2019 IEEE International Symposium on Information Theory

Learning to Detect an Odd Markov Arm

P. N. Karthik and Rajesh Sundaresan

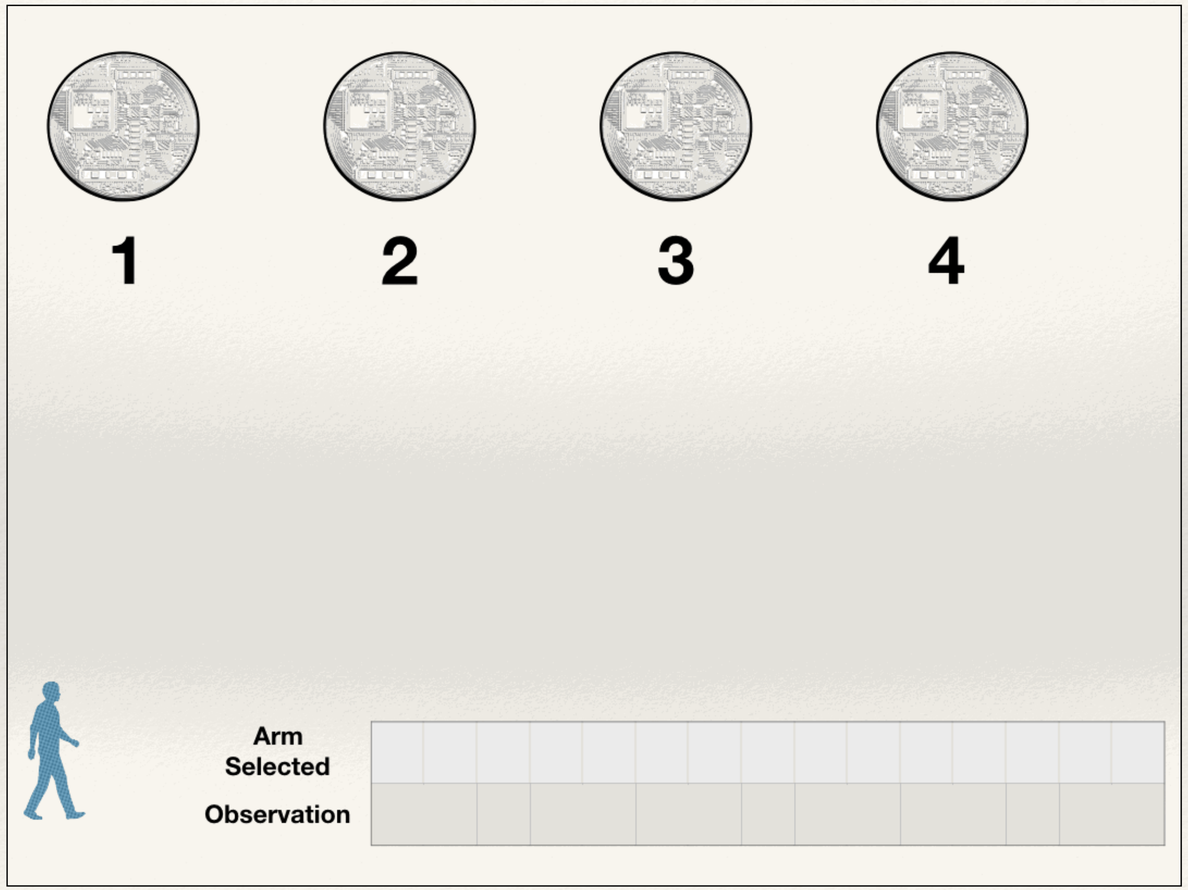
Department of Electrical Communication Engineering

Indian Institute of Science, Bangalore

Problem Setup

- A multi-armed bandit with K independent arms
- Arm: an ergodic DTMC on a finite state space
- Common state space for all the arms
- One of the arms (**odd arm**) is governed by tpm P_1
- Rest of the arms governed by tpm P_2 ($P_2 \neq P_1$)
- Only one arm may be observed at any time
- The observed arm undergoes transition. The unobserved arms remain frozen (rested)

GOAL:
Given an error tolerance level $\epsilon > 0$, identify the odd arm as quickly as possible without the knowledge of P_1 or P_2, subject to $P_e \leq \epsilon$



Comparison With Existing Works

		Problem Setting		Nature of observations		Arm distributions	
		Regret Minimisation	Optimal Stopping	IID	Markov (rested)	Known	Unknown
Odd Arm Identification	Gittins[1]	✓	✗	✗	✓	✓	✗
	Agarwal et al.[2] Anantharam et al.[3]	✓	✗	✗	✓	✗	✓
	Vaidhiyan et al.[4] Prabhu et al.[5]	✗	✓	✓	✗	✗	✓
	Current work	✗	✓	✗	✓	✗	✓

Finite parameter space

Uncountably Infinite parameter space

[1] J. C. Gittins, "Bandit processes and dynamic allocation indices," Journal of the Royal Statistical Society. Series B (Methodological), pp. 148–177, 1979.

