Optimal Best Markovian Arm Identification with Fixed Confidence

Vrettos Moulos  
Department of Electrical Engineering and Computer Sciences  
University of California Berkeley  
vrettos@berkeley.edu

Abstract

We give a complete characterization of the sampling complexity of best Markovian arm identification in one-parameter Markovian bandit models. We derive instance specific nonasymptotic and asymptotic lower bounds which generalize those of the IID setting. We analyze the Track-and-Stop strategy, initially proposed for the IID setting, and we prove that asymptotically it is at most a factor of four apart from the lower bound. Our one-parameter Markovian bandit model is based on the notion of an exponential family of stochastic matrices for which we establish many useful properties. For the analysis of the Track-and-Stop strategy we derive a novel and optimal concentration inequality for Markov chains that may be of interest in its own right.

1 Introduction

1.1 Contributions

This paper is about optimal best Markovian arm identification with fixed confidence. We consider the setting known as stochastic multi-armed bandits in reference to the gambling game. In particular we have \(K\) independent options which are referred to as arms, and our goal is to find the best arm as quickly as possible. Each arm \(a\) is associated with a discrete time stochastic process governed by the probability law \(P_a\). In prior work [EDMM06, JMN14, GK16] the discrete time stochastic process associated with each arm \(a\) is assumed to be an IID process. Here we go one step further and we study more complicated dependent processes, which allow us to use more expressive models in the stochastic multi-armed bandits framework. More specifically we consider the case that each \(P_a\) is the law of an irreducible finite state Markov chain associated with a stationary mean \(\mu_a\). Given some confidence level \(\delta \in (0, 1)\), our task is to identify the arm with the highest stationary mean as fast as possible and with confidence at least \(1 - \delta\). In this work we establish nonasymptotic (Theorem 2) and asymptotic (Corollary 1) lower bounds for the expected sample complexity, as well as an analysis of the Track-and-Stop strategy, proposed for the IID setting in [GK16], which shows (Theorem 4) that asymptotically the Track-and-Stop strategy in the Markovian dependence setting attains a sample complexity which is at most a factor of four apart from our asymptotic lower bound. Both our lower and upper bounds extend the work of [GK16] in the more complicated and more general Markovian dependence setting.

The abstract framework of multi-armed bandits has numerous applications in areas like clinical trials, ad placement, adaptive routing, resource allocation, gambling etc. For more context we refer the interested reader to the survey of [BCB12]. Here we generalize this model to allow for the presence of Markovian dependence, enabling this way the practitioner to use richer and more expressive models for the various applications. In particular, Markovian dependence allows models where the distribution of next sample depends on the sample just observed, and this way it can capture assertions
of the form: ‘if treatment \(a\) was effective the \(n\)-th time used, then it is very likely that it will be effective again the \(n + 1\)-th time used’.

Our key technical contributions stem from the large deviations theory for Markov processes [Mil61, DV75, Ell84, DZ98]. In particular we utilize the concept of an exponential family of stochastic matrices [Mil61] in order to model our one-parameter Markovian bandit model. Many properties of the family are established which are then used for our analysis of the Track-and-Stop strategy. The most prominent one is an optimal concentration inequality for the empirical means of Markov chains (Theorem 1). We are able to establish this inequality for a large class of Markov chains, including those that all the transitions have positive probability, and we improve on all the prior work [DLS81, Gil93, Din95, Lez98, LP04, CLLM12] providing essentially an unimprovable concentration inequality up to constant prefactors. This result may be of independent interest due to the wide applicability of Markov chains in many aspects of learning theory such as reinforcement learning, Markov decision processes, Markov chain Monte Carlo and others.

1.2 Related Work

The cornerstone of the stochastic multi-armed bandits literature is the seminal paper of Lai and Robbins [LR85]. In their work they consider \(K\) IID process with the objective being to maximize the expected value of the sum of the observed samples, or minimize the so called regret. In the same spirit [AVW87a, AVW87b] examine the generalization where one is allowed to get multiple samples at each time-step first in the case that processes are IID [AVW87a], and then in the case that the processes are irreducible and aperiodic Markov chains [AVW87b]. A survey about the regret minimization objective in the multi-armed bandits framework can be found in [BCB12].

