A Lower Bound for Identifying the Best Markovian Arm with Fixed Confidence

Vrettos Moulos

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2019-33
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2019/EECS-2019-33.html

May 14, 2019

Acknowledgement

We are grateful to Venkat Anantharam, Jim Pitman and Satish Rao for many helpful lectures and discussion.

A Lower Bound for Identifying the Best Markovian Arm with Fixed Confidence

by Vrettos Moulos

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

_____

Professor Satish Rao
Research Advisor

_____

(Date)

* * * * * * *

_____

Professor Jiantao Jiao
Second Reader

_____

(Date)

A Lower Bound for Identifying the Best Markovian Arm with Fixed Confidence

Vrettos Moulos VRETTOS@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

Abstract

We consider the problem of best Markovian arm identification, where we sequentially collect samples from K Markov chains with our goal being to identify the one with the largest stationary mean with some fixed level of confidence. In Theorem 4 we derive an instance specific non-asymptotic lower bound for the sample complexity, which in the high confidence regime (Corollary 5) generalizes the asymptotic lower bound of Garivier and Kaufmann (2016) which deals with the special case where the K stochastic processes are i.i.d. processes.

Keywords: multi-armed bandits, best Markovian arm identification with fixed confidence, Markov chains

1. Introduction

Consider the following simple setting: we have K stochastic processes/arms which are stationary or converge to stationarity, with stationary means $\mu_1, \ldots, \mu_K$, and at each time-step we select the process/arm from which we want to observe a sample while the others stay still.

In the stochastic multi-armed bandits literature one objective that has been extensively studied is the one of maximizing the expected value of the sum of the observed samples, or minimize the so called regret. The seminal work of Lai and Robbins (1985) popularized this problem, in the case that each of the K process is an i.i.d. process. Anantharam et al. (1987a) first generalized the problem to the scenario when one is allowed to get multiple samples at each time-step, and then in Anantharam et al. (1987b) they dropped the i.i.d. assumption and considered K irreducible and aperiodic finite state Markov chains. We refer the reader to the survey of Bubeck and Cesa-Bianchi (2012) for more details.

An alternative objective, which only recently draw the attention of the research community, is the one of identifying the process with the highest/best mean as fast as and as accurately as possible, notions which are made precise in Section 3. In the i.i.d. setting, Even-Dar et al. (2006) establish an elimination based algorithm in order to find an approximately best arm, and Mannor and Tsitsiklis (2004) provide a matching lower bound. Jamieson et al. (2014) propose an upper confidence strategy, inspired by the law of iterated logarithm, for exact best arm identification given some fixed level of confidence. In the asymptotic high confidence regime, the problem is settled by the work of Garivier and Kaufmann (2016), who provide instance specific matching lower and upper bounds. For their upper bound they propose the Track-and-Stop strategy which is further explored in the work of Kaufmann and Koolen (2018).

In this work we consider K irreducible and positive recurrent Markov chains over a countable state space, with the objective being to identify the best arm with fixed confidence. In Section 4 we establish an instance specific non-asymptotic lower bound, which in the high confidence regime generalizes the lower bound of the i.i.d. setting established by Garivier and Kaufmann (2016). In our subsequent work Moulos (2019) we provide an analysis of the Track-and-Stop strategy in the Markovian setting with asymptotic sample complexity that is at most four times the asymptotic lower bound. We note that the Markov chains that we consider in this work are more general than the Markov chains considered in Moulos (2019), since here we allow the state space to be countable infinite and we don't enforce all K Markov chains to come from an exponential family of transition probability functions.

2. Preliminaries

Let $S \subset \mathbb{R}$ be a countable subset of the real numbers, which will serve as our state space. On this state space we consider a one-parameter family of transition probability functions

$$\mathcal{P} := \{P_\theta : S \times S \to [0,1] : \theta \in \Theta \subseteq \mathbb{R}\}.$$

For the one-parameter family we impose the following assumptions:

1. For all $\theta \in \Theta$, P_θ is the transition probability function of an irreducible and positive recurrent Markov chain. Therefore, there exists a unique stationary distribution π_θ corresponding to P_θ.

2. Assume that for all $\theta \in \Theta$, $\sum_{x \in S} |x| \pi_\theta(x) < \infty$, and denote by μ_θ the stationary mean of the Markov chain with transition probability function P_θ, i.e.

$$\mu_\theta := \sum_{x \in S} x \pi_\theta(x).$$

3. The map $\theta \mapsto \mu_\theta$ is a bijection of Θ to an interval $\mathcal{M}$. Therefore, with some abuse of notation for $\mu \in \mathcal{M}$ we may write P_μ in order to denote P_θ, where θ is the preimage of μ.

