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 $\mathbb{P}_a$. In prior work [EDMM06, JMNB14, 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 $\mathbb{P}_a$ is the law of an irreducible finite state Markov chain associated with a stationary mean μ_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

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

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. [JMNB14] 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 $\phi : S \to \mathbb{R}$ be a non-constant function, with $M := \max_x \phi(x)$ and $m := \min_x \phi(x)$. Based on ϕ 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 $\mathcal{P}(\phi, S)$. For example, a positive stochastic matrix, i.e. one where all the transition probabilities are positive, belongs in $\mathcal{P}(\phi, S)$.

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 $P \in \mathcal{P}(\phi, S)$. 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 $\widetilde{(P)}_\theta(x, y)$ 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, parametrized by θ, 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

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

In addition their stationary distributions are given by

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

since

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

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 $u_{A'^{-1}(\mu)}, v_{A'^{-1}(\mu)}, P_{A'^{-1}(\mu)}, \pi_{A'^{-1}(\mu)}$.

Lemma 1.

(a) $\rho(\theta),\ A(\theta),\ u_\theta$ *and* v_θ *are analytic functions of* θ *on* $\mathbb{R}$.

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

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

(d) $\mathcal{M} = (m, M)$.

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. $\mathcal{M} = (m, M)$. 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 $\mathbb{Q}, \mathbb{P}$ over the same measure space we define

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

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 π_Q, and a Markov chain with stochastic matrix P, which can be defined as

$$D\left(Q \parallel P\right) := D\left(\pi_Q \odot Q \parallel \pi_Q \odot P\right) = \sum_{x,y} \pi_Q(x) Q(x,y) \log \frac{Q(x,y)}{P(x,y)},$$

where $\pi_Q \odot Q \in \mathcal{M}_1\left(S\right)$ denotes the bi-variate distribution

$$(\pi_Q \odot Q)(x,y) := \pi_Q(x) Q(x,y),$$

and we use the standard notational conventions $\log 0 = \infty,\ \log \frac{\alpha}{0} = \infty$ if $\alpha > 0,\ 0 \log 0 = 0 \log \frac{0}{0} = 0$. For Markov chains coming from the exponential family we will simplify our notation as $D\left(\theta_1 \parallel \theta_2\right) := D\left(\pi_{\theta_1} \odot P_{\theta_1} \parallel \pi_{\theta_1} \odot P_{\theta_2}\right)$, and $D\left(\mu_1 \parallel \mu_2\right) := D\left(\pi_{\mu_1} \odot P_{\mu_1} \parallel \pi_{\mu_1} \odot P_{\mu_2}\right)$.

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

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

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 μ.

Theorem 1. *Let* S *be a finite state space, and let* $\phi : \mathcal{X} \to \mathbb{R}$ *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* $q \in \mathcal{M}_1\left(\mathcal{X}\right)$, *an irreducible stochastic matrix* $P \in \mathcal{P}(\phi, S)$, *and stationary distribution* $\pi \in \mathcal{M}_1\left(S\right)$. *There exists a constant* $c = c(P, \phi) \geq 1$ *such that*

$$\left.\begin{array}{ll} \text{for any } \mu \in [\pi(\phi), M], & \mathbb{P}\left(\phi(X_1) + \ldots + \phi(X_n) \geq n\mu\right) \\ \text{for any } \mu \in [m, \pi(\phi)], & \mathbb{P}\left(\phi(X_1) + \ldots + \phi(X_n) \leq n\mu\right) \end{array}\right\} \leq c e^{-nD(\mu \parallel \pi(\phi))}.$$

If in addition P *is a positive stochastic matrix then we can take* $c = \max_{x,y,z} \frac{P(y,z)}{P(x,z)}$.

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

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

where $N = N(q, \pi)$ is a constant depending only on the initial distribution q and the stationary distribution π and $\epsilon(P^*P)$ 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

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

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 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in \mathcal{T}$ 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 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in \mathcal{T}$ 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

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

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 $a^*(\boldsymbol{\mu})$ 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 ($\boldsymbol{\delta}$-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,*

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

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 $(\boldsymbol{q},\boldsymbol{\mu}) \in \mathcal{I}\times\mathcal{T}$, and we define the set of models that exhibit a different behavior

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

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

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

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.

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

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

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

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

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

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

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

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

Combining those 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 δ-PC strategy and any Markovian bandit model $(\boldsymbol{q}, \boldsymbol{\mu}) \in \mathcal{I} \times \mathcal{T}$ we have that*

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

where

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

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 δ goes to 0, an immediate corollary of Theorem 2 is the following asymptotic lower bound.

Corollary 1.

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

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

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

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

Those weights $w^*(\boldsymbol{\mu})$ play an important role in the derivation of the (α, δ)-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, $w_a^*(\boldsymbol{\mu})$ fraction of the time.

6 Upper Bound on the Sample Complexity: the (α, δ)-Track-and-Stop Strategy

The (α, δ)-Track-and-Stop strategy, which was proposed in [GK16] in order to tackle the IID setting, tries to track the optimal weights $w_a^*(\boldsymbol{\mu})$. Not having access to $\boldsymbol{\mu}$, it tries to approximate it using sample means. Let $\hat{\boldsymbol{\mu}}(t) := (\hat{\mu}_1(N_1(t)), \ldots, \hat{\mu}_K(N_K(t)))$ 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 $\hat{\boldsymbol{\mu}}(t)$ will almost surely converge to the true means $\boldsymbol{\mu}$, as $t \to \infty$.

We proceed by briefly describing the three components of the (α, δ)-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:

$$A_{t+1} \in \begin{cases} \operatorname*{arg\,min}_{a \in U_t} N_a(t), & \text{if } U_t \neq \emptyset \quad \text{(forced exploration)} \\ \operatorname*{arg\,max}_{a=1,\ldots,K} \left\{ w_a^*(\hat{\boldsymbol{\mu}}(t)) - \dfrac{N_a(t)}{t} \right\}, & \text{otherwise} \quad \text{(direct tracking)} \end{cases}$$

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

6.2 Stopping Rule: (α, δ)-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

$$Z_{a,b}(t) := \frac{N_a(t)}{N_a(t) + N_b(t)} D\left(\hat{\mu}_a(N_a(t)) \parallel \hat{\mu}_{a,b}(N_a(t), N_b(t))\right) + \frac{N_b(t)}{N_a(t) + N_b(t)} D\left(\hat{\mu}_b(N_b(t)) \parallel \hat{\mu}_{a,b}(N_a(t), N_b(t))\right),$$