An alternative objective 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 4. In the IID setting, [EDMM06] establish an elimination based algorithm in order to find an approximate best arm, and [MT04] provide a matching lower bound. [JMN14] 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 [GK16], 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 [KK18].

The earliest reference for the exponential family of stochastic matrices which is being used to model the Markovian arms can be found in the work of [Mil61]. Exponential families of stochastic matrices lie in the heart of the theory of large deviations for Markov processes, which was popularized with the pioneering work of [DV75]. A comprehensive overview of the theory can be found in the book [DZ98]. Naturally they also show up when one conditions on the second-order empirical distribution of a Markov chain, see the work of [CCC87] about conditional limit theorems. A variant of the exponential family that we are going to discuss has been developed in the context of hypothesis testing in [NK93]. A more recent development by [Nag05] gives an information geometry perspective to this concept, and the work [HW16] examines parameter estimation for the exponential family. Our development of the exponential family of stochastic matrices tries to parallel the development of simple exponential families of probability distributions of [WJ08].

Regarding concentration inequalities for Markov chains one of the earliest works [DLS81] is based on counting, and is able to capture the optimal rate of exponential decay dictated by the theory of large deviations, but has a suboptimal polynomial prefactor. More recent approaches follow the line of work started by [Gil93], who used matrix perturbation theory to derive a bound for reversible Markov chains. His bound attains a constant prefactor but with a suboptimal rate of exponential decay which depends on the spectral gap of the transition matrix. His work was later extended by [Din95, Lez98] but still with a sub-optimal rate. The work of [LP04] reduces the problem to a two state Markov chain, and attains the optimal rate only for the case of a two state Markov chain. [CLLM12] obtain rates that depend on the mixing time of the chain rather than the spectral gap, but which are still suboptimal. We improve upon all these works by providing an optimal concentration inequality in Theorem 1.
2 Exponential Family of Stochastic Matrices

We will be interested in discrete time Markov chains on a finite state space $S$. Let $\phi : S \rightarrow \mathbb{R}$ be a non-constant function, with $M := \max_x \phi(x)$ and $m := \min_x \phi(x)$. Based on $\phi$ we define two set of states, $S_M := \{x \in S : \phi(x) = M\}$ and $S_m := \{x \in S : \phi(x) = m\}$. Our goal is to create a family of Markov chains which can realize any stationary mean in the interval $(m, M)$, which will be later used in order to model the Markovian arms. Towards this goal we use as a 'generator' for our family, a stochastic matrix $P$ which satisfies the following properties:

1. $P$ is irreducible;
2. the sub-matrix of $P$ with rows and columns in $S_M$ is irreducible;
3. for every $x \in S - S_M$, there is a $y \in S_M$ such that $P(x, y) > 0$;
4. the sub-matrix of $P$ with rows and columns in $S_m$ is irreducible;
5. for every $x \in S - S_m$, there is a $y \in S_m$ such that $P(x, y) > 0$.

Denote this set of stochastic matrices by $P(\phi, S)$. For example, a positive stochastic matrix, i.e. one where all the transition probabilities are positive, belongs in $P(\phi, S)$.

Fix an initial distribution $q \in \mathcal{M}_1 (S)$, where with $\mathcal{M}_1 (S)$ we denote the set of probability distributions over $S$, and a stochastic matrix $P \in P(\phi, S)$. Let $X_0, X_1, \ldots, X_n, \ldots$ be a Markov chain which evolves according to the initial distribution $q$ and the stochastic matrix $P$. Since $P$ is irreducible it has a unique stationary distribution $\pi \in \mathcal{M}_1 (S)$, and we denote the stationary mean of the chain with $\pi(\phi) := \mathbb{E}_{X \sim \pi} \phi(X)$.