[2] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space," IEEE Trans. on Automatic Control, vol. 34, no. 12, pp. 1249–1259, 1989.

[3] Anantharam V, Varaiya P, Walrand J. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part II: Markovian rewards. IEEE Trans. on Automatic Control. 1987 Nov, 32(11), pp. 977-982.

[4] N. K. Vaidhiyan and R. Sundaresan, "Learning to detect an oddball target," IEEE Trans. on Information Theory, vol. 64, no. 2, pp. 831–852, 2018.

[5] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Learning to detect an oddball target with observations from an exponential family," 2017.

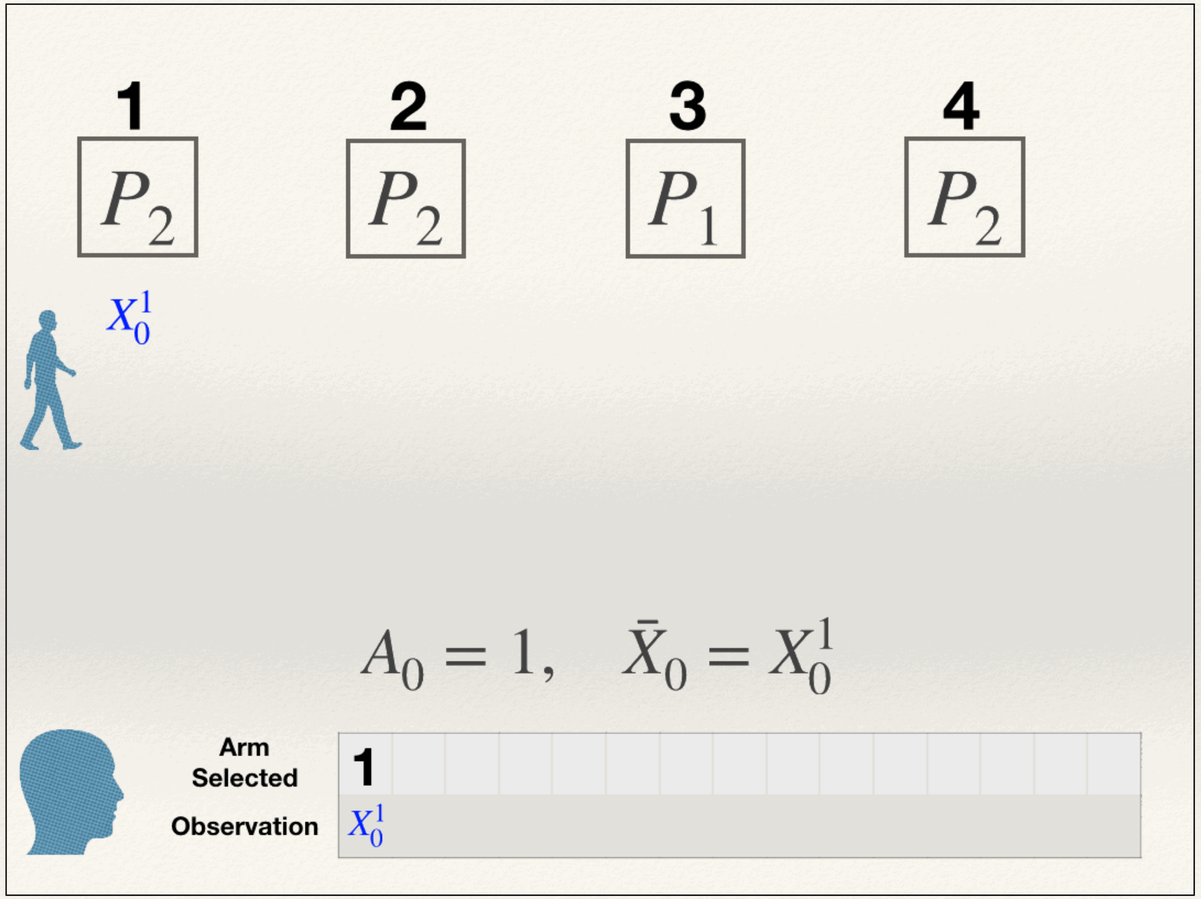
Main Contributions

- An asymptotic[1] lower bound on the expected number of samples required to identify the odd arm as a function of error tolerance
- An asymptotically optimal scheme which respects the desired error tolerance
- The scheme is a modification of the classical GLRT with forced exploration
- Key challenges in the Markov setting identified. A first step towards solving the more difficult case of “restless” Markov arms

[1]The asymptotics is as the error tolerance vanishes.