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 (α, δ)-Chernoff stopping rule following [GK16]

$$\tau_{\alpha,\delta} := \inf \left\{ t \in \mathbb{N} : \exists a \in \{1, \dots, K\} \, \forall b \neq a, \; Z_{a,b}(t) > (0 \vee \beta_{\alpha,\delta}(t)) \right\},$$

where $\beta_{\alpha,\delta}(t) := 2 \log \frac{C t^\alpha}{\delta}$, $C := \dfrac{2\alpha K c^2}{\alpha - 1}$, 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 $c := \max_{x,y,z} \frac{P(y,z)}{P(x,z)}$. It is important to notice that the constant $c = c(P)$ does not depend on the bandit instance $(\boldsymbol{q}, \boldsymbol{\mu})$ or the confidence level δ, 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 $(\boldsymbol{q}, \boldsymbol{\mu})$, 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 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ 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 (α, δ)-Chernoff stopping rule:

$$\hat{a}_{\tau_{\alpha,\delta}} := \underset{a=1,\dots,K}{\arg\max} \; \hat{\mu}_a(N_a(\tau_{\tau_{\alpha,\delta}})).$$

6.4 Sample Complexity Analysis

The contributions of this sections is establishing that the (α, δ)-Track-and-Stop strategy is δ-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 $(\boldsymbol{q}, \boldsymbol{\mu}) \in \mathcal{I} \times \mathcal{S}$, $\delta \in (0, 1)$, and $\alpha > 1$. For any sampling rule, if we use the (α, δ)-Chernoff's stopping rule and the best sample mean decision rule we are guaranteed that*

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

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 (α, δ)-Track-and-Stop strategy is δ-PC.*

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

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

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

References

[AVW87a] Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. I. I.I.D. rewards. *IEEE Trans. Automat. Control*, 32(11):968–976, 1987.

[AVW87b] Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. II. Markovian rewards. *IEEE Trans. Automat. Control*, 32(11):977–982, 1987.

[BCB12] 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.

[CCC87] Imre Csiszár, Thomas M. Cover, and Byoung Seon Choi. Conditional limit theorems under Markov conditioning. *IEEE Trans. Inform. Theory*, 33(6):788–801, 1987.

[CLLM12] Kai-Min Chung, Henry Lam, Zhenming Liu, and Michael Mitzenmacher. Chernoff-Hoeffding bounds for Markov chains: generalized and simplified. In *29th International Symposium on Theoretical Aspects of Computer Science*, volume 14 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages 124–135. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2012.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006.

[Din95] I. H. Dinwoodie. A probability inequality for the occupation measure of a reversible Markov chain. *Ann. Appl. Probab.*, 5(1):37–43, 1995.

[DLS81] Lee D. Davisson, Giuseppe Longo, and Andrea Sgarro. The error exponent for the noiseless encoding of finite ergodic Markov sources. *IEEE Trans. Inform. Theory*, 27(4):431–438, 1981.

[Dur10] Rick Durrett. *Probability: theory and examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fourth edition, 2010.

[DV75] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. I. II. *Comm. Pure Appl. Math.*, 28:1–47; ibid. 28 (1975), 279–301, 1975.

[DZ98] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998.

[EDMM06] 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, 2006.

[Ell84] Richard S. Ellis. Large deviations for a general class of random vectors. *Ann. Probab.*, 12(1):1–12, 1984.

[Gil93] David Gillman. A Chernoff bound for random walks on expander graphs. In *34th Annual Symposium on Foundations of Computer Science (Palo Alto, CA, 1993)*, pages 680–691. IEEE Comput. Soc. Press, Los Alamitos, CA, 1993.

[GK16] 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.

[HJ13] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[HW16] Masahito Hayashi and Shun Watanabe. Information geometry approach to parameter estimation in Markov chains. *Ann. Statist.*, 44(4):1495–1535, 2016.

[JMNB14] 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.

[KK18] Emilie Kaufmann and Wouter Koolen. Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals. 2018.

[Lax07] Peter D. Lax. *Linear algebra and its applications*. Pure and Applied Mathematics (Hoboken). Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2007.

[Lez98] Pascal Lezaud. Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.*, 8(3):849–867, 1998.

[LP04] Carlos A. León and François Perron. Optimal Hoeffding bounds for discrete reversible Markov chains. *Ann. Appl. Probab.*, 14(2):958–970, 2004.

[LR85] T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.*, 6(1):4–22, 1985.

[Mil61] H. D. Miller. A convexity property in the theory of random variables defined on a finite Markov chain. *Ann. Math. Statist.*, 32:1260–1270, 1961.

[MT04] Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.*, 5:623–648, 2003/04.

[Nag05] H. Nagaoka. The exponential family of Markov chains and its information geometry. In *Proceedings of The 28th Symposium on Information Theory and Its Applications (SITA2005)*, pages 1091–1095, Okinawa, Japan, November 2005.

[NK93] Kenji Nakagawa and Fumio Kanaya. On the converse theorem in statistical hypothesis testing for Markov chains. *IEEE Trans. Inform. Theory*, 39(2):629–633, 1993.

[Ort90] James M. Ortega. *Numerical analysis*, volume 3 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1990. A second course.

[WJ08] M. J. Wainwright and M. I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, January 2008.

A Exponential Family of Stochastic Matrices

We start by establishing parts (a), (b) and (c) of Lemma 1.

Proof. (Lemma 1)

(a) Each entry of $\tilde{P}_\theta$ is a real analytic function of θ, and for each θ_0 the Perron-Frobenius eigenvalue $\rho(\theta_0)$ is simple with a unique corresponding left and right eigenvectors $u_{\theta_0},\ v_{\theta_0}$ and such that they are both positive, $\sum_x u_{\theta_0}(x) = 1$ and $\sum_x u_{\theta_0}(x)v_{\theta_0}(x) = 1$. The conclusion follows by standard implicit function theorem type of arguments. See for example Theorem 7 and Theorem 8 in Chapter 9 from the book of [Lax07].