Let $\theta \in \mathbb{R}$ be the canonical parameter of the family and define the non-negative irreducible matrix $\tilde{P}_\theta(x, y) := P(x, y)e^{\theta \phi(y)}$ or $(\tilde{P})_\theta(x, y)$ more generally, which has the same transition structure as $P$. In order to normalize the nonnegative matrix $\tilde{P}_\theta$ and turn it into a stochastic matrix, we invoke the Perron-Frobenius theory. Let $\rho(\theta)$ (or $\rho(\tilde{P}_\theta)$) be the spectral radius of $\tilde{P}_\theta$, which from the Perron-Frobenius theory we know that is a simple eigenvalue of $\tilde{P}_\theta$, called the Perron-Frobenius eigenvalue, associated with unique left and right eigenvectors $u_\theta$, $v_\theta$ (or $u_{\tilde{P}_\theta}$, $v_{\tilde{P}_\theta}$) such that they are both positive, $\sum_x u_\theta(x) = 1$, and $\sum_x u_\theta(x)v_\theta(x) = 1$, see for instance Theorem 8.4.4 in the book [HJ13]. Define $A(\theta) := \log \rho(\theta)$ the log-Perron-Frobenius eigenvalue which plays a role similar to that of a log-moment-generating function. From $\tilde{P}_\theta$ we can define a family of non-negative irreducible matrices, parameterized by $\theta$, in the following way

$$P_\theta(x, y) := \frac{\tilde{P}_\theta(x, y)v_\theta(y)}{\rho(\theta)v_\theta(x)} = e^{\theta \phi(y) - A(\theta)} P(x, y) \frac{v_\theta(y)}{v_\theta(x)},$$

which are also stochastic, since

$$\sum_y P_\theta(x, y) = \frac{1}{\rho(\theta)v_\theta(x)} \sum_y \tilde{P}_\theta(x, y)v_\theta(y) = 1, \text{ for all } x \in S.$$

In addition their stationary distributions are given by

$$\pi_\theta(x) := u_\theta(x)v_\theta(x), \text{ for } x \in S,$$

since

$$\sum_x \pi_\theta(x)P_\theta(x, y) = \frac{v_\theta(y)}{\rho(\theta)} \sum_x u_\theta(x)\tilde{P}_\theta(x, y) = u_\theta(y)v_\theta(y) = \pi_\theta(y), \text{ for all } y \in S.$$

Note that the ‘generator’ stochastic matrix $P$, is the member of the family that corresponds to $\theta = 0$, i.e. $P = P_0$, $A(0) = 0$, and $\pi_0 = \pi$.

The following Lemma suggests that the family can be reparametrized using the mean parameters $\mu = \pi_\theta(\phi)$. More specifically $A'$ is a strictly increasing bijection between the set $\mathbb{R}$ of canonical parameters and the set $\mathcal{M} := \{\mu \in (m, M) : \pi_\theta(\phi) = \mu, \text{ for some } \theta \in \mathbb{R}\}$ of mean parameters. Therefore with some abuse of notation, for any $\mu \in \mathcal{M}$ we will write $u_\mu, v_\mu, P_\mu, \pi_\mu$ for $u_{A^\prime (\mu)}, v_{A^\prime (\mu)}, P_{A^\prime (\mu)}, \pi_{A^\prime (\mu)}$.

Lemma 1.
(a) $\rho(\theta)$, $A(\theta)$, $u_0$ and $v_0$ are analytic functions of $\theta$ on $\mathbb{R}$.

(b) $A'(\theta) = \pi_{\theta}(\phi)$, for all $\theta \in \mathbb{R}$.

(c) $A'(\theta)$ is strictly increasing.

(d) $M = (m, M)$.

In Lemma 1 we actually show that any mean in the interval $(m, M)$ can be realized as a stationary mean of some member of the family, i.e. $M = (m, M)$. We do this by analyzing the limiting behavior of the family, based on the assumptions that we made on $P$. In Appendix A we derive an extension of the Perron-Frobenius theory for matrices which may not be irreducible, and we use continuity of simple eigenvalues and eigenvectors to completely characterize the limiting elements of the family.