4. For all $x, y \in S$ and all $\theta, \theta' \in \Theta$, $P_\theta(x, y) > 0 \;\Rightarrow\; P_{\theta'}(x, y) > 0$.

For two probability measures ν, ξ over the same measure space $(\Omega, \mathcal{F})$, we define the *relative entropy* from ξ to ν as

$$D(\nu \parallel \xi) := \begin{cases} \int \log\left(\frac{d\nu}{d\xi}\right) d\nu, & \text{if } \nu \ll \xi \\ \infty, & \text{otherwise} \end{cases}$$

Let $\mathcal{M}_1(S^k)$ be the set of all probability distributions on S^k. For $\nu, \xi \in \mathcal{M}_1(S^k)$ the relative entropy from ξ to ν simplifies to

$$D(\nu \parallel \xi) = \sum_{x \in S^k} \nu(x) \log \frac{\nu(x)}{\xi(x)},$$

with the standard notational conventions $\log 0 = \infty$, $\log \frac{\alpha}{0} = \infty$ if $\alpha > 0$, $0 \log 0 = 0 \log \frac{0}{0} = 0$.

We denote the binary relative entropy from $\mathrm{Bernoulli}(q)$ to $\mathrm{Bernoulli}(p)$ with

$$D_2(p \parallel q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

We will be mostly interested in bi-variate distributions, which arise from stationary Markov chains. Let $q \in \mathcal{M}_1(S)$, and Q be a transition probability function on S. We define the bi-variate distribution $q \odot Q \in \mathcal{M}_1(S^2)$ as

$$q \odot Q(x, y) := q(x) Q(x, y).$$

So in particular the *stationary relative entropy* between two Markov chains of the family $\mathcal{P}$ is given by

$$D\left(\theta \,\middle\|\, \theta'\right), D(\mu_\theta \parallel \mu_{\theta'}) := D(\pi_\theta \odot P_\theta \parallel \pi_\theta \odot P_{\theta'}) = \sum_{x \in S} \pi_\theta(x) D(P_\theta(x, \cdot) \parallel P_{\theta'}(x, \cdot)).$$

3. Markovian Bandit Model

Our Markovian bandit model consists of K irreducible and positive recurrent Markov chains, each determined by an initial distribution and a transition probability function.

Let $\mathcal{I} := \mathcal{M}_1(S)^K$ be the set of all possible initial distributions of K Markov chains. Let $\mathcal{T} \subset \mathcal{M}^K$ be a set of vectors such that for each $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in \mathcal{T}$ there exists an $a^*(\boldsymbol{\mu}) \in \{1, \ldots, K\}$ such that $\mu_{a^*(\boldsymbol{\mu})} > \mu_a$ for all $a \neq a^*(\boldsymbol{\mu})$. Each $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in \mathcal{S}$ should be though of as a vector $(P_{\mu_1}, \ldots, P_{\mu_K}) \in \mathcal{P}^K$ of K irreducible and positive recurrent Markov chains, with a single of them possessing the highest stationary mean.

A Markovian bandit model is a pair

$$(\boldsymbol{q}, \boldsymbol{\mu}) = ((q_1, \dots, q_K), (\mu_1, \dots, \mu_K)) \in \mathcal{I} \times \mathcal{T}.$$

The evolution of each arm $a = 1, \dots, K$ is completely determined by its initial distribution $q_a \in \mathcal{M}_1(S)$ and its transition probability function $P_{\mu_a} \in \mathcal{P}$. We will denote samples coming from arm a as $X_{a,0}, X_{a,1}, \dots, X_{a,n}, \dots$. In addition after observing t samples over all, let $N_a(t)$ be the number of transitions coming from the Markovian arm a. We will define $\mathcal{F}_t$ to be the observed information up to and including the t-th sample, i.e.

$$\mathcal{F}_t := \sigma\left(X_{1,0}, X_{1,1}, \dots, X_{1,N_1(t)}, \dots, X_{K,0}, X_{K,1}, \dots, X_{K,N_K(t)}\right).$$

Our goal is to identify the best arm $a^*(\boldsymbol{\mu})$ as fast and as accurately as possible, where by accuracy we mean that given a level $\delta \in (0,1)$ we have to find the best arm with probability at least $1-\delta$. To this end for the given level δ we need to come up with a strategy $\mathcal{A}_\delta$ which is a triple $\mathcal{A}_\delta = ((A_t)_{t \in \mathbb{Z}_{>0}}, \tau_\delta, \hat{a}_{\tau_\delta})$ consisting of:

- a *sampling rule* $(A_t)_{t \in \mathbb{Z}_{>0}}$, which based on the past observations $\mathcal{F}_t$, determines which arm A_{t+1} we should sample next, so A_{t+1} is $\mathcal{F}_t$-measurable;
- a *stopping rule* τ_δ, which denotes the end of the data collection phase and is stopping time with respect to $(\mathcal{F}_t)_{t \in \mathbb{Z}_{>0}}$, such that $\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}\, \tau_\delta < \infty$ for all $(\boldsymbol{q}, \boldsymbol{\mu}) \in \mathcal{I} \times \mathcal{T}$;
- a *decision rule* $\hat{a}_{\tau_\delta}$, which is $\mathcal{F}_{\tau_\delta}$-measurable, and determines the arm that we estimate to be the best one.

So if we use strategy $\mathcal{A}_\delta$, after observing t samples the number of transitions coming from arm a is

$$N_a(t) = \sum_{s=1}^{t} 1\{A_s = a\} - 1.$$

Of course our strategies need to perform well across all possible bandit instances, therefore we need to restrict our strategies to a class of 'uniformly accurate' strategies. This motivates the following standard definition.

Definition 1 (δ-PC) *Given a level $\delta \in (0,1)$, a strategy $\mathcal{A}_\delta = ((A_t)_{t \in \mathbb{Z}_{>0}}, \tau_\delta, \hat{a}_{\tau_\delta})$ is called δ-PC (Probably Correct) if,*

$$\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}(\hat{a}_{\tau_\delta} \neq a^*(\boldsymbol{\mu})) \leq \delta, \text{ for all } (\boldsymbol{q}, \boldsymbol{\mu}) \in \mathcal{I} \times \mathcal{T}.$$

Overall given a bandit model $(\boldsymbol{q}, \boldsymbol{\mu}) \in \mathcal{I} \times \mathcal{T}$, and an accuracy level $\delta \in (0,1)$ our goal is to derive an instance specific lower bound for the quantity

$$\inf_{\mathcal{A}_\delta:\ \delta\text{-PC}} \mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[\tau_\delta].$$

4. Lower Bound on the Sample Complexity

Deriving lower bounds in the multi-armed bandits setting is a task performed by change of measure arguments which roughly speaking say that in order to identify the best arm we should at least be able to differentiate between two bandit models that exhibit different best arms but are statistically similar, a technique popularized by Lai and Robbins (1985). For our purposes we use a variant developed by Garivier and Kaufmann (2016) which combines several change of measures at once.

Let $(\boldsymbol{q}, \boldsymbol{\mu}) \in \mathcal{I} \times \mathcal{T}$ be a bandit model, and $(\boldsymbol{q}, \boldsymbol{\lambda}) \in \mathcal{I} \times \text{Alt}(\boldsymbol{\mu})$ be an alternative bandit model. The key link between the two bandit models is their log-likelihood ratio up to time t, which can be written in the following way

$$\log\left(\frac{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\,|_{\mathcal{F}_t}}{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\,|_{\mathcal{F}_t}}\right) = \sum_{a=1}^{K}\sum_{s=0}^{N_a(t)-1}\log\frac{P_{\mu_a}(X_{a,s},X_{a,s+1})}{P_{\lambda_a}(X_{a,s},X_{a,s+1})} = \sum_{a=1}^{K}\sum_{x,y}N_a(x,y,0,t)\log\frac{P_{\mu_a}(x,y)}{P_{\lambda_a}(x,y)},$$

where $N_a(x,y,0,t)$ denotes the number of transitions from state x to state y that occurred from time 0 up to time t in the Markov chain with initial distribution q_a and transition probability function P_{μ_a}, i.e.

$$N_a(x,y,0,t) := \sum_{s=0}^{t-1} 1\{X_{a,s} = x, X_{a,s+1} = y\}.$$

Using the likelihood ratio we can perform a change of measure from the $(\boldsymbol{q}, \boldsymbol{\mu})$ model to the $(\boldsymbol{q}, \boldsymbol{\lambda})$ model for every fixed t, as well as for random stopping times as the following identity suggests. Its proof can be found in Appendix A.1.