(b) For any $x, y \in S$ such that $P(x, y) > 0$ we have that

$$\log P_\theta(x, y) = \theta\phi(y) - A(\theta) + \log v_\theta(y) - \log v_\theta(x) + \log P(x, y).$$

Differentiating with respect to θ, and taking expectation with respect to $\pi_\theta \odot P_\theta$ we obtain

$$\mathbb{E}_{(X,Y)\sim\pi_\theta\odot P_\theta} \frac{d}{d\theta} \log P_\theta(X, Y) = \pi_\theta(\phi) - A'(\theta),$$

where the logarithms cancel out since $\pi_\theta \odot P_\theta$ has identical marginals. The conclusion follows because

$$\mathbb{E}_{(X,Y)\sim\pi_\theta\odot P_\theta} \frac{d}{d\theta} \log P_\theta(X, Y) = \sum_x \pi_\theta(x) \frac{d}{d\theta} \left(\sum_y P_\theta(x, y) \right) = 0.$$

(c) For any $x, y \in S$ such that $P(x, y) > 0$ we have that

$$\frac{d^2}{d\theta^2} \log P_\theta(x, y) = -A''(\theta) + \frac{d^2}{d\theta^2} \log v_\theta(y) - \frac{d^2}{d\theta^2} \log v_\theta(x).$$

Taking expectation with respect to $\pi_\theta \odot P_\theta$ we obtain

$$A''(\theta) = -\mathbb{E}_{(X,Y)\sim\pi_\theta\odot P_\theta} \frac{d^2}{d\theta} \log P_\theta(X, Y) = \mathbb{E}_{(X,Y)\sim\pi_\theta\odot P_\theta} \left(\frac{d}{d\theta} \log P_\theta(X, Y) \right)^2 \geq 0.$$

This ensures that $A'(\theta)$ is increasing.

Assume, towards contradiction, that $A''(\theta) = 0$ in a neighborhood of θ_0. Then P_θ does not depend on θ in a neighborhood of θ_0. Tthe S_M component is irreducible so we can find $x_1, \ldots, x_{l+1} \in S_M$ such that $P(x_i, x_{i+1}) > 0$ for $i = 1, \ldots, l$ and $x_1 = x_{l+1}$, and so

$$P_\theta(x_1, x_2) \ldots P_\theta(x_l, x_{l+1}) = \frac{P(x_1, x_2) \ldots P(x_l, x_{l+1}) e^{\theta l M}}{\rho(\theta)^l},$$

and the S_m component is irreducible as well so we can find $y_1, \ldots, y_{k+1} \in S_m$ such that $P(y_i, y_{i+1}) > 0$ for $i = 1, \ldots, k$ and $y_1 = y_{k+1}$, and so

$$P_\theta(y_1, y_2) \ldots P_\theta(y_l, y_{k+1}) = \frac{P(y_1, y_2) \ldots P(y_k, y_{k+1}) e^{\theta k m}}{\rho(\theta)^k}.$$

This means that the ratio $\frac{(P_\theta(x_1,x_2)\ldots P_\theta(x_l,x_{l+1}))^{1/l}}{(P_\theta(y_1,y_2)\ldots P_\theta(y_k,y_{k+1}))^{1/k}} = \frac{P(x_1,x_2)\ldots P(x_l,x_{l+1})}{P_\theta(y_1,y_2)\ldots P_\theta(y_k,y_{k+1})} e^{\theta(M-m)}$ depends on θ. This contradicts the assumption that P_θ does not depend on θ on a neighborhood of θ_0.

Therefore, $A''(\theta)$ does not vanish on any non-empty open interval of $\mathbb{R}$, and so we conclude that $A'(\theta)$ is strictly increasing.

□

Showing part (d) of Lemma 1 requires the study of the limiting behavior of the family which we do in the following two Lemmata. The first is a simple extension of the Perron-Frobenius theory.

Lemma 4. *Let $W \in \mathbb{R}_{\geq 0}^{n\times n}$ be a non-negative matrix consisting of: a non-negative irreducible square block $A \in \mathbb{R}_{\geq 0}^{k\times k}$, and a non-negative rectangular block $B \in \mathbb{R}_{\geq 0}^{(n-k)\times k}$ such that none of the rows of B is zero, for some $k \in \{1, \ldots, n\}$, assembled together in the following way:*

$$W = \begin{bmatrix} A & 0 \\ B & 0 \end{bmatrix},$$

Then, $\rho(W) = \rho(A)$ is a simple eigenvalue of W, which we call the Perron-Frobenius eigenvalue, and is associated with unique left and right eigenvectors u_W, v_W such that u_W has its first k coordinates positive and its last $n-k$ coordinates equal to zero, v_W is positive, $\sum_{x=1}^{n} u_W(x) = 1$, and $\sum_{x=1}^{n} u_W(x)v_W(x) = 1$.

Proof. Let u_A, v_A be the unique left and right eigenvectors of A corresponding to the Perron-Frobenius eigenvalue $\rho(A)$, such that both of them are positive, $\sum_{x=1}^{k} u_A(x) = 1$ and $\sum_{x=1}^{k} u_A(x)v_A(x) = 1$. Observe that the vectors

$$u_W = \begin{bmatrix} u_A \\ 0 \end{bmatrix}, \text{ and } v_W = \begin{bmatrix} v_A \\ Bv_A/\rho(A) \end{bmatrix},$$

are left and right eigenvectors of W with associated eigenvalue $\rho(A)$, and satisfy all the conditions. In addition, $\rho(W)$ being greater than $\rho(A)$, or $\rho(W)$ not being a simple eigenvalue, or u_W, v_W not being unique would contradict the Perron-Frobenius Theorem for the non-negative irreducible matrix A. □

Now we define the matrix $\overline{P}_\infty := \lim_{\theta\to\infty} e^{-\theta M}\tilde{P}_\theta$, i.e. the matrix P where we keep the columns $y \in X_M$ intact, and we zero out all the other columns. After suitable permutation of the states Lemma 4 applies for $\overline{P}_\infty$, and so $\rho(\overline{P}_\infty)$ is a simple eigenvalue of $\overline{P}_\infty$, which is associated with unique left and right eigenvectors u_∞, v_∞ such that $u_\infty(x) > 0$ for $x \in S_M$ and $u_\infty(x) = 0$ for $x \notin S_M$, v_∞ is positive, $\sum_x u_\infty(x) = 1$ and $\sum_x u_\infty(x)v_\infty(x) = 1$. Similarly, we define $\overline{P}_{-\infty} := \lim_{\theta\to-\infty} e^{-\theta m}\tilde{P}_\theta$, with Perron-Frobenius eigenvalue $\rho(\overline{P}_{-\infty})$, which is associated with unique left and right eigenvectors $u_{-\infty}, v_{-\infty}$ such that $u_{-\infty}(x) > 0$ for $x \in S_m$ and $u_{-\infty}(x) = 0$ for $x \notin S_m$, $v_{-\infty}$ is positive, $\sum_x u_{-\infty}(x) = 1$ and $\sum_x u_{-\infty}(x)v_{-\infty}(x) = 1$.

The following Lemma characterizes the limiting stochastic matrices $P_\infty,\ P_{-\infty}$ of the exponential family, and proves part (d) of Lemma 1.

Lemma 5.