The notion of relative entropy will play a key role in the sample complexity of the best Markovian arm identification. For two probability distributions $Q, P$ over the same measure space we define

$$D(Q || P) := \begin{cases} \mathbb{E}_Q \left[ \log \frac{dQ}{dP} \right], & \text{if } Q \ll P \\ \infty, & \text{otherwise} \end{cases}$$

In particular for Markov chains we will be interested in the relative entropy rate between a Markov chain with irreducible stochastic matrix $Q$ and stationary distribution $\pi_Q$, and a Markov chain with stochastic matrix $P$, which can be defined as

$$D(Q || P) := D(\pi_Q \circ Q \parallel \pi_Q \circ P) = \sum_{x,y} \pi_Q(x)Q(x, y) \log \frac{Q(x, y)}{P(x, y)},$$

where $\pi_Q \circ Q \in M_1(S)$ denotes the bi-variate distribution $(\pi_Q \circ Q)(x, y) := \pi_Q(x)Q(x, y)$, and we use the standard notational conventions $\log 0 = \infty$, $\log \frac{a}{0} = \infty$ if $a > 0$, $0 \log 0 = 0 \log \frac{0}{a} = 0$. For Markov chains coming from the exponential family we will simplify our notation as $D(\theta_1 || \theta_2) := D(\pi_{\theta_1} \circ P_{\theta_1} \parallel \pi_{\theta_1} \circ P_{\theta_2})$, and $D(\mu_1 || \mu_2) := D(\pi_{\mu_1} \circ P_{\mu_1} \parallel \pi_{\mu_1} \circ P_{\mu_2})$.

In Appendix A various properties of this relative entropy rate are established, including primal and dual representations (Lemma 7), monotonicity (Corollary 2) and conjugate duality with the log-Perron-Frobenius eigenvalue $A(\theta)$ (Lemma 8), which are later used for the analysis of the Track-and-Stop strategy for the best Markovian arm identification.

3 Concentration for Markov Chains

In this section we introduce our key technical piece, which is later on used in Section 6 to analyze the Track-and-Stop strategy and derive an upper bound on its sample complexity.

We first recall that the large deviations theory for Markov processes [Mil61, DV75, Ell84, DZ98] suggests that if we pick any $\mu \in [\pi(\phi), M]$, then the probability of the large deviation event $\{\phi(X_1) + \ldots + \phi(X_n) \geq n\mu\}$ asymptotically is an exponential decay with the rate of the decay given by a relative entropy. In particular Theorem 3.1.2. from [DZ98] in our context suggests that

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(\phi(X_1) + \ldots + \phi(X_n) \geq n\mu) = -D(\mu \parallel \pi(\phi)), \text{ for any } \mu \in [\pi(\phi), M].$$

In the following Theorem we present a concentration inequality for Markov chains which attains the rate of exponential decay prescribed from the large deviations theory, as well as a constant prefactor which is independent of $\mu$.

**Theorem 1.** Let $S$ be a finite state space, and let $\phi : X \to \mathbb{R}$ be a non-constant function on the state space, with maximum value $M$ and minimum value $m$. Let $X_0, X_1, \ldots, X_n, \ldots$ be a Markov chain on $S$, with initial distribution $q \in M_1(X)$, an irreducible stochastic matrix $P \in \mathcal{P}(\phi, S)$, and stationary distribution $\pi \in M_1(S)$. There exists a constant $c = c(P, \phi) \geq 1$ such that

$$\text{for any } \mu \in [\pi(\phi), M], \quad \mathbb{P}(\phi(X_1) + \ldots + \phi(X_n) \geq n\mu) \leq \exp(-cD(\mu \parallel \pi(\phi))).$$

If in addition $P$ is a positive stochastic matrix then we can take $c = \max_{x,y,z} \frac{P(x,y,z)}{P(x,z)}$.
We note that in the special case that the process is an IID process the constant $c(P, \phi)$ can be taken to be 1, and thus our Theorem 1 generalizes the classical Cramer-Chernoff inequality from the IID setting.