Lemma 2 *Let τ be an almost surely finite stopping time with respect to $(\mathcal{F}_t)_{t\in\mathbb{Z}_{>0}}$, for both $(\boldsymbol{q}, \boldsymbol{\mu})$ and $(\boldsymbol{q}, \boldsymbol{\lambda})$. For every $X \in \mathcal{F}_\tau$ we have that*

$$\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\lambda})}[X] = \mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}\left[X\frac{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\,|_{\mathcal{F}_\tau}}{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\,|_{\mathcal{F}_\tau}}\right],$$

and in particular if we instantiate this with $X = 1_{\mathcal{E}}$ for some $\mathcal{E} \in \mathcal{F}_\tau$ we get that

$$\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}(\mathcal{E}) = \mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}\left[1_{\mathcal{E}}\frac{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\,|_{\mathcal{F}_\tau}}{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\,|_{\mathcal{F}_\tau}}\right].$$

In order to extend the lower bounding technique of Garivier and Kaufmann (2016), to the context of Markov chains we need the following Lemma which is a variant of Lemma 2.1 in Anantharam et al. (1987b), and its proof is based on a renewal argument given in Appendix A.3.

Lemma 3 *Let $X_0, X_1, \ldots, X_n, \ldots$ be an irreducible and positive recurrent Markov chain on a countable state space S, with initial distribution q, transition probability function P, and stationary distribution π. Assume that the mean return time of the chain is finite*

$$R = \mathbb{E}_q[\inf\{n > 0 : X_n = X_0\}] < \infty,$$

and define

$$N(x,y,n,m) = \sum_{s=n}^{m-1} 1\{X_s = x, X_{s+1} = y\},$$

the number of transitions from x to y that occurred from time n up to time m.

In addition let $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$ be the observe information up to time n, and let $\mathcal{G}$ be a σ-algebra which is independent of $\sigma(\cup_{n=0}^{\infty}\mathcal{F}_n)$. Let τ be a stopping time with respect to $(\sigma(\mathcal{F}_n \cup G))_{n\in\mathbb{N}}$, with $\mathbb{E}_q\,\tau < \infty$. Then

$$\mathbb{E}_q\, N(x, y, 0, \tau) \leq \pi(x)P(x, y)(\mathbb{E}_q\,\tau + R), \text{ for all } x, y \in S.$$

Some more notation is needed in order to be able to express those bandit models that exhibit a different best arm than the model in consideration. We define

$$\text{Alt}(\boldsymbol{\mu}) := \{\boldsymbol{\lambda} \in \mathcal{T} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}.$$

We are now ready to establish our non-asymptotic lower bound in the Markovian bandit setting.

Theorem 4 *Let $\delta \in (0, 1)$. For any δ-PC strategy and any Markovian bandit model $(\boldsymbol{q}, \boldsymbol{\mu}) \in \mathcal{I}\times\mathcal{T}$ such that the mean return times are finite*

$$R_a = \mathbb{E}_{q_a}\left[\inf\{n > 0 : X_{a,n} = X_{a,0}\}\right] < \infty, \text{ for } a = 1, \ldots, K,$$

we have that

$$T^*(\boldsymbol{\mu})D_2\left(\delta \parallel 1-\delta\right) - \sum_{a=1}^{K} R_a \leq \mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[\tau_\delta],$$

where

$$T^*(\boldsymbol{\mu})^{-1} := \sup_{w\in\mathcal{M}_1([K])} \inf_{\boldsymbol{\lambda}\in Alt(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a D\left(\mu_a \parallel \lambda_a\right).$$

Proof *Fix a δ-PC strategy $\mathcal{A}_\delta = ((A_t), \tau_\delta, \hat{a}_{\tau_\delta})$ and a bandit model $(\boldsymbol{q}, \boldsymbol{\mu}) \in \mathcal{I} \times \mathcal{T}$. Consider an alternative bandit model $(\boldsymbol{q}, \boldsymbol{\lambda}) \in \mathcal{I} \times Alt(\boldsymbol{\mu})$.*

The data processing inequality (see the book of Cover and Thomas (2006) for some context on the inequality) give us as a way to lower bound the divergence of the two models $\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\,|_{\mathcal{F}_{\tau_\delta}}$ and $\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\,|_{\mathcal{F}_{\tau_\delta}}$.

$$D_2\left(\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}(\mathcal{E}) \parallel \mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}(\mathcal{E})\right) \leq D\left(\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\,|_{\mathcal{F}_{\tau_\delta}} \,\middle\|\, \mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\,|_{\mathcal{F}_{\tau_\delta}}\right), \text{ for any } \mathcal{E} \in \mathcal{F}_{\tau_\delta}.$$