(a) $\theta M - A(\theta) \to -\log\rho(\overline{P}_\infty),\ u_\theta \to u_\infty,\ v_\theta \to v_\infty$, as $\theta \to \infty$, and so

$$\lim_{\theta\to\infty} P_\theta(x,y) = \frac{\overline{P}_\infty(x,y)v_\infty(y)}{\rho(\overline{P}_\infty)v_\infty(x)} =: P_\infty(x,y),$$

and $\pi_\theta(\phi) \to M$ as $\theta \to \infty$.

(b) $\theta m - A(\theta) \to -\log\rho(\overline{P}_{-\infty}),\ u_\theta \to u_{-\infty},\ v_\theta \to v_{-\infty}$, as $\theta \to -\infty$, and so

$$\lim_{\theta\to-\infty} P_\theta(x,y) = \frac{\overline{P}_{-\infty}(x,y)v_{-\infty}(y)}{\rho(\overline{P}_{-\infty})v_{-\infty}(x)} =: P_{-\infty}(x,y),$$

and $\pi_\theta(\phi) \to m$ as $\theta \to -\infty$.

Proof. Both parts are a straightforward application of the continuity of the function $P \mapsto (\rho(P),\ u_P,\ v_P)$, at $\overline{P}_\infty$ and $\overline{P}_{-\infty}$. The continuity of eigenvalues and eigenvectors is due to the fact that the Perron-Frobenius eigenvalue $\rho(P)$ is a simple eigenvalue and more details can be found in Chapter 3 of the book [Ort90]. □

This Lemma suggests that we can extend the domain of $A'(\theta)$ by continuity over the set of extended real numbers $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, by defining $A'(\infty) = M$ and $A'(-\infty) = m$. This way we have a one-to-one and onto correspondence of $\overline{\mathbb{R}}$ with the closed interval $[m, M]$, with the limit stochastic

matrices being P_∞ and $P_{-\infty}$, which represent degenerate Markov chains where all the transitions lead into states $y \in S_M$ when $\theta = \infty$, and into states $y \in S_m$ when $\theta = -\infty$.

We proceed by deriving some alternative representations for the relative entropy between elements of the exponential family. The following Lemma is needed in order to derive the asymptotic relative entropies.

Lemma 6.

(a) $\theta A'(\theta) - A(\theta) \to -\log \rho(\overline{P}_\infty)$, *as* $\theta \to \infty$.

(b) $\theta A'(\theta) - A(\theta) \to -\log \rho(\overline{P}_{-\infty})$, *as* $\theta \to -\infty$.

Proof. Let $M_2 = \max_{x \notin S_M} \phi(x)$. Fix $x \in S$ and $y \notin S_M$. Pick $y_M \in S_M$ such that $P(x, y_M) > 0$. Using Lemma 9 we see that there is a constant $c = c(P, \phi)$ such that

$$P_\theta(x, y) \leq c e^{-\theta(M - \phi(y))} P_\theta(x, y_M) \leq c e^{-\theta(M - M_2)}.$$

Therefore the stationary probability of any such y is at most $\pi_\theta(y) \leq c e^{-\theta(M-M_2)}$, and so

$$\pi_\theta(\phi) \geq (1 - c|S|e^{-\theta(M-M_2)})M + c|S|e^{-\theta(M-M_2)}m.$$

From this we obtain that

$$0 \leq \theta(M - \pi_\theta(\phi)) \leq c|S|\theta e^{-\theta(M-M_2)}(M - m), \text{ for any } \theta \geq 0,$$

which yields that $\theta(A'(\theta) - M) \to 0$, as $\theta \to \infty$. Part (a) now follows, since Lemma 5 suggests that $\theta M - A(\theta) \to -\log \rho(\overline{P}_\infty)$, as $\theta \to \infty$. The second limit follows by the same argument. □

Having this in our possession we state and prove alternative representations for the relative entropy.

Lemma 7.

(a) For all $\theta_1, \theta_2 \in \mathbb{R}$,

$$\begin{aligned} D(\theta_1 \parallel \theta_2) &= \theta_1 A'(\theta_1) - A(\theta_1) - (\theta_2 A'(\theta_1) - A(\theta_2)); \\ D(\infty \parallel \theta_2) &= -\log \rho(\overline{P}_\infty) - (\theta_2 M - A(\theta_2)); \\ D(-\infty \parallel \theta_2) &= -\log \rho(\overline{P}_{-\infty}) - (\theta_2 m - A(\theta_2)). \end{aligned}$$

(b) For all $\mu_1, \mu_2 \in (m, M)$,

$$\begin{aligned} D(\mu_1 \parallel \mu_2) &= A'^{-1}(\mu_1)\mu_1 - A(A'^{-1}(\mu_1)) - (A'^{-1}(\mu_2)\mu_1 - A(A'^{-1}(\mu_2))); \\ D(M \parallel \mu_2) &= -\log \rho(\overline{P}_\infty) - (A'^{-1}(\mu_2)M - A(A'^{-1}(\mu_2))); \\ D(m \parallel \mu_2) &= -\log \rho(\overline{P}_{-\infty}) - (A'^{-1}(\mu_2)m - A(A'^{-1}(\mu_2))). \end{aligned}$$

Proof. For $\theta_1, \theta_2 \in \mathbb{R}$ we have that

$$\begin{aligned} D(\theta_1 \parallel \theta_2) &= \mathbb{E}_{(X,Y) \sim \pi_{\theta_1} \odot P_{\theta_1}} \log \frac{P_{\theta_1}(X, Y)}{P_{\theta_2}(X, Y)} \\ &= A(\theta_2) - A(\theta_1) - (\theta_2 - \theta_1)A'(\theta_1) + \mathbb{E}_{(X,Y) \sim \pi_{\theta_1} \odot P_{\theta_1}} \left[\log \frac{v_{\theta_1}(Y)}{v_{\theta_1}(X)} - \log \frac{v_{\theta_2}(Y)}{v_{\theta_2}(X)} \right] \\ &= \theta_1 A'(\theta_1) - A(\theta_1) - (\theta_2 A'(\theta_1) - A(\theta_2)), \end{aligned}$$

and the third equality follows due to the fact that $\pi_{\theta_1} \odot P_{\theta_1}$ has identical marginals and so the expectation vanishes.