Moreover our inequality is optimal up to the constant prefactor, since the exponential decay is unimprovable due to the large deviations theory, while with respect to the prefactor we can not expect anything better than a constant because otherwise we would contradict the central limit theorem for Markov chains. In particular, when our conditions on the stochastic matrix $P$ are met, our bound improves on all the prior work [DLS81, Gil93, Din95, Lez98, LP04, CLLM12].

For example in the work of [Lez98] matrix perturbation theory is used in order to obtain in Theorem 3.3, that when $P^*P$ is irreducible, where $P^*(x, y) = \pi(y)P(y, x)/\pi(x)$, (a condition that will be true if $P(x, x) > 0$ for all $x \in S$) we have the following bound

$$
\mathbb{P}(\phi(X_1) + \ldots + \phi(X_n) \geq n\mu) \leq N \exp \left\{ -n \frac{(\mu - \pi(\phi))^2}{24 \max\{m^2, M^2\}} \right\}, \text{ for all } \mu \in [\pi(\phi), M],
$$

where $N = N(q, \pi)$ is a constant depending only on the initial distribution $q$ and the stationary distribution $\pi$ and $c(P^*P)$ denotes the spectral gap of the matrix $P^*P$. The large deviations theory suggest that the rate of exponential decay is suboptimal, and the following Pinsker type inequality holds true

$$
D(\mu \parallel \pi(\phi)) \geq \frac{(\mu - \pi(\phi))^2}{24 \max\{m^2, M^2\}}, \text{ for all } \mu \in [m, M].
$$

We give a full proof of Theorem 1 in Appendix B, where the main techniques involved are a uniform upper bound on the ratio of the entries of the right Perron-Frobenius eigenvector, as well as an approximation of the log-Perron-Frobenius eigenvalue using the log-moment-generating function.

### 4 The Markovian Bandit Model

Our Markovian bandit model consists of $K$ irreducible Markov chains, each determined by an initial distribution and a stochastic matrix from the exponential family with $\phi(x) = x$ being just the identity 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 the set of vectors such that for each $\mu = (\mu_1, \ldots, \mu_K) \in \mathcal{T}$ there exists an $a^*(\mu) \in \{1, \ldots, K\}$ such that $\mu_{a^*(\mu)} > \mu_q$ for all $a \neq a^*(\mu)$. Each $\mu = (\mu_1, \ldots, \mu_K) \in \mathcal{T}$ should be thought of as a vector $(P_{\mu_1}, \ldots, P_{\mu_K})$ of $K$ elements of the exponential family of stochastic matrices, with a single of them possessing the highest stationary mean.

A Markovian bandit model is a pair

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

The evolution of each arm $a = 1, \ldots, K$ is completely determined by its initial distribution $q_a \in \mathcal{M}_1(S)$ and its stochastic matrix $P_{\mu_a} \in \mathcal{P}(x \mapsto x, S)$. Each arm $a$ transitions only when it is pulled, and samples coming from arm $a$ are denoted by $X_{a,0}, X_{a,1}, \ldots, X_{a,n}, \ldots$. 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.

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

Our goal is to identify the best arm $a^*(\mu)$ 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 $\delta$ we need to come up with a strategy $\mathcal{A}_\delta$ which is a triple $\mathcal{A}_\delta = ((A_t)_{t \in \mathbb{Z}_{t>0}}, \tau_\delta, \hat{a}_\tau)$ consisting of:

- a **sampling rule** $(A_t)_{t \in \mathbb{Z}_{t>0}}$, 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** $\tau_\delta$, which denotes the end of the data collection phase and is stopping time with respect to $(\mathcal{F}_t)_{t \in \mathbb{Z}_{t>0}}$, such that $\mathbb{E}(q, \mu) \tau_\delta < \infty$ for all $(q, \mu) \in \mathcal{I} \times \mathcal{T}$;
- a **decision rule** $\hat{a}_\tau$, which is $\mathcal{F}_{\tau_\delta}$-measurable, and determines the arm that we estimate to be the best one.
So if we use strategy $A_δ$, 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 $δ \in (0, 1)$, a strategy $A_δ = ((A_t)_{t \in \mathbb{Z}_{\geq 0}}, \tau_δ, \hat{a}_\tau)$ is called δ-PC (Probably Correct) if,

$$P((q,\mu) (\hat{a}_{\tau_δ} \neq a^*(\mu)) \leq δ, \text{ for all } (q, \mu) \in \mathcal{I} \times \mathcal{T}.\]$$