We apply this inequality with the event $\mathcal{E} = \{\hat{a}_{\tau_\delta} \neq a^*(\boldsymbol{\mu})\} \in \mathcal{F}_{\tau_\delta}$. The fact that the strategy $\mathcal{A}_\delta$ is δ-PC implies that

$$\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}(\mathcal{E}) \leq \delta, \quad \text{and} \quad \mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}(\mathcal{E}) \geq 1-\delta,$$

hence

$$D_2\left(\delta \parallel 1-\delta\right) \leq D\left(\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\,|_{\mathcal{F}_{\tau_\delta}} \,\middle\|\, \mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\,|_{\mathcal{F}_{\tau_\delta}}\right),$$

Using Lemma 3 we further have that

$$D\left(\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\,|_{\mathcal{F}_{\tau_\delta}} \,\middle\|\, \mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\,|_{\mathcal{F}_{\tau_\delta}}\right) = \mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}\left[\log\frac{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\,|_{\mathcal{F}_{\tau_\delta}}}{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\,|_{\mathcal{F}_{\tau_\delta}}}\right] \leq \sum_{a=1}^{K}\left(\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[N_a(\tau_\delta)] + R_a\right) D\left(\mu_a \parallel \lambda_a\right).$$

Combining those two we get that

$$D_2\left(\delta \parallel 1-\delta\right) \leq \sum_{a=1}^{K} \left(\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[N_a(\tau_\delta)] + R_a\right) D\left(\mu_a \parallel \lambda_a\right), \text{ for all } \boldsymbol{\lambda} \in Alt(\boldsymbol{\mu}).$$

The fact that $\sum_{a=1}^{K} N_a(\tau_\delta) \leq \tau_\delta$ *gives*

$$D_2\left(\delta \parallel 1-\delta\right) \leq \left(\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[\tau_\delta] + \sum_{a=1}^{K} R_a\right) \sum_{a=1}^{K} \frac{\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[N_a(\tau_\delta)] + R_a}{\sum_{b=1}^{K}\left(\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[N_b(\tau_\delta)] + R_b\right)} D\left(\mu_a \parallel \lambda_a\right),$$

and now we follow the technique of combining multiple alternative models $\boldsymbol{\lambda}$ *in order to obtain*

$$\begin{aligned} D_2\left(\delta \parallel 1-\delta\right) &\leq \left(\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[\tau_\delta] + \sum_{a=1}^{K} R_a\right) \inf_{\boldsymbol{\lambda} \in Alt(\boldsymbol{\mu})} \sum_{a=1}^{K} \frac{\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[N_a(\tau_\delta)] + R_a}{\sum_{b=1}^{K}\left(\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[N_b(\tau_\delta)] + R_b\right)} D\left(\mu_a \parallel \lambda_a\right) \\ &\leq \left(\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[\tau_\delta] + \sum_{a=1}^{K} R_a\right) \sup_{w \in \mathcal{M}_1([K])} \inf_{\boldsymbol{\lambda} \in Alt(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a D\left(\mu_a \parallel \lambda_a\right). \end{aligned}$$

■

Corollary 5 $D_2\left(\delta \parallel 1-\delta\right) \sim \log \frac{1}{\delta}$ *as* δ *goes to* 0*, and so Theorem 4 yields the asymptotic lower bound*

$$T^*(\boldsymbol{\mu}) \leq \liminf_{\delta \to 0} \frac{\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[\tau_\delta]}{\log \frac{1}{\delta}}.$$

It is worth mentioning that our asymptotic lower bound has no dependence on the initial distributions of the Markov chains, as one would expect because in the long run the effect of the initial distributions vanishes. In addition it generalizes the asymptotic lower bound of Garivier and Kaufmann (2016), where each arm is an i.i.d. sequence, to the Markovian setting.

As shown in Garivier and Kaufmann (2016) the supremum in the definition of $T^*(\mu)^{-1}$ is actually a maximum and we define

$$w^*(\boldsymbol{\mu}) := \operatorname*{arg\,max}_{w \in \mathcal{M}_1([K])} \inf_{\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a D\left(\mu_a \parallel \lambda_a\right).$$

Those weights $w^*(\boldsymbol{\mu})$ play an important role in the derivation of the (α, δ)-Track-and-Stop strategy as they represent the optimal proportions that we should sample the K Markov chains. For a development of the (α, δ)-Track-and-Stop strategy in the Markovian setting the interested reader can see Moulos (2019).

5. Conclusion

We developed instance specific non-asymptotic and asymptotic lower bounds for the problem of identifying the best Markovian arm with fixed confidence. In our subsequent work Moulos (2019) we analyze the (α, δ)-Track-and-Stop strategy in the Markovian setting, and we derive an upper bound which is a factor of four apart from the lower bound. A direction for future research is to eliminate this factor of four, and establish the exact sample complexity of the best Markovian arm identification problem.

