Learning to Detect an Odd Markov Arm

P. N. Karthik and Rajesh Sundaresan
Department of Electrical Communication Engineering, Indian Institute of Science

Problem Setup

We consider a multi-armed bandit with K arms in which each arm is viewed as an irreducible, aperiodic and time homogeneous discrete time Markov process evolving on a finite state space $\mathcal{S}$. The state space $\mathcal{S}$ is common to all the arms, and the arms are independent of each other. The transition probability matrix of one of the arms (which we refer to in the sequel as the odd arm) is P_1, whereas that of all the remaining arms is P_2, where $P_2 \neq P_1$. A learner who has no knowledge of P_1 or P_2, but possesses the knowledge that one of the arms is odd, wishes to identify the index of the odd arm as quickly as possible. In order to do so, he devises sequential tests that involve selecting one of the K arms at each time and observing the state of the selected arm. At any given time, the unobserved arms remain frozen (or "rested") at their last observed state values. The goal of the learner is to identify the index of the odd arm with as few number of arm selections as possible, while ensuring that probability of making error is below a certain specified tolerance.

Illustration

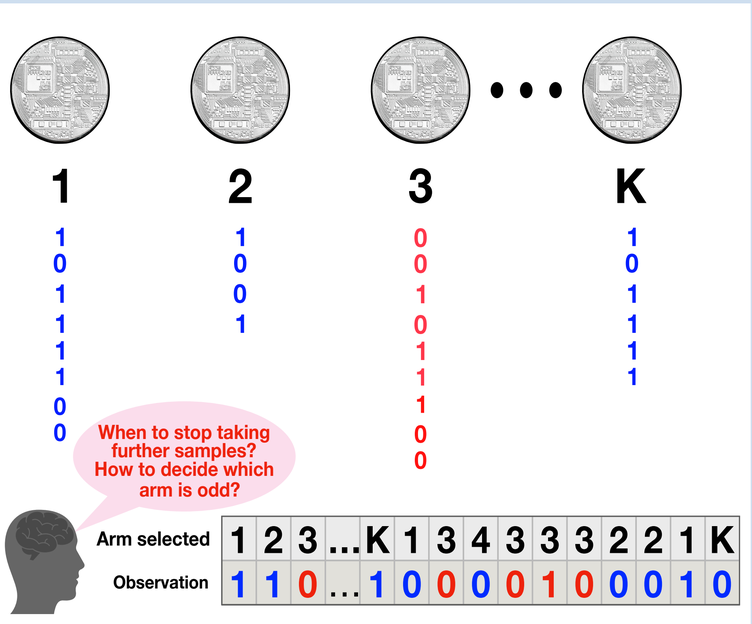

Figure 1: Illustration depicting a multi-armed bandit with K arms in which each arm is a coin. The coins are visually identical. One of the coins has a bias different from the common bias of the remaining coins. A learner who has no knowledge of the coin biases wishes to identify the "odd coin" by selecting one at each time and observing its outcome. The learner wishes to accomplish his task by taking as few observations as possible.

Main Contributions

- We derive an asymptotic lower bound on the expected number of arm selections required by the learner to identify the index of the odd arm. Here, the asymptotics is in the regime of vanishing error tolerance.
- We propose a scheme to detect the odd Markov arm, and show that its expected number of arm selections comes arbitrarily close to the lower bound as the error tolerance vanishes (asymptotic optimality).

Notations

- $C = (h, P_1, P_2)$: an underlying configuration of the arms. Here, h is the index of the odd arm, P_1 is the transition probability matrix of the odd arm Markov process and $P_2 \neq P_1$ that of the remaining non-odd processes.
- π: A policy or scheme used by the learner to identify the index of the odd arm.
- $\tau(\pi)$: stopping time of policy π.
- $\Pi(\epsilon)$: set of all policies π whose probability of error at stoppage is below an error tolerance ϵ for all arm configurations $C = (h, P_1, P_2)$.

Converse (Lower Bound)

For any two transition probability matrices P and Q of dimension $|\mathcal{S}| \times |\mathcal{S}|$, and a probability distribution μ on $\mathcal{S}$, define $D(P||Q|\mu)$ as the quantity

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

with the convention $0 \log 0 = 0 = 0 \log \frac{0}{0}$.
Suppose $C = (h, P_1, P_2)$ is the underlying configuration of the arms. Further, let μ_i denote the unique stationary distribution of P_i, $i = 1, 2$. Then,

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

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

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

In the above equation, 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: Modified GLRT with Forced Exploration [2]

1

$$\lambda_{opt}(h, P_1, P_2)(a) = \begin{cases} \lambda^*, & a = h, \\ \frac{1-\lambda^*}{K-1}, & a \neq h, \end{cases}$$

where λ^* is the optimal value of λ that maximizes $D^*(h, P_1, P_2)$

2

$(X_k^a)_{k\geq 0}$: **Markov process of arm** a
$(A_k)_{k\geq 0}$: **arm selection process**
$(\bar{X}_k)_{k\geq 0}$: **observation process**
$(A_0^k, \bar{X}_0^k)$: **history up to time** k
$\{A_k = a\} \in \sigma(A_0^{k-1}, \bar{X}_0^{k-1})$ **for all** $k \geq 0$