Now let $\theta_2 \in \mathbb{R}$. Using the continuity of relative entropy, the formula that we just established, and Lemma 6 we obtain

$$\begin{aligned} D(\infty \parallel \theta) &= \lim_{\theta_1 \to \infty} D(\theta_1 \parallel \theta_2) \\ &= \lim_{\theta_1 \to \infty} (\theta_1 A'(\theta_1) - A(\theta_1)) - \lim_{\theta_1 \to \infty} (\theta_2 A'(\theta_1) - A(\theta_2)) \\ &= -\log \rho(\overline{P}_\infty) - (\theta_2 M - A(\theta_2)). \end{aligned}$$

We argue in the same way for $D(-\infty \parallel \theta)$, and part (b) directly follows from part (a). □

As a direct consequence of these representation we obtain the following monotonicity properties of the relative entropy.

Corollary 2.

(a) For fixed $\theta_2 \in \mathbb{R}$, the function $\theta_1 \mapsto D(\theta_1 \parallel \theta_2)$ is strictly increasing in the interval $[\theta_2, \infty]$ and strictly decreasing in the interval $[-\infty, \theta_2]$.

(b) For fixed $\mu_2 \in (m, M)$, the function $\mu_1 \mapsto D(\mu_1 \parallel \mu_2)$ is strictly increasing in the interval $[\mu_2, M]$ and strictly decreasing in the interval $[m, \mu_2]$.

We close this appendix by establishing that relative entropy is the convex conjugate of the log-Perron-Frobenius eigenvalue.

Lemma 8.

$$D(\mu \parallel \pi(\phi)) = \sup_{\theta \in \mathbb{R}} \{\theta\mu - A(\theta)\} = \begin{cases} \sup_{\theta \geq 0} \{\theta\mu - A(\theta)\}, & \text{if } \mu \in [\pi(\phi), M] \\ \sup_{\theta \leq 0} \{\theta\mu - A(\theta)\}, & \text{if } \mu \in [m, \pi(\phi)]. \end{cases}$$

Proof. Fix $\mu \in (m, M)$. The function $\theta \mapsto \theta\mu - A(\theta)$ is strictly concave and its derivative vanishes at $\theta = {A'}^{-1}(\mu)$, which belong in $[0, \infty)$ when $\mu \in [\pi(\phi), M)$ and in $(-\infty, 0]$ when $\mu \in (m, \pi(\phi)]$. Therefore, using Lemma 7 we obtain

$$\sup_{\theta \in \mathbb{R}} \{\theta\mu - A(\theta)\} = {A'}^{-1}(\mu)\mu - A({A'}^{-1}(\mu)) = D(\mu \parallel \pi(\phi)).$$

Similarly when $\mu = M$ or $\mu = m$, the derivative only vanishes at ∞ and $-\infty$ respectively, and so from a combination of Lemma 5 and Lemma 7 we obtain

$$\sup_{\theta \in \mathbb{R}} \{\theta M - A(\theta)\} = \lim_{\theta \to \infty} (\theta M - A(\theta)) = D(M \parallel \pi(\phi)),$$

and

$$\sup_{\theta \in \mathbb{R}} \{\theta m - A(\theta)\} = \lim_{\theta \to -\infty} (\theta m - A(\theta)) = D(m \parallel \pi(\phi)).$$

□

B Concentration for Markov Chains

We first use continuity in order to get a uniform bound on the ratio of the entries of the right Perron-Frobenius eigenvector.

Lemma 9. *Let $\phi : S \to \mathbb{R}$ be a non-constant function. For any $P \in \mathcal{P}(\phi, S)$, there exists a constant $c = c(P, \phi) \geq 1$ such that*

$$c^{-1} \leq \sup_{\theta \in \mathbb{R}, x, y \in S} \frac{v_\theta(y)}{v_\theta(x)} \leq c.$$

If in addition P is a positive stochastic matrix then we can take $c = \max_{x,y,z} \frac{P(y,z)}{P(x,z)}$.

Proof. For any $x, y \in S$, the ratio $\frac{v_\theta(y)}{v_\theta(x)}$ is a positive real number, and due to Lemma 1 a continuous function of θ. In addition Lemma 4 and Lemma 5 suggest that its limit points $\frac{v_\infty(y)}{v_\infty(x)}$, $\frac{v_{-\infty}(y)}{v_{-\infty}(x)}$ are positive real numbers as well, hence we can take c to be $c = \sup_{\theta \in \mathbb{R}, x, y \in S} \frac{v_\theta(y)}{v_\theta(x)} \geq 1$, which is guaranteed to be finite.

In the special case that P is a positive stochastic matrix, we use the fact that v_θ is a right Perron-Frobenius eigenvector of $\tilde{P}_\theta$ in order to write

$$\frac{v_\theta(y)}{v_\theta(x)} = \frac{\sum_w \tilde{P}_\theta(y, w) v_\theta(w)}{\sum_w \tilde{P}_\theta(x, w) v_\theta(w)}, \text{ for all } x, y \in S.$$

Now using the simple inequality

$$\left(\min_z \frac{\tilde{P}_\theta(y,z)}{\tilde{P}_\theta(x,z)}\right)\tilde{P}_\theta(x,w) \le \tilde{P}_\theta(y,w) \le \left(\max_z \frac{\tilde{P}_\theta(y,z)}{\tilde{P}_\theta(x,z)}\right)\tilde{P}_\theta(x,w), \text{ for all } x,y,w \in S,$$

and observing that $\frac{\tilde{P}_\theta(y,z)}{\tilde{P}_\theta(x,z)} = \frac{P(y,z)}{P(x,z)}$ we obtain

$$\min_z \frac{P(y,z)}{P(x,z)} \le \frac{v_\theta(y)}{v_\theta(x)} \le \max_z \frac{P(y,z)}{P(x,z)}.$$

□

Next we establish a Proposition which gives us an approximation of the log-Perron-Frobenius eigenvalue using the log-moment-generating-function

$$A_n(\theta) := \frac{1}{n}\log \mathbb{E}_{(q,P)} \exp\left\{\theta(\phi(X_1) + \ldots + \phi(X_n))\right\}$$

Proposition 1. *There exists a constant $C = C(P,\phi) \ge 0$ such that*

$$|A_n(\theta) - A(\theta)| \le \frac{C}{n}, \textit{ for all } \theta \in \mathbb{R}.$$

If in addition P is a positive stochastic matrix then we can take $C = \log\left(\max_{x,y,z} \frac{P(y,z)}{P(x,z)}\right)$.