References

V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multi-armed bandit problem with multiple plays-Part I: I.I.D. rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, November 1987a.

V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part II: Markovian rewards. *IEEE Transactions on Automatic Control*, 32(11):977–982, November 1987b.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. ISSN 1935-8237.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006. ISBN 978-0-471-24195-9; 0-471-24195-4.

Rick Durrett. *Probability: theory and examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fourth edition, 2010. ISBN 978-0-521-76539-8. doi: 10.1017/CBO9780511779398.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *J. Mach. Learn. Res.*, 7: 1079–1105, December 2006.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. *Proceedings of the 29th Conference On Learning Theory*, 49:1–30, January 2016.

Kevin G. Jamieson, Matthew Malloy, Robert D. Nowak, and Sébastien Bubeck. lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits. In *COLT*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 423–439, 2014.

Emilie Kaufmann and Wouter Koolen. Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals. 2018.

T.L Lai and Herbert Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985.

Shie Mannor and John N. Tsitsiklis. The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Vrettos Moulos. Optimal Best Markovian Arm Identification with Fixed Confidence. 2019. manuscript.

Appendix A. Lower Bound on the Sample Complexity

A.1. Proof of Lemma 2

It is straightforward to see that for each fixed t and $X \in \mathcal{F}_t$ we have that

$$\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\lambda})}[X] = \mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}\left[X \frac{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\,|_{\mathcal{F}_t}}{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\,|_{\mathcal{F}_t}}\right].$$

Now let τ be an almost surely finite stopping time with respect to $(\mathcal{F}_t)_{t\in\mathbb{Z}_{>0}}$, for both $\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}$ and $\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}$, and $X \in \mathcal{F}_\tau$. Then

$$\begin{aligned}\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\lambda})}[X] &= \sum_{t=0}^{\infty} \mathbb{E}_{(\boldsymbol{q},\boldsymbol{\lambda})}\left[\underbrace{X1\{\tau = t\}}_{\in\mathcal{F}_t})\right] \\ &= \sum_{t=0}^{\infty} \mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}\left[X1\{\tau = t\}\frac{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\mid_{\mathcal{F}_t}}{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\mid_{\mathcal{F}_t}}\right] \\ &= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}\left[X\frac{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\mid_{\mathcal{F}_\tau}}{d\,\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\mid_{\mathcal{F}_\tau}}\right].\end{aligned}$$

A.2. Markov Chains

Here we establish some facts about Markov chains, which are needed in the argument for the the lower bound. Let $(S, \mathcal{S})$ be a measure space, with S a countable state space, and $\mathcal{S} = 2^S$ the σ-field containing all subsets of S. This measure space can be extended to a product measure space on $(n+1)$-tuples $(\Omega_n, \mathcal{F}_n) = (S^{n+1}, \mathcal{S}^{\otimes(n+1)})$, as well as on sequences $(\Omega_\infty, \mathcal{F}_\infty) = (S^\infty, \mathcal{S}^{\otimes\infty})$. We let $(X_n,\ n \in \mathbb{N})$ be the coordinate process, i.e. $X_n(\omega) = \omega_n$ for $\omega \in S^\infty$. We fix a transition probability function $P : S \times \mathcal{S} \to [0, 1]$. For each initial probability distribution q on $(\Omega_0, \mathcal{F}_0)$ we define a probability distribution $\mathbb{P}_q\mid_{\mathcal{F}_n}$ on $(\Omega_n, \mathcal{F}_n)$ as

$$\mathbb{P}_q\mid_{\mathcal{F}_n}(X_0 = x_0, \ldots, X_n = x_n) = q(x_0)P(x_0, x_1)\cdots P(x_{n-1}, x_n).$$

By Kolmogorov's extension Theorem we can extend the finite dimensional distributions to a unique probability distribution $\mathbb{P}_q$ on $(\Omega_\infty, \mathcal{F}_\infty)$.

The fundamental *Markov property* can be written as

$$\mathbb{P}_q(X_{n+1} \in B \mid \mathcal{F}_n) = P(X_n, B), \text{ for any } B \in \mathcal{S}.$$

The Markov property can be extended to the so called *strong Markov property* which asserts that if τ is a $\mathbb{P}_q$-a.s. finite stopping time with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, then

$$\mathbb{P}_q(X_{\tau+1} \in B \mid \mathcal{F}_\tau) = P(X_\tau, B), \text{ for any } B \in \mathcal{S}.$$

For a more thorough discussion on the Markov and the strong Markov properties the interested reader is referred to Durrett (2010).