Notations

- $(X_k^a)_{k\geq 0}$: Markov process of arm a
- TPM of odd arm: P_1; TPM of other arms: P_2
- $(A_k)_{k\geq 0}$: arm selection process
- $(\bar{X}_k)_{k\geq 0}$: observation process
- $A_k \in \sigma(\bar{X}_0^{k-1}, A_0^{k-1})$ for all k



The Lower Bound

Suppose $C = (h, P_1, P_2)$ denotes the underlying configuration of the arms. Let μ_i denote the stationary distribution of P_i for $i = 1,2$.

For any $\epsilon > 0$, let $\Pi(\epsilon)$ denote the set of policies

$$\Pi(\epsilon) = \left\{ \pi : P^{\pi}(\text{error} \,|\, C) \leq \epsilon \;\forall\; C = (h, P_1, P_2) \right\}$$

Lower bound:

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

$$D^*(h, P_1, P_2) = \max_{0 \leq \lambda \leq 1} \left\{ \lambda \, D(P_1 \,||\, P_\lambda \,|\, \mu_1) + (1 - \lambda) \frac{(K-2)}{(K-1)} D(P_2 \,||\, P_\lambda \,|\, \mu_2) \right\}$$

$$P_\lambda(j \,|\, i) = \frac{\lambda_1 \mu_1(i) P_1(j \,|\, i) + (1 - \lambda_1) \frac{(K-2)}{(K-1)} \mu_2(i) P_2(j \,|\, i)}{\lambda_1 \mu_1(i) + (1 - \lambda_1) \frac{(K-2)}{(K-1)} \mu_2(i)}$$

For any two transition matrices P and Q, and a probability distribution μ on the states, define the quantity $D(P \,||\, Q \,|\, \mu)$ as

$$D(P \,||\, Q \,|\, \mu) = \sum_i \mu(i) \sum_j P(j \,|\, i) \log \frac{P(j \,|\, i)}{Q(j \,|\, i)}$$

Remarks

- $D^*(h, P_1, P_2)$ **is a measure of effort required to guard against the nearest alternative**
- **For small values of error tolerance** ϵ**,**

$$E^{\pi}[\tau(\pi) \,|\, C] \gtrapprox \frac{\log(1/\epsilon)}{D^*(h, P_1, P_2)}, \quad \pi \in \Pi(\epsilon)$$

Lower Bound: Key Steps in the Proof

- If $C = (h, P_1, P_2)$ and $C' = (h', P_1', P_2')$, where $h' \neq h$, then using a result of Kaufmann et al.[1], we have

$$\min_{C'=(h',P_1',P_2')} E^{\pi}\left[\sum_{t=0}^{\tau(\pi)} \log \frac{f(\bar{X}_t \,|\, \bar{X}_0^{t-1}, A_0^{t-1}, C)}{f(\bar{X}_t \,|\, \bar{X}_0^{t-1}, A_0^{t-1}, C')}\right] \quad \geq \quad \underbrace{\epsilon \log \frac{\epsilon}{1-\epsilon} + (1-\epsilon)\log \frac{1-\epsilon}{\epsilon}}_{d(\epsilon, 1-\epsilon)}$$

- For any given arm, the long-term fraction of exits of a state i is equal to the long-term fraction of entries into i. This common fraction is the stationary probability of observing i. **This is a consequence of the rested nature of the arms**

- Wald's identity not applicable. A generalisation of a change of measure argument in Kaufmann et al.[1] to Markov processes used

- $\dfrac{d(\epsilon, 1-\epsilon)}{\log(1/\epsilon)} \to 1 \text{ as } \epsilon \downarrow 0$

Lower bound:

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

$$D^*(h, P_1, P_2) = \max_{0 \leq \lambda \leq 1} \left\{ \lambda \, D(P_1 \,||\, P_\lambda \,|\, \mu_1) + (1-\lambda)\frac{(K-2)}{(K-1)} D(P_2 \,||\, P_\lambda \,|\, \mu_2) \right\}$$

$$P_\lambda(j \,|\, i) = \frac{\lambda_1 \mu_1(i) P_1(j \,|\, i) + (1-\lambda_1)\frac{(K-2)}{(K-1)}\mu_2(i) P_2(j \,|\, i)}{\lambda_1 \mu_1(i) + (1-\lambda_1)\frac{(K-2)}{(K-1)}\mu_2(i)}$$

$$E^{\pi}\left[\sum_{t=0}^{\tau(\pi)} \log \frac{f(\bar{X}_t \,|\, \bar{X}_0^{t-1}, A_0^{t-1}, C)}{f(\bar{X}_t \,|\, \bar{X}_0^{t-1}, A_0^{t-1}, C')}\right] = E^{\pi}[\tau(\pi)] \cdot E^{\pi}\left[\log \frac{f(\bar{X}_t \,|\, A_t, C)}{f(\bar{X}_t \,|\, A_t, C')}\right]$$

Wald's identity for IID processes

[1] E. Kaufmann, O. Cappe, and A. Garivier, "On the complexity of best-arm identification in multi-armed bandit models," The Journal of Machine Learning Research, vol. 17, no. 1, pp. 1–42, 2016.

Upper Bound: Preliminaries

- The learner does not know the underlying configuration of the arms
- Suppose the configuration is $C = (h, P_1, P_2)$
- Suppose observations $\bar{X}_0, \ldots, \bar{X}_n$ have been obtained under arm selections $A_0, \ldots, A_n$
- The joint likelihood of all observations, given arm selections, under config. C is

$$f(\bar{X}_0^n \,|\, A_0^n, C) = \prod_{a=1}^{K} \nu(X_0^a) \prod_{i,j} (P_1(j\,|\,i))^{N_h(n,i,j)} \; (P_2(j\,|\,i))^{\sum_{a \neq h} N_a(n,i,j)}$$

- Introduce a joint prior on (P_1, P_2) as $\mathscr{D}(P_1, P_2) = \Gamma(P_1) \cdot \Gamma(P_2)$
- **Average likelihood** of all observations, given actions, when h is the odd arm is

$$f_h(\bar{X}_0^n \,|\, A_0^n) = \prod_{a=1}^{K} \nu(X_0^a) \int_{(P_1,P_2)} \prod_{i,j} (P_1(j\,|\,i))^{N_h(n,i,j)} \; (P_2(j\,|\,i))^{\sum_{a \neq h} N_a(n,i,j)} \, \mathscr{D}(P_1, P_2) \, dP_1(j\,|\,i) \, dP_2(j\,|\,i)$$

$N_h(n, i, j)$: no. of $i \to j$ transitions observed on arm h

Prior $\Gamma(\,\cdot\,)$ on the space of all TPMs:

Pick each row of a TPM independently according to uniform distribution on the probability simplex.

Cannot be computed since the actual configuration is not known

Can be computed for each odd arm index h

Upper Bound: Preliminaries

- **Average likelihood** of all observations, given actions, when h is the odd arm is

$$f_h(\bar{X}_0^n \mid A_0^n) = \prod_{a=1}^{K} \nu(X_0^a) \int_{(P_1,P_2)} \prod_{i,j} (P_1(j \mid i))^{N_h(n,i,j)} \, (P_2(j \mid i))^{\sum_{a \neq h} N_a(n,i,j)} \, \mathscr{D}(P_1, P_2) \, dP_1(j \mid i) \, dP_2(j \mid i)$$

- When h is the odd arm, the **maximum likelihood** of all observations, given actions, is

$$\hat{f}_h(\bar{X}_0^n \mid A_0^n) = \prod_{a=1}^{K} \nu(X_0^a) \prod_{i,j} \left(\frac{N_h(n,i,j)}{N_h(n,i)} \right)^{N_h(n,i,j)} \left(\frac{\sum_{a \neq h} N_a(n,i,j)}{\sum_{a \neq h} N_a(n,i)} \right)^{\sum_{a \neq h} N_a(n,i,j)}$$

When h is the odd arm:

$\frac{N_h(n,i,j)}{\sum_j N_h(n,i,j)}$ = ML estimate of (i,j) entry of the TPM of h

$\frac{\sum_{a \neq h} N_a(n,i,j)}{\sum_{a \neq h} \sum_j N_h(n,i,j)}$ = ML estimate of (i,j) entry of the TPM of $a \neq h$

Modified GLR Statistic

$$M_{ha}(n) = \log \frac{f_h(\bar{X}_0^n \mid A_0^n)}{\hat{f}_a(\bar{X}_0^n \mid A_0^n)}, \quad a \neq h$$

Usual GLR Statistic

$$M_{ha}(n) = \log \frac{\hat{f}_h(\bar{X}_0^n \mid A_0^n)}{\hat{f}_a(\bar{X}_0^n \mid A_0^n)}, \quad a \neq h$$