### 5 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 [LR85].

We fix some Markovian bandit model $(q, \mu) \in \mathcal{I} \times \mathcal{T}$, and we define the set of models that exhibit a different behavior

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

Then we consider an alternative model $(q, \lambda) \in \mathcal{I} \times \mathrm{Alt}(\mu)$ and we write down their log-likelihood ratio up to time $t$

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

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 $P_{\mu_a}$, i.e.

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

The likelihood ratio enables us to perform the change of measure for fixed times $t$, and more generally for almost surely finite stopping times $\tau$ with respect to $(\mathcal{F}_t)_{t \in \mathbb{Z}_{\geq 0}}$, through the following change of measure formula

$$P_{(q,\lambda)}(\mathcal{E}) = E_{(q,\mu)} \left[ 1_{\mathcal{E}} \frac{d P_{(q,\lambda)}(\mathcal{F}_\tau) \mid \mathcal{F}_\tau}{d P_{(q,\mu)}(\mathcal{F}_\tau) \mid \mathcal{F}_\tau} \right], \text{ for any } \mathcal{E} \in \mathcal{F}_\tau.$$ 

In order to derive our lower bound we use a technique developed for the IID case by [GK16] which combines several change of measures at once. To make this technique work in the Markovian setting we need the following inequality which we derive in Appendix C using a renewal argument for Markov chains. In our nonasymptotic lower bound a quantity that shows up in the expected sample complexity is the mean return time of each Markovian arm

$$R_a := E_{(q_a,\mu_a)} \left[ \inf \{n > 0 : X_{a,n} = X_{a,0}\} \right], \text{ for } a = 1, \ldots, K,$$

which of course is finite since our Markov chains are over a finite state space and they are irreducible.

**Lemma 2.** Let $(q, \mu) \in \mathcal{I} \times \mathcal{T}$ and $(q, \lambda) \in \mathcal{I} \times \mathrm{Alt}(\mu)$ be two Markovian bandit models. Then

$$D \left( P_{(q,\mu)}(\mathcal{F}_{\tau_δ}) \mid P_{(q,\lambda)}(\mathcal{F}_{\tau_δ}) \right) \leq \sum_{a=1}^{K} (E_{(q,\mu)}[N_a(\tau_δ)] + R_a) D (\mu_a \parallel \lambda_a).$$

Combining those ingredients with the data processing inequality we derive our nonasymptotic instance specific lower bound for the Markovian bandit identification problem in Appendix C.
Theorem 2. Let $\delta \in (0, 1)$. For any $\delta$-PC strategy and any Markovian bandit model $(\mathbf{q}, \mathbf{\mu}) \in \mathcal{I} \times \mathcal{T}$ we have that

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

where

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

and with $D_2 (\delta \parallel 1 - \delta)$ we denote the binary relative entropy

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

Since $D_2 (\delta \parallel 1 - \delta) \sim \log \frac{1}{\delta}$ as $\delta$ goes to 0, an immediate corollary of Theorem 2 is the following asymptotic lower bound.

Corollary 1.

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

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 [GK16], where each arm is an IID sequence, to the Markovian setting.

The supremum in the definition of $T^*(\mathbf{\mu})^{-1}$ is actually a maximum and we define

$$w^*(\mathbf{\mu}) := \arg \max_{\mathbf{w} \in \mathcal{M}_1([K])} \inf_{\lambda \in \mathcal{A}(\mathbf{\mu})} \sum_{a=1}^{K} w_a D (\mu_a \parallel \lambda_a).$$

Those weights $w^*(\mathbf{\mu})$ play an important role in the derivation of the $(\alpha, \delta)$-Track-and-Stop strategy which gives the upper bound on the sample complexity. It is the form of the lower bound which reveals that an optimal strategy should pull arm $a$, $w^*_a(\mathbf{\mu})$ fraction of the time.