Proof. We start with the following calculation

$$\begin{aligned} e^{nA_n(\theta)} &= \sum_{x_0,x_1,\ldots,x_{n-1},x_n} q(x_0)P(x_0,x_1)e^{\theta\phi(x_1)}\cdots P(x_{n-1},x_n)e^{\theta\phi(x_n)} \\ &= \sum_{x_0,x_n} q(x_0)\tilde{P}_\theta^n(x_0,x_n). \end{aligned}$$

From this using the simple inequality

$$\frac{v_\theta(y)}{\max_x v_\theta(x)} \le 1 \le \frac{v_\theta(y)}{\min_x v_\theta(x)}, \text{ for all } y \in S,$$

together with the fact that v_θ is a right Perron-Frobenius eigenvector of $\tilde{P}_\theta$ we obtain

$$\min_{x,y} \frac{v_\theta(y)}{v_\theta(x)} e^{nA(\theta)} \le e^{nA_n(\theta)} \le \max_{x,y} \frac{v_\theta(y)}{v_\theta(x)} e^{nA(\theta)}.$$

The conclusion now follows by applying Lemma 9 □

Having this approximation we can derive Theorem 1 by a simple application of Markov's inequality combined with the convex conjugate result of Lemma 8.

Proof. (Theorem 1)
Let $\mu \in [\pi(\phi), M]$. Then for any $\theta \ge 0$ we have that

$$\begin{aligned} \mathbb{P}_{(q,P)}\left(\phi(X_1) + \ldots + \phi(X_n) \ge n\mu\right) &\le \mathbb{P}_{(q,P)}\left(e^{\theta(\phi(X_1)+\ldots+\phi(X_n))} \ge e^{\theta n\mu}\right) \\ &\le e^{-n(\theta\mu - A_n(\theta))} \\ &\le ce^{-n(\theta\mu - A(\theta))}, \end{aligned}$$

where the second inequality is Markov's inequality, and the third is the estimate from Proposition 1 with $c = c(P,\phi) \ge 1$, which we can take as $c = \max_{x,y,z} \frac{P(y,z)}{P(x,z)}$ in the special case that P is a positive stochastic matrix. The conclusion now follows by optimizing over $\theta \ge 0$ and applying Lemma 8. The case $\mu \in [m, \pi(\phi)]$ follows in the same way. □

C Lower Bound on the Sample Complexity

First we establish the change of measure formula presented in Section 5.

Lemma 10. *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*

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

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

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

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

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

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

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

□

Our next goal is to prove Lemma 2, for which we will apply a renewal argument. Let q be an initial distribution, and P be an irreducible stochastic matrix which govern the evolution of a Markov chain $X_0, X_1, \ldots, X_n, \ldots$. Using the *strong Markov property* we can derive the following standard decomposition of the Markov chain in IID blocks.

Lemma 11. *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. Those random times partition the Markov chain in a sequence $v_1, v_2, \ldots, v_k, \ldots$ of IID 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 $\mathbb{P}_{(q,P)}$-a.s. finite, which will enable us to apply the strong Markov property. In addition observe that if we let $\theta_n(\omega_0, \omega_1, \ldots) = (\omega_n, \omega_{n+1}, \ldots)$ be the shift operators on the sequence space, then $v_k = v_1 \circ \theta_{\tau_{k-1}}$. In addition 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

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

and so

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

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 [Dur10], that the stationary distribution π of the Markov chain satisfies the relation

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

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

Lemma 12.

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

Proof. Since $\pi(x) = \frac{\mathbb{E}_{(x_0,P)}\, N(x,0,\tau_1)}{\mathbb{E}_{(x_0,P)}\, \tau_1}$ for any $x_0 \in S$, it is enough to show that

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

or expanding out the definitions that

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

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

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

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

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

□

The following Lemma, which is a variant of Lemma 2.1 in [AVW87b], is the place that we use the IID block structure of the Markov chain.

Lemma 13. *Define the mean return time of the (q, P) chain*

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

Let $\mathcal{F}_n := \sigma(X_0, X_1, \ldots, X_n)$ be the observed 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 $\mathbb{E}_{(q,P)}\,\tau < \infty$. Then

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

Proof. We use the k-th return times

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

in order to decompose $N(x, y, 0, \tau_k)$ in k i.i.d. summands according to Lemma 11

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

Now let $\kappa = \inf\{k > 0 : \tau_k \ge \tau\}$, 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

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

Taking expectations in the first one we obtain

$$\mathbb{E}_{(q,P)}[\tau_\kappa - \tau] \le \mathbb{E}_{(q,P)}[\tau_\kappa - \tau_{\kappa-1}] = \mathbb{E}_{(q,P)}\,r_1 = R,$$

which also gives that

$$\mathbb{E}_{(q,P)}\,\tau_\kappa \le \mathbb{E}_{(q,P)}\,\tau + R < \infty.$$

This allows us to use Wald's identity, followed by Lemma 12, followed by Wald's identity again, in order to get

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

Therefore

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

□

Now Lemma 2 follows directly from Lemma 13.

The last part of Appendix C involves the proof of Theorem 2.

Proof. (Theorem 2) 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 $(\boldsymbol{q}, \boldsymbol{\lambda}) \in \mathcal{I} \times \text{Alt}(\boldsymbol{\mu})$.

The data processing inequality (see the book of [CT06] 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 $\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\lambda})}\,|_{\mathcal{F}_{\tau_\delta}}$.

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

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

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

hence

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

Combining this with Lemma 2 we get that

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

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

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

and now we follow the technique of [GK16] which combines multiple alternative models $\boldsymbol{\lambda}$

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

□

D Upper Bound on the Sample Complexity: the (α, δ)-Track-and-Stop Strategy

In order to obtain the uniform over time deviation Lemma 3, we need a uniform bound on the constants $c(P_\theta)$ of the exponential family as θ ranges over $\mathbb{R}$. We derive this in the following Lemma.

Lemma 14.

$$\sup_{\theta\in\mathbb{R}} c(P_\theta) \le c(P)^2.$$

Proof. Recall that

$$c(P_{\theta_2}) = \sup_{\theta_1\in\mathbb{R},x,y\in S} \frac{v_{\widetilde{(P_{\theta_2})}_{\theta_1}}(y)}{v_{\widetilde{(P_{\theta_2})}_{\theta_1}}(x)}.$$

We claim that

$$\frac{v_{\widetilde{(P_{\theta_2})}_{\theta_1}}(y)}{v_{\widetilde{(P_{\theta_2})}_{\theta_1}}(x)} = \frac{v_{\tilde{P}_{\theta_1+\theta_2}}(y) v_{\tilde{P}_{\theta_2}}(x)}{v_{\tilde{P}_{\theta_1+\theta_2}}(x) v_{\tilde{P}_{\theta_2}}(y)}.$$

