restless Markov arm as quickly and accurately as possible without the knowledge of the arm TPMs

learning

Unknowns : (h, P_1, P_2)

Uncountably many possibilities

Estimate the unknowns

At time n :

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

Certainty Equivalence

Next arm is selected assuming

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

is the true arms configuration

Achievability: A Summary of the Key Ideas

Lower Bound : $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n)) \longrightarrow (h, P_1, P_2)$

Identification / Identifiability

Goal: find the odd restless Markov arm as quickly and accurately as possible without the knowledge of the arm TPMs

learning

Unknowns : (h, P_1, P_2)

Uncountably many possibilities

Estimate the unknowns

At time n :

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

Certainty Equivalence

Next arm is selected assuming

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

is the true arms configuration

Achievability: A Summary of the Key Ideas

Maximum likelihood estimation

Continuous selection assumption

Regularity assumption on P_1 and P_2

Lower Bound : $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n)) \longrightarrow (h, P_1, P_2)$

Identification / Identifiability

Goal: find the odd restless Markov arm as quickly and accurately as possible without the knowledge of the arm TPMs

learning

Unknowns : (h, P_1, P_2)

Uncountably many possibilities

Estimate the unknowns

At time n :

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

Certainty Equivalence

Next arm is selected assuming

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

is the true arms configuration

Achievability: A Summary of the Key Ideas

$\pi^{\star}(L, \delta)$

Maximum likelihood estimation

Continuous selection assumption

Regularity assumption on P_1 and P_2

Lower Bound : $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h, P_1, P_2)}$

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n)) \longrightarrow (h, P_1, P_2)$

Identification / Identifiability

Goal: find the odd restless Markov arm as quickly and accurately as possible without the knowledge of the arm TPMs

learning

Unknowns : (h, P_1, P_2)

Uncountably many possibilities

Estimate the unknowns

At time n :

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

Certainty Equivalence

Next arm is selected assuming

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

is the true arms configuration

Achievability: A Summary of the Key Ideas

$L = \frac{1}{\epsilon} \implies \pi^\star(L,\delta) \in \Pi(\epsilon)$

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

Maximum likelihood estimation

Continuous selection assumption

Regularity assumption on P_1 and P_2

Lower Bound : $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi)\,|\,C]}{\log(1/\epsilon)} \geq \frac{1}{R^*(h,P_1,P_2)}$

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n)) \longrightarrow (h, P_1, P_2)$

Identification / Identifiability

Goal: find the odd restless Markov arm as quickly and accurately as possible without the knowledge of the arm TPMs

learning

Unknowns : (h, P_1, P_2)

Uncountably many possibilities

Estimate the unknowns

At time n :

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

Certainty Equivalence

Next arm is selected assuming

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

is the true arms configuration

Achievability: A Summary of the Key Ideas

$L = \frac{1}{\epsilon} \implies \pi^\star(L,\delta) \in \Pi(\epsilon)$

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

$$\limsup_{L\to\infty} \frac{E[\tau(\pi^\star(L,\delta))\,|\,C]}{\log L} \le \frac{(1+\delta)^2}{R^*(h,P_1,P_2)}$$

Upper Bound : $\lim_{\delta\downarrow 0} \limsup_{L\to\infty} \frac{E[\tau(\pi^\star(L,\delta))\,|\,C]}{\log L} \le \frac{1}{R^*(h,P_1,P_2)}$

Maximum likelihood estimation

Continuous selection assumption

Regularity assumption on P_1 and P_2

Lower Bound : $\lim_{\epsilon\downarrow 0} \inf_{\pi\in\Pi(\epsilon)} \frac{E[\tau(\pi)\,|\,C]}{\log(1/\epsilon)} \ge \frac{1}{R^*(h,P_1,P_2)}$

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n)) \longrightarrow (h, P_1, P_2)$

Identification / Identifiability

Goal: find the odd restless Markov arm as quickly and accurately as possible without the knowledge of the arm TPMs

learning

Unknowns : (h, P_1, P_2)

Uncountably many possibilities

Estimate the unknowns

At time n :

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

Certainty Equivalence

Next arm is selected assuming

$(\hat{h}(n), \hat{P}_1(n), \hat{P}_2(n))$

is the true arms configuration

Summary

Summary

- We derived a lower bound on the expected time to find the odd arm

Summary

- We derived a lower bound on the expected time to find the odd arm
- We devised a policy based on the principle of certainty equivalence with ML estimation and showed that it achieves the lower bound

