Learning to Detect an Odd Markov Arm ICTS Program on Applied Probability

PN Karthik and Rajesh Sundaresan

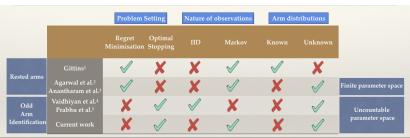
Indian Institute of Science, Bangalore

August 17, 2019

Problem Setting

- A multi-armed bandit with K independent arms
- Each arm is a time homogeneous and ergodic DTMC on a finite state space
- The state space is common to all arms
- One of the arms is "odd": governed by different TPM than the rest
- The transition matrices of the odd arm and the non-odd arms are not known
- Use sequential tests to identify the odd arm as quickly as possible
- Only the arm selected undergoes state evolution. States of the remaining arms "rested" or frozen at their last observed values
- PAC framework

Review of Known Results



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, 1989.

³ Anantharam V, <u>Varaiya</u> P, <u>Walrand</u> J. Asymptotically efficient allocation rules for the <u>multiarmed</u> bandit problem with multiple plays-Part II: Markovian rewards. IEEE Trans. on Automatic Control. 1987.

⁴ N. K. <u>Vaidhiyan</u> 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.

Contributions

- Asymptotic lower bound on the expected number of samples required to identify the odd arm as a function of error tolerance, where the asymptotics is in the regime of vanishing error tolerance
- Asymptotically optimal scheme: modified GLRT + forced exploration
- Key challenges in the rested Markov setting identified
- Our analysis serves as a first step towards analysing the more difficult problem of odd arm identification in restless Markov bandits

Notations

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

$$D(P||Q|\mu) \coloneqq \sum_{i \in \mathcal{S}} \mu(i) \sum_{j \in \mathcal{S}} P(j|i) \log \frac{P(j|i)}{Q(j|i)}$$

- A triplet $C = (h, P_1, P_2)$ denotes a configuration of the arms in which the odd arm index is h, the TPM of the odd arm Markov process is P_1 and the TPMs of each of the non-odd arms is P_2 ; here, $P_2 \neq P_1$
- Given $\epsilon > 0$, define the set of policies

$$\Pi(\epsilon) = \left\{ \pi : P^{\pi}(\mathsf{error}|C) \leq \epsilon \ \forall \ C = (h, P_1, P_2), \ \mathsf{where} \ h \in \mathcal{A} \ \mathsf{and} \ P_2 \neq P_1 \right\}$$

Lower Bound

Proposition

Let $C = (h, P_1, P_2)$ denote the underlying configuration of the arms. Then,

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

where $D^*(h, P_1, P_2)$ is a configuration-dependent constant that is a function only of P_1 and P_2 , and is given by

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

In (1), P_{λ} is a transition probability matrix whose entry in the ith row and jth column is given by

$$P_{\lambda}(j|i) = \frac{\lambda \mu_{1}(i) P_{1}(j|i) + (1 - \lambda) \frac{(K-2)}{(K-1)} \mu_{2}(i) P_{2}(j|i)}{\lambda \mu_{1}(i) + (1 - \lambda) \frac{(K-2)}{(K-1)} \mu_{2}(i)}$$

Achievability - 1

- We devise a sequential test that is a modification of the classical GLRT with forced exploration
- The modification is obtained by replacing the max in the numerator of the classical GLR statistic by an average computed with respect to an artificial prior
- lacksquare Our scheme $\pi^{\star}(L,\delta)$ takes as input two parameters $L\geq 1$ and $\delta\in(0,1)$
- lacksquare Parameter L controls error probability, while δ controls the amount of forced exploration
- Setting $L = 1/\epsilon$ ensures $\pi^*(L, \delta) \in \Pi(\epsilon)$ for all $\delta \in (0, 1)$

Achievability - 2

Proposition

For the policy
$$\pi = \pi^*(L, \delta)$$
,

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