To see this we just need to verify that

$$v_{\tilde{P}_{\theta_2}}(x) v_{\widetilde{(P_{\theta_2})}_{\theta_1}}(x),\ x \in S,$$

is a right eigenvector of $\tilde{P}_{\theta_1+\theta_2}$, with associated eigenvalue $\rho(\tilde{P}_{\theta_2})\rho\left(\widetilde{(P_{\theta_2})}_{\theta_1}\right)$, which from the Perron-Frobenuous theory has to be the Perron-Frobenious eigenvalue since the associated eigenvector has positive entries. The verification is straight forward

$$\begin{aligned} \sum_y \tilde{P}_{\theta_1+\theta_2}(x,y) v_{\tilde{P}_{\theta_2}}(y) v_{\widetilde{(P_{\theta_2})}_{\theta_1}}(y) &= \rho(\tilde{P}_{\theta_2}) v_{\tilde{P}_{\theta_2}}(x) \sum_y \widetilde{(P_{\theta_2})}_{\theta_1}(x,y) v_{\widetilde{(P_{\theta_2})}_{\theta_1}}(y) \\ &= \rho(\tilde{P}_{\theta_2})\rho\left(\widetilde{(P_{\theta_2})}_{\theta_1}\right) v_{\tilde{P}_{\theta_2}}(x) v_{\widetilde{(P_{\theta_2})}_{\theta_1}}(x), \text{ for all } x \in S. \end{aligned}$$

From this we see that

$$\sup_{\theta_1,\theta_2\in\mathbb{R},x,y\in S} \frac{v_{\widetilde{(P_{\theta_2})}_{\theta_1}}(y)}{v_{\widetilde{(P_{\theta_2})}_{\theta_1}}(x)} \le \left(\sup_{\theta_1,\theta_2\in\mathbb{R},x,y\in S} \frac{v_{\tilde{P}_{\theta_1+\theta_2}}(y)}{v_{\tilde{P}_{\theta_1+\theta_2}}(x)}\right)\left(\sup_{\theta_2\in\mathbb{R},x,y\in S} \frac{v_{\tilde{P}_{\theta_2}}(x)}{v_{\tilde{P}_{\theta_2}}(y)}\right) = c(P)^2.$$

□

The proof of Lemma 3 uses the concentration Theorem 1, combined with the monotonicty of the relative entropy that we say in Corollary 2, and the uniform bound on the constants of the family Lemma 14.

Proof. (Theorem 1)
We first note the following inclusion of events

$$\bigcup_{t=1}^{\infty}\bigcup_{n=1}^{t}\left\{N_a(t)D\left(\hat{\mu}_a(N_a(t))\parallel\mu_a\right)\geq\beta_{\alpha,\delta}(t)/2,\ N_a(t)=n\right\}\subseteq\bigcup_{t=1}^{\infty}\bigcup_{n=1}^{t}\left\{nD\left(\hat{\mu}_a(n)\parallel\mu_a\right)\geq\beta_{\alpha,\delta}(t)/2\right\}$$
$$=\bigcup_{t=1}^{\infty}\left\{tD\left(\hat{\mu}_a(t)\parallel\mu_a\right)\geq\beta_{\alpha,\delta}(t)/2\right\},$$

where the last equality follows because we have by the monotonicity of $t\mapsto\beta_{\alpha,\delta}(t)/2$ that for each $n\in\mathbb{Z}_{>0}$

$$\left\{nD\left(\hat{\mu}_a(n)\parallel\mu_a\right)\geq\beta_{\alpha,\delta}(t)/2\right\}\subseteq\left\{nD\left(\hat{\mu}_a(n)\parallel\mu_a\right)\geq\beta_{\alpha,\delta}(n)/2\right\},\ \text{for all } t=n,n+1,\ldots$$

This together with a union bound gives

$$\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\left(\exists t\in\mathbb{Z}_{>0}:N_a(t)D\left(\hat{\mu}_a(N_a(t))\parallel\mu_a\right)\geq\beta_{\alpha,\delta}(t)/2\right)\leq\mathbb{P}_{(q_a,P_{\mu_a})}\left(\exists t\in\mathbb{Z}_{>0}:tD\left(\hat{\mu}_a(t)\parallel\mu_a\right)\geq\beta_{\alpha,\delta}(t)/2\right)$$
$$\leq\sum_{t=1}^{\infty}\mathbb{P}_{(q_a,P_{\mu_a})}\left(D\left(\hat{\mu}_a(t)\parallel\mu_a\right)\geq\frac{\beta_{\alpha,\delta}(t)}{2t}\right).$$

Each summand $\mathbb{P}_{(q_a,P_{\mu_a})}\left(D\left(\hat{\mu}_a(t)\parallel\mu_a\right)\geq\frac{\beta_{\alpha,\delta}(t)}{2t}\right)$ is less or equal than

$$\mathbb{P}_{(q_a,P_{\mu_a})}\left(D\left(\hat{\mu}_a(t)\parallel\mu_a\right)\geq\frac{\beta_{\alpha,\delta}(t)}{2t},\ \hat{\mu}_a(t)\geq\mu_a\right)+\mathbb{P}_{(q_a,P_{\mu_a})}\left(D\left(\hat{\mu}_a(t)\parallel\mu_a\right)\geq\frac{\beta_{\alpha,\delta}(t)}{2t},\ \hat{\mu}_a(t)\leq\mu_a\right).$$

We focus on the first term and we let $\mu_{a,t}$ be the unique (due to Corollary 2) solution (if no solution exists then the probability is already zero) of the equations

$$D\left(\mu_{a,t}\parallel\mu_a\right)=\frac{\beta_{\alpha,\delta}(t)}{2t},\quad\text{and}\quad\mu_a\leq\mu_{a,t}\leq M,$$

where $M:=\max_{x\in S}\ x$ and $m:=\min_{x\in S}\ x$. Then a combination of Corollary 2 and Theorem 1 gives us

$$\mathbb{P}_{(q_a,P_{\mu_a})}\left(D\left(\hat{\mu}_a(t)\parallel\mu_a\right)\geq\frac{\beta_{\alpha,\delta}(t)}{2t},\ \hat{\mu}_a(t)\geq\mu_a\right)=\mathbb{P}_{(q_a,P_{\mu_a})}\left(\hat{\mu}_a(t)\geq\mu_{a,t}\right)\leq\frac{\delta}{C}\frac{1}{t^{\alpha}}c(P_{\mu_a}).$$

We further upper bound the constant $c(P_{\mu_a})$ by $c(P)^2$ using Lemma 14, in order to obtain a uniform upper bound for any Markovian arm coming from the family.