6 Upper Bound on the Sample Complexity: the $(\alpha, \delta)$-Track-and-Stop Strategy

The $(\alpha, \delta)$-Track-and-Stop strategy, which was proposed in [GK16] in order to tackle the IID setting, tries to track the optimal weights $w^*_a(\mathbf{\mu})$. Not having access to $\mathbf{\mu}$, it tries to approximate it using sample means. Let $\hat{\mathbf{\mu}}(t) := (\hat{\mu}_1(N_1(t)), \ldots, \hat{\mu}_K(N_K(t)))$ be the sample means of the $K$ Markov chains when $t$ samples have been observed overall and the calculation of the very first sample from each Markov chain is excluded from the calculation of its sample mean, i.e.

$$\hat{\mu}_a(t) := \frac{1}{N_a(t)} \sum_{s=1}^{N_a(t)} X_{a,s}.$$  

By imposing sufficient exploration the law of large numbers for Markov chains will kick in and the sample means $\hat{\mu}(t)$ will almost surely converge to the true means $\mathbf{\mu}$, as $t \to \infty$.

We proceed by briefly describing the three components of the $(\alpha, \delta)$-Track-and-Stop strategy.

6.1 Sampling Rule: Tracking the Optimal Proportions

For initialization reasons the first $2K$ samples that we are going to select are $X_{1,0}, X_{1,1}, \ldots, X_{K,0}, X_{K,1}$. After that, for $t \geq 2K$ we let $U_t = \{ a : N_a(t) < \sqrt{t - K} / 2 \}$ and we follow the tracking rule:

$$A_{t+1} \in \begin{cases} 
\arg \min_{a \in U_t} N_a(t), & \text{if } U_t \neq \emptyset \quad \text{(forced exploration)} \\
\arg \max_{a=1, \ldots, K} \left\{ w^*_a(\hat{\mathbf{\mu}}(t)) - \frac{N_a(t)}{t} \right\}, & \text{otherwise (direct tracking)} 
\end{cases}$$

7
The forced exploration step is there to ensure that \( \hat{\mu}(t) \overset{a.s.}{\rightarrow} \mu \) as \( t \to \infty \). Then the continuity of \( \mu \mapsto w_a^*(\mu) \), combined with the direct tracking step guarantees that almost surely the frequencies \( \frac{N_a(t)}{t} \) converge to the optimal weights \( w_a^*(\mu) \) for all \( a = 1, \ldots, K \).

6.2 Stopping Rule: \((\alpha, \delta)\)-Chernoff’s Stopping Rule

For the stopping rule we will need the following statistics: for any two distinct arms \( a, b \) if \( \hat{\mu}_a(N_a(t)) \geq \hat{\mu}_b(N_b(t)) \), we define

\[
Z_{a,b}(t) := \frac{N_a(t)}{N_a(t) + N_b(t)} D(\hat{\mu}_a(N_a(t)) \parallel \hat{\mu}_{a,b}(N_a(t), N_b(t))) + \frac{N_b(t)}{N_a(t) + N_b(t)} D(\hat{\mu}_b(N_b(t)) \parallel \hat{\mu}_{a,b}(N_a(t), N_b(t)));
\]

while if \( \hat{\mu}_a(N_a(t)) < \hat{\mu}_b(N_b(t)) \), we define \( Z_{a,b}(t) := -Z_{b,a}(t) \), where

\[
\hat{\mu}_{a,b}(N_a(t), N_b(t)) := \frac{N_a(t)}{N_a(t) + N_b(t)} \hat{\mu}_a(N_a(t)) + \frac{N_b(t)}{N_a(t) + N_b(t)} \hat{\mu}_b(N_b(t)).
\]

Note that the statistics \( Z_{a,b}(t) \) do not arise as the closed form solutions of the Generalized Likelihood Ratio statistics for Markov chains, as it is the case in the IID bandits setting.