Policy: Modified GLRT + Forced Exploration

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

- Fix parameters $L \geq 1$ and $\delta \in (0,1)$
- Select each arm once in the first K time instants
- Repeat the following steps until stoppage for all $n \geq K-1$:
  - Compute

$$\hat{h}(n) = \arg\max_{h} \min_{a \neq h} M_{ha}(n)$$

  - if $\min_{a \neq \hat{h}(n)} M_{\hat{h}(n)a}(n) \geq \log((K-1)L)$, stop and declare $\hat{h}(n)$ as the odd arm
  - if $\min_{a \neq \hat{h}(n)} M_{\hat{h}(n)a}(n) < \log((K-1)L)$, select next arm as follows:
    - Toss a coin with bias δ
    - If coin falls heads, select the next arm uniformly at random
    - If coin falls tails, select the next arm according to distribution $\lambda_{opt}(\hat{h}(n), \hat{P}_1^n, \hat{P}_2^n)$

L : error tolerance control parameter

δ : forced exploration parameter

Modified GLR Statistic

$$M_{ha}(n) = \log \frac{f_h(\bar{X}_0^n \,|\, A_0^n)}{\hat{f}_a(\bar{X}_0^n \,|\, A_0^n)}, \quad a \neq h$$

$$D^*(h, P_1, P_2) = \max_{0 \leq \lambda \leq 1} \left\{ \lambda\, D(P_1 \,||\, P_\lambda \,|\, \mu_1) + (1-\lambda) \frac{(K-2)}{(K-1)} D(P_2 \,||\, P_\lambda \,|\, \mu_2) \right\}$$

$$P_\lambda(j \,|\, i) = \frac{\lambda_1 \mu_1(i) P_1(j \,|\, i) + (1-\lambda_1)\frac{(K-2)}{(K-1)} \mu_2(i) P_2(j \,|\, i)}{\lambda_1 \mu_1(i) + (1-\lambda_1)\frac{(K-2)}{(K-1)} \mu_2(i)}$$

Suppose $\hat{\lambda} = \arg\max \lambda$ in the expression for $D^*(\hat{h}(n), \hat{P}_1^n, \hat{P}_2^n)$

Then, let

$$\lambda_{opt}(\hat{h}(n), \hat{P}_1^n, \hat{P}_2^n)(a) = \begin{cases} \hat{\lambda}, & a = \hat{h}(n), \\ \frac{1-\hat{\lambda}}{K-1}, & a \neq \hat{h}(n)\,. \end{cases}$$

Salient Features of Policy $\pi^\star(L, \delta)$

- The policy stops in finite time a.s.
- If $C = (h, P_1, P_2)$ is the actual configuration, then a.s.:

$$\hat{h}(n) \to h$$
$$\hat{P}_1^n(j \mid i) \to P_1(j \mid i) \quad \text{for all } i, j$$
$$\hat{P}_2^n(j \mid i) \to P_2(j \mid i) \quad \text{for all } i, j$$

- Given error tolerance $\epsilon > 0$, for any $\delta \in (0,1)$,

$$L = (1/\epsilon) \implies P_e(\pi^\star(L, \delta)) \leq \epsilon$$

- **Asymptotic optimality**:

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

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

- Fix parameters $L \geq 1$ and $\delta \in (0,1)$
- Select each arm once in the first K time instants
- Repeat the following until stop for all $n \geq K - 1$:
  - Compute

$$\hat{h}(n) = \arg\max_h \min_{a \neq h} M_{ha}(n)$$

  - if $M_{\hat{h}(n)}(n) \geq \log((K-1)L)$, stop and declare $\hat{h}(n)$ as the odd arm
  - if $M_{\hat{h}(n)}(n) < \log((K-1)L)$, select next arm as follows:
    - Toss a coin with bias δ
    - If coin falls heads, select the next arm uniformly at random
    - If coin falls tails, select the next arm according to distribution $\lambda_{opt}(\hat{h}(n), \hat{P}_1^n, \hat{P}_2^n)$

Conclusions and Future Work

- Analysed the problem of odd arm identification for rested Markov arms
- Provided an asymptotic lower bound on the expected number of arm selections needed to identify the odd arm
- Proposed a scheme based on a modification of GLRT with forced exploration, and proved that it is asymptotically optimal
- Future work: to analyse the more difficult problem of odd arm identification in the setting of "restless" Markov arms. The analysis in the current paper serves as a key first step towards this

Acknowledgements

- Science and Engineering Research Board (SERB), Department of Science and Technology (grant no. EMR/2016/002503), Govt. of India

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

Thank You