Furthermore, in the special case that P is a positive stochastic matrix, we can take $c(P_{\mu_a})=\max_{x,y,z}\frac{P_{\mu_a}(y,z)}{P_{\mu_a}(x,z)}$, for which we can give an explicit uniform upper bound using Lemma 9.

$$\max_{x,y,z}\frac{P_{\mu_a}(y,z)}{P_{\mu_a}(x,z)}\leq\left(\max_{x,y,z}\frac{P(y,z)}{P(x,z)}\right)\left(\max_{x,y}\frac{v_{\mu_a}(x)}{v_{\mu_a}(y)}\right)\leq\left(\max_{x,y,z}\frac{P(y,z)}{P(x,z)}\right)^2.$$

Similarly we get the same upper bound for the second term, and then the conclusion follows by summing up over all t and using the simple integral based estimate

$$\sum_{t=1}^{\infty}\frac{1}{t^{\alpha}}\leq\frac{\alpha}{1-\alpha}.$$

□

Embarking on the proof of the fact that the (α,δ)-Track-and-Stop strategy is δ-PC we first show that the error probability is at most δ no matter the bandit model.

Proposition 2. *Let $(\boldsymbol{q}, \boldsymbol{\mu}) \in \mathcal{I} \times \mathcal{T}$, $\delta \in (0,1)$, and $\alpha > 1$. For any sampling rule, if we use the (α, δ)-Chernoff's stopping rule and the best sample mean decision rule we are guaranteed that*

$$\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}(\tau_{\alpha,\delta} < \infty, \hat{a}_{\tau_{\alpha,\delta}} \neq a^*(\boldsymbol{\mu})) \leq \delta.$$

Proof. The following lemma which is easy to check, and its proof is omitted, will be useful in our proof of Proposition 2.

Lemma 15. *The generalized Jensen-Shannon divergence*

$$I_a(\mu, \lambda) := aD\left(\mu \parallel a\mu + (1-a)\lambda\right) + (1-a)D\left(\lambda \parallel a\mu + (1-a)\lambda\right), \text{ for } a \in [0,1]$$

satisfies the following variational characterization

$$I_a(\mu, \lambda) = \inf_{\mu' < \lambda'} \left\{aD\left(\mu \parallel \mu'\right) + (1-a)D\left(\lambda \parallel \lambda'\right)\right\}.$$

If $\tau_{\alpha,\delta} < \infty$ and $\hat{a}_{\tau_{\alpha,\delta}} \neq a^*(\boldsymbol{\mu})$, then there $\exists t \in \mathbb{Z}_{>0}$ and there $\exists a \neq a^*(\boldsymbol{\mu})$ such that $Z_{a,a^*(\boldsymbol{\mu})}(t) > \beta_{\alpha,\delta}(t)$. In this case we also have

$$\begin{aligned}
\beta_{\alpha,\delta}(t) &< Z_{a,a^*(\boldsymbol{\mu})}(t) \\
&= N_a(t) D\left(\hat{\mu}_a(N_a(t)) \,\middle\|\, \hat{\mu}_{a,a^*(\boldsymbol{\mu})}(N_a(t), N_{a^*(\boldsymbol{\mu})}(t))\right) + \\
&\quad N_{a^*(\boldsymbol{\mu})}(t) D\left(\hat{\mu}_{a^*(\boldsymbol{\mu})}(N_{a^*(\boldsymbol{\mu})}(t)) \,\middle\|\, \hat{\mu}_{a,a^*(\boldsymbol{\mu})}(N_a(t), N_{a^*(\boldsymbol{\mu})}(t))\right) \\
&= (N_a(t) + N_{a^*(\boldsymbol{\mu})}(t)) I_{\frac{N_a(t)}{N_a(t)+N_{a^*(\boldsymbol{\mu})}(t)}}(\hat{\mu}_a(N_a(t)), \hat{\mu}_{a^*(\boldsymbol{\mu})}(N_{a^*(\boldsymbol{\mu})}(t))) \\
&= \inf_{\mu'_a < \mu''_a} \left\{N_a(t) D\left(\hat{\mu}_a(N_a(t)) \parallel \mu'_a\right) + N_{a^*(\boldsymbol{\mu})}(t) D\left(\hat{\mu}_{a^*(\boldsymbol{\mu})}(N_{a^*(\boldsymbol{\mu})}(t)) \,\middle\|\, \mu''_a\right)\right\} \\
&\leq N_a(t) D\left(\hat{\mu}_a(N_a(t)) \parallel \mu_a\right) + N_{a^*(\boldsymbol{\mu})}(t) D\left(\hat{\mu}_{a^*(\boldsymbol{\mu})}(N_{a^*(\boldsymbol{\mu})}(t)) \,\middle\|\, \mu_{a^*(\boldsymbol{\mu})}\right),
\end{aligned}$$

where the third equality follows from the variational formula for the generalized Jensen-Shannon divergence given in Lemma 15, and the last inequality follows from the fact that $\mu_a < \mu_{a^*(\boldsymbol{\mu})}$.

This in turn implies that

$$\beta_{\alpha,\delta}(t)/2 < N_a(t) D\left(\hat{\mu}_a(N_a(t)) \parallel \mu_a\right), \quad \text{or} \quad \beta_{\alpha,\delta}(t)/2 < N_{a^*(\boldsymbol{\mu})}(t) D\left(\hat{\mu}_{a^*(\boldsymbol{\mu})}(N_{a^*(\boldsymbol{\mu})}(t)) \,\middle\|\, \mu_{a^*(\boldsymbol{\mu})}\right).$$

Therefore by union bounding over the K arms we obtain

$$\mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}(\tau_\delta < \infty, \hat{a}_{\tau_\delta} \neq a^*(\boldsymbol{\mu})) \leq \sum_{a=1}^{K} \mathbb{P}_{(\boldsymbol{q},\boldsymbol{\mu})}\left(\exists t \in \mathbb{Z}_{>0} : N_a(t) D\left(\hat{\mu}_a(N_a(t)) \parallel \mu_a\right) \geq \beta_{\alpha,\delta}(t)/2\right).$$

The conclusion now follows by applying Lemma 3. □

Proof. (Theorem 3)
Following the proof of Proposition 13 in [GK16], and observing that in their proof they show that $\tau_{\alpha,\delta}$ is essentially bounded we obtain that

$$\mathbb{E}_{(\boldsymbol{q},\boldsymbol{\mu})}[\tau_{\alpha,\delta}] < \infty.$$

This combined with Proposition 2 establishes that the (α, δ)-Track-and-Stop strategy is δ-PC. □

Proof. (Theorem 4)
Finally for the proof the sample complexity of the (α, δ)-Track-and-Stop strategy in Theorem 4 we follow the proof of Theorem 14 in [GK16], where we substitute the usage of the law of large numbers with the law of large numbers for Markov chains, and in order to establish their Lemma 19 we use our concentration bound in Theorem 1. □