For a confidence level \( \delta \in (0, 1) \), and a convergence parameter \( \alpha > 1 \) we define the \((\alpha, \delta)\)-Chernoff stopping rule following [GK16]

\[
\tau_{a,\delta} := \inf \{ t \in \mathbb{N} : \exists a \in \{1, \ldots, K\} \forall b \neq a, Z_{a,b}(t) > (0 \wedge \beta_{a,\delta}(t)) \},
\]

where \( \beta_{a,\delta}(t) := 2 \log \frac{C_{1,\alpha}}{2} \), \( C := \frac{2\alpha K c^2}{\alpha - 1} \), and \( c = c(P) \) is the constant coming from the concentration inequality Theorem 1. In the special case that \( P \) is a positive stochastic matrix we can explicitly set \( c := \max_{x,y,z} P(y,z) \frac{P(x,z)}{P(x,z)} \). It is important to notice that the constant \( c = c(P) \) does not depend on the bandit instance \((q, \mu)\) or the confidence level \( \delta \), but only on the generator stochastic matrix \( P \) of the exponential family. It is worth mentioning that the knowledge of \( c = c(P) \) does not reveal any information about the bandit instance \((q, \mu)\), but it’s rather a property of the exponential family generated by the stochastic matrix \( P \) and the identity reward function \( \phi(x) = x \). In the same way that relative entropy rate, \( D(\mu \parallel \mu') \), between two members of the family only depends on the exponential family, and it doesn’t leak any information about the particular \( \mu = (\mu_1, \ldots, \mu_K) \) used as the stationary means of the \( K \) arms.

6.3 Decision Rule: Best Sample Mean

Fix an arm \( a \). Clearly \( \min_{b \neq a} Z_{a,b}(t) > 0 \) iff \( \hat{\mu}_a(N_a(t)) > \hat{\mu}_b(N_b(t)) \) for all \( b \neq a \). Hence the following simple decision rule is well defined when used in conjunction with the \((\alpha, \delta)\)-Chernoff stopping rule:

\[
\hat{a}_{\tau_{a,\delta}} := \arg \max_{a=1, \ldots, K} \hat{\mu}_a(N_a(\tau_{a,\delta})).
\]

6.4 Sample Complexity Analysis

The contributions of this sections is establishing that the \((\alpha, \delta)\)-Track-and-Stop strategy is \( \delta \)-PC, as well as an upper bound on its expected sample complexity. For both of them we use our Markovian concentration bound Theorem 1.

We first use it in order to establish the following uniform over time deviation Lemma.

Lemma 3. Let \((q, \mu) \in \mathcal{I} \times S\), \( \delta \in (0, 1) \), and \( \alpha > 1 \). For any sampling rule, if we use the \((\alpha, \delta)\)-Chernoff’s stopping rule and the best sample mean decision rule we are guaranteed that

\[
\mathbb{P}(q, \mu) \left( \exists t \in \mathbb{Z}_{>0} : N_a(t) D(\hat{\mu}_a(N_a(t)) \parallel \mu_a) \geq \beta_{a,\delta}(t)/2 \right) \leq \frac{\delta}{K}, \text{ for any } a = 1, \ldots, K.
\]

With this in our possession we are able to prove in Appendix D that
Theorem 3. Let $\delta \in (0, 1)$, and $\alpha \in (1, e/4]$. The $(\alpha, \delta)$-Track-and-Stop strategy is $\delta$-PC.

Finally, we obtain that in the high confidence regime of $\delta \to 0$, the $(\alpha, \delta)$-Track-and-Stop strategy has a sample complexity which is at most $4\alpha$ times the asymptotic lower bound that we established in Corollary 1.

Theorem 4. Let $(q, \mu) \in I \times S$, and $\alpha \in (1, e/4]$. The $(\alpha, \delta)$-Track-and-Stop strategy has asymptotic expected sample complexity at most

$$\limsup_{\delta \to 0} \frac{\mathbb{E}[q, \mu][\tau_{\alpha, \delta}]}{\log \frac{1}{\delta}} \leq 4\alpha T^* (\mu).$$

References