Using the strong Markov property we can prove a fundamental Lemma about Markov chains that reveals the i.i.d. structure that is present. Define recursively the k-th return time to the initial state as

$$\begin{cases}\tau_0 &= 0 \\ \tau_k &= \inf\{n > \tau_{k-1} : X_n = X_0\}, \text{ for } k \geq 1,\end{cases}$$

and for $k \geq 1$ let $r_k = \tau_k - \tau_{k-1}$ be the residual time.

Lemma 6 *If we further assume that the Markov chain is irreducible and recurrent, then those random times partition the Markov chain in a sequence $v_1, v_2, \ldots, v_k, \ldots$ of i.i.d. random blocks given by*

$$v_1 = (r_1, X_{\tau_0}, \ldots, X_{\tau_1 - 1}),\ v_2 = (r_2, X_{\tau_1}, \ldots, X_{\tau_2 - 1}), \ldots, v_k = (r_k, X_{\tau_{k-1}}, \ldots, X_{\tau_k - 1}), \ldots.$$

Proof *First note that due to recurrence τ_k is $\mathbb{P}_q$-a.s. finite*

$$\mathbb{P}_q(\tau_k < \infty) = \sum_{x \in S} q(X_0 = x) \underbrace{\mathbb{P}_x(\tau_k < \infty)}_{=1} = 1,$$

which will enable us to apply the strong Markov property. In addition observe that $v_k = v_1 \circ \theta_{\tau_{k-1}}$, and the block random variable v_k is a discrete random variable, since it can take on only countably many values. Let v be such a possible value, then the strong Markov property informs us that

$$\mathbb{P}_x(v_k = v \mid \mathcal{F}_{\tau_{k-1}}) = \mathbb{P}_x(v_1 \circ \theta_{\tau_{k-1}} = v \mid \mathcal{F}_{\tau_{k-1}}) = \mathbb{P}_x(v_1 = v), \text{ for each } x \in S,$$

and so

$$\mathbb{P}_q(v_k = v \mid \mathcal{F}_{\tau_{k-1}}) = \mathbb{P}_q(v_1 = v),$$

which means that for each $k \geq 1$, v_k is independent of $\mathcal{F}_{\tau_{k-1}}$, and so independent of $v_1, \ldots, v_{k-1}$, and has the same distribution as v_1.

■

We let $N(x, n, m)$ be the number of visits to x that occurred from time n up to (but not including) time m, and $N(x, y, n, m)$ to be the number of transitions from x to y that occurred from time n up to time m:

$$N(x, n, m) = \sum_{s=n}^{m-1} 1\{X_s = x\};$$

$$N(x, y, n, m) = \sum_{s=n}^{m-1} 1\{X_s = x, X_{s+1} = y\}.$$

It is well know, for instance see Durrett (2010), that if the Markov chain is irreducible and positive recurrent then it posses a unique stationary distribution π which satisfies the relation

$$\pi(x) = \frac{\mathbb{E}_{x_0} N(x, 0, \tau_1)}{\mathbb{E}_{x_0} \tau_1} = \frac{1}{\mathbb{E}_x \tau_1}, \text{ for all } x_0, x \in S.$$

In the following Lemma we establish a similar relation for the invariant distribution over pairs of the Markov chain.

Lemma 7 *Further assume that the Markov chain is irreducible and positive recurrent, so it posses a unique stationary distributions π. Then*

$$\pi(x)P(x, y) = \frac{\mathbb{E}_{x_0} N(x, y, 0, \tau_1)}{\mathbb{E}_{x_0} \tau_1}, \text{ for any } x_0, x, y \in S.$$

If in addition the initial distribution q is such that $\mathbb{E}_q \tau_1 < \infty$, then

$$\pi(x)P(x, y) = \frac{\mathbb{E}_q N(x, y, 0, \tau_1)}{\mathbb{E}_q \tau_1}, \text{ for any } x, y \in S.$$

Remark 8 *The assumption $\mathbb{E}_q\,\tau_1 < \infty$ is essential because $\pi(x) = \frac{1}{\mathbb{E}_x\,\tau_1}$, and so if we take $q = \pi$, and S is countably infinite, then $\mathbb{E}_\pi\,\tau_1 = \infty$.*

Proof *Since $\pi(x) = \frac{\mathbb{E}_{x_0} N(x,0,\tau_1)}{\mathbb{E}_{x_0}\,\tau_1}$, it is enough to show that*

$$\mathbb{E}_{x_0}\,N(x,0,\tau_1)P(x,y) = \mathbb{E}_{x_0}\,N(x,y,0,\tau_1),$$

or expanding out the definitions that

$$\mathbb{E}_{x_0}\sum_{n=0}^{\tau_1-1} 1\{X_n = x\}P(x,y) = \mathbb{E}_{x_0}\sum_{n=0}^{\tau_1-1} 1\{X_n = x, X_{n+1} = y\}.$$