Summary

Main Result： Under $C = (h, P_1, P_2)$, $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} = \frac{1}{R^*(h, P_1, P_2)}$

- We derived a lower bound on the expected time to find the odd arm
- We devised a policy based on the principle of certainty equivalence with ML estimation and showed that it achieves the lower bound

Summary

Main Result： Under $C = (h, P_1, P_2)$, $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} = \frac{1}{R^*(h, P_1, P_2)}$

- We derived a lower bound on the expected time to find the odd arm
- We devised a policy based on the principle of certainty equivalence with ML estimation and showed that it achieves the lower bound
- Key challenge: resolving the parameter identification problem in the context of a countable-state MDP
  - Continuous selection assumption
  - Regularity assumption on the unknown TPMs

Summary

Main Result : Under $C = (h, P_1, P_2)$, $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} = \frac{1}{R^*(h, P_1, P_2)}$

- We derived a lower bound on the expected time to find the odd arm
- We devised a policy based on the principle of certainty equivalence with ML estimation and showed that it achieves the lower bound
- Key challenge: resolving the parameter identification problem in the context of a countable-state MDP
  - Continuous selection assumption
  - Regularity assumption on the unknown TPMs
- We assume that the trembling hand parameter $\eta > 0$
  It would be interesting to study the case $\eta = 0$

Summary

Main Result : Under $C = (h, P_1, P_2)$, $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} = \frac{1}{R^*(h, P_1, P_2)}$

- We derived a lower bound on the expected time to find the odd arm
- We devised a policy based on the principle of certainty equivalence with ML estimation and showed that it achieves the lower bound
- Key challenge: resolving the parameter identification problem in the context of a countable-state MDP
  - Continuous selection assumption
  - Regularity assumption on the unknown TPMs
- We assume that the trembling hand parameter $\eta > 0$
  It would be interesting to study the case $\eta = 0$
- The ML estimates cannot be computed in closed form

Summary

Main Result : Under $C = (h, P_1, P_2)$, $\lim_{\epsilon \downarrow 0} \inf_{\pi \in \Pi(\epsilon)} \frac{E[\tau(\pi) \mid C]}{\log(1/\epsilon)} = \frac{1}{R^*(h, P_1, P_2)}$

- We derived a lower bound on the expected time to find the odd arm
- We devised a policy based on the principle of certainty equivalence with ML estimation and showed that it achieves the lower bound
- Key challenge: resolving the parameter identification problem in the context of a countable-state MDP
  - Continuous selection assumption
  - Regularity assumption on the unknown TPMs
- We assume that the trembling hand parameter $\eta > 0$
  It would be interesting to study the case $\eta = 0$
- The ML estimates cannot be computed in closed form
- Why not a simpler policy: sample the arms repeatedly to estimate the TPMs?

Learning to Detect an Odd Restless Markov Arm with a Trembling Hand

P. N. Karthik and Rajesh Sundaresan

Abstract

This paper studies the problem of finding an anomalous arm in a multi-armed bandit when (a) each arm is a finite-state Markov process, and (b) the arms are restless. Here, anomaly means that the transition probability matrix (TPM) of one of the arms (the odd arm) is different from the common TPM of each of the non-odd arms. The TPMs are unknown to a decision entity that wishes to find the index of the odd arm as quickly as possible, subject to an upper bound on the error probability. We derive a problem instance specific asymptotic lower bound on the expected time required to find the odd arm index, where the asymptotics is as the error probability vanishes. Further, we devise a policy based on the principle of certainty equivalence, and demonstrate that under a continuous selection assumption and a certain regularity assumption on the TPMs, the policy achieves the lower bound arbitrarily closely. Thus, while the lower bound is shown for all problem instances, the upper bound is shown only for those problem instances satisfying the regularity assumption. Our achievability analysis is based on resolving the identifiability problem in the context of a certain countable-state controlled Markov process.

Index Terms

Odd arm identification, restless multi-armed bandits, controlled Markov process, certainty equivalence, identifiability.

Answers to these questions in the full version

https://arxiv.org/pdf/2105.03603.pdf

Acknowledgements

- Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India. (Grant number EMR/2016/002503)

- Centre for Networked Intelligence, Indian Institute of Science

- Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science

Thank You!