3

$$N_a(n) = \sum_{k=0}^{n} 1_{\{A_k=a\}}$$

$$N_a(n, i) = \sum_{k=0}^{N_a(n)-2} 1_{\{X_k^a=i\}}, \quad i \in \mathcal{S}$$

$$N_a(n, i, j) = \sum_{k=0}^{N_a(n)-2} 1_{\{X_k^a=i,\, X_{k+1}^a=j\}}, \quad i, j \in \mathcal{S}$$

7

The Modified GLR Statistic

$$M_h(n) = \log \frac{f(\bar{X}_0^n | A_0^n, H_h)}{\max_{a\neq h} f(\bar{X}_0^n | A_0^n, H_a)}$$

8

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

Fix parameters $L \geq 1$ **and** $\delta \in (0,1)$
$A_n = n + 1$ **for** $n = 0,1,2,\ldots, K-1$

For $n \geq K-1$, **repeat until stop** :

$$h^*(n) = \arg \max_{h\in\{1,2,\ldots,K\}} M_h(n)$$

If $M_{h^*(n)}(n) \geq \log((K-1)L)$, **stop** $h^*(n)$ **is the odd arm**

If $M_{h^*(n)}(n) < \log((K-1)L)$
choose $A_{n+1} = a$ **as follows:**
w.p. δ **uniformly**
w.p. $1-\delta$ **according to**
$\lambda_{opt}(h^*(n), \hat{P}^n_{h^*(n),1}, \hat{P}^n_{h^*(n),2})(a)$

6

Fix parameters $\alpha_1, \ldots, \alpha_{|\mathcal{S}|}$.
Average likelihood when the odd arm is h :

$$f(\bar{X}_0^n | A_0^n, H_h) = C \cdot \prod_{a=1}^{K} \nu(X_0^a) \cdot \int_{P,Q} \prod_{i\in\mathcal{S}} \prod_{i\in\mathcal{S}} (P(j|i))^{\alpha_j - 1} (Q(j|i))^{\alpha_j - 1}\, dP(j|i)\, dQ(j|i)$$

Maximum likelihood when the odd arm is $h' \neq h$:

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

5

Suppose P **and** Q **are** $|\mathcal{S}| \times |\mathcal{S}|$ **transition probability matrices. The following is a joint pdf on** (P, Q) **with parameters** $\alpha_1, \ldots, \alpha_{|\mathcal{S}|}$:

$$\mathscr{D}(P, Q) = C \cdot \prod_{i\in\mathcal{S}} \prod_{j\in\mathcal{S}} (P(j|i))^{\alpha_j - 1} (Q(j|i))^{\alpha_j - 1}$$

where C **is the normalisation constant.**

4

Maximum likelihood estimates of transition probability matrices P_1 and P_2 for arms configuration $C = (h, P_1, P_2)$

$$\hat{P}^n_{h,1}(j|i) = \frac{N_h(n,i,j)}{N_h(n,i)}, \quad \hat{P}^n_{h,2}(j|i) = \frac{\sum_{a\neq h} N_a(n,i,j)}{\sum_{a\neq h} N_a(n,i)}, \quad i, j \in \mathcal{S}$$

Lower Bound: Key Ideas

- For each arm, the long-term fraction of exits out of any state is equal to the long-term fraction of entries into the same state, and both these are equal to the stationary probability of observing this state.
- Wald's identity for iid processes not applicable. Generalisation of a result in Kaufmann et al. [1] to Markov processes used.

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

- With probability 1, policy $\pi^\star(L, \delta)$ stops in finite time for any value of $L \geq 1$ and $\delta \in (0, 1)$.
- For any fixed $\delta \in (0, 1)$,

$$L = (1/\epsilon) \implies \pi^\star(L, \delta) \in \Pi(\epsilon).$$

- Given $C = (h, P_1, P_2)$, for any sequence $\{L_n\}_{n\geq 1}$ satisfying $L_n \to \infty$ as $n \to \infty$,

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

Conclusions

- We analysed the problem of odd arm identification for the case of "rested" Markov arms.
- We provided an asymptotic lower bound on the expected number of observations required to identify the odd arm, and an asymptotically optimal policy, where the asymptotics is as error tolerance vanishes.
- In a future work, we hope to extend the above analysis to the case of "restless" Markov arms where the unobserved arms continue to evolve.

Acknowledgements

This work was supported by the Science and Engineering Research Board, Department of Science and Technology (grant no. EMR/2016/002503), and by the Robert Bosch Centre for Cyber Physical Systems at the Indian Institute of Science.

References

1. Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. The Journal of Machine Learning Research, 17(1), 1-42.

2. Albert, A. E. (1961). The sequential design of experiments for infinitely many states of nature. The Annals of Mathematical Statistics, 32(3), 774-799.