Conditioning over the possible values of τ_1 and using Fubini's Theorem we obtain

$$\begin{aligned}
\mathbb{E}_{x_0}\sum_{n=0}^{\tau_1-1} 1\{X_n = x\}P(x,y) &= \sum_{t=1}^{\infty}\mathbb{P}_{x_0}(\tau_1 = t)\sum_{n=0}^{t-1}\mathbb{P}_{x_0}(X_n = x \mid \tau_1 = t)P(x,y)\\
&= \sum_{n=0}^{\infty}\sum_{t=n+1}^{\infty}\mathbb{P}_{x_0}(X_n = x, \tau_1 = t)P(x,y)\\
&= \sum_{n=0}^{\infty}\mathbb{P}_{x_0}(X_n = x, \tau_1 > n)P(x,y)\\
&= \sum_{n=0}^{\infty}\mathbb{P}_{x_0}(X_n = x, X_{n+1} = y)\,\mathbb{P}_{x_0}(\tau_1 > n \mid X_n = x)\\
&= \sum_{n=0}^{\infty}\mathbb{P}_{x_0}(X_n = x, X_{n+1} = y, \tau_1 > n)\\
&= \mathbb{E}_{x_0}\sum_{n=0}^{\tau_1-1} 1\{X_n = x, X_{n+1} = y\},
\end{aligned}$$

where the second to last equality holds true because because Markov chains satisfy a reversed Markov property as well and so

$$\mathbb{P}_{x_0}(\tau_1 > n \mid X_n = x, X_{n+1} = y) = \mathbb{P}_{x_0}(\tau_1 > n \mid X_n = x).$$

Finally, under the assumption $\mathbb{E}_q\,\tau_1 < \infty$ we conclude that

$$\pi(x)P(x,y) = \frac{\mathbb{E}_q\,N(x,y,0,\tau_1)}{\mathbb{E}_q\,\tau_1}, \text{ for any } x, y \in S.$$

■

A.3. Proof of Lemma 3

We are going to use the k-th return times

$$\begin{cases}\tau_0 &= 0\\ \tau_k &= \inf\{n > \tau_{k-1} : X_n = X_0\}, \text{ for } k \geq 1.\end{cases}$$

in order to decompose $N(x, y, 0, \tau_k)$ in k i.i.d. summands according to Lemma 6

$$N(x, y, 0, \tau_k) = \sum_{i=0}^{k-1} N(x, y, \tau_i, \tau_{i+1}).$$

Now let $\kappa = \inf\{k > 0 : \tau_k \geq \tau\}$, so that τ_κ is the first return time to the initial state after or at time τ. By definition of τ_κ we have the following two inequalities

$$\tau_\kappa - \tau \leq \tau_\kappa - \tau_{\kappa-1}, \text{ and } N(x, y, 0, \tau) \leq N(x, y, 0, \tau_\kappa).$$

Taking expectations in the first one we obtain

$$\mathbb{E}_q[\tau_\kappa - \tau] \leq \mathbb{E}_q[\tau_\kappa - \tau_{\kappa-1}] = \mathbb{E}_q\, r_1 = R,$$

which also gives that

$$\mathbb{E}_q\, \tau_\kappa \leq \mathbb{E}_q\, \tau + R < \infty.$$

This allows us to use Wald's identity, followed by Lemma 7, followed by Wald's idenity again, in order to get

$$\begin{aligned}
\mathbb{E}_q\, N(x, y, 0, \tau_\kappa) &= \mathbb{E}_q \sum_{i=0}^{\kappa-1} N(x, y, \tau_i, \tau_{i+1}) \\
&= \mathbb{E}_q\, N(x, y, 0, \tau_1)\, \mathbb{E}_q\, \kappa \\
&= p(x)P(x, y)\, \mathbb{E}_q\, \tau_1\, \mathbb{E}_q\, \kappa \\
&= p(x)P(x, y)\, \mathbb{E}_q\, \tau_\kappa.
\end{aligned}$$

Therefore

$$\mathbb{E}_q\, N(x, y, 0, \tau) \leq \mathbb{E}_q\, N(x, y, 0, \tau_\kappa) = \pi(x)P(x, y)\, \mathbb{E}_q\, \tau_\kappa \leq \pi(x)P(x, y)(\mathbb{E}_q\, \tau + R).$$