Title
Concentration and Sequential Decision Making in Markovian Environments

Author
Moulos, Vrettos

Publication Date
2020

Peer reviewed|Thesis/dissertation

Concentration and Sequential Decision Making in Markovian Environments

by

Vrettos Moulos

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

and the Designated Emphasis

in

Communication, Computation, and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Satish Rao, Chair
Professor Peter Bartlett
Associate Professor Aditya Guntuboyina

Fall 2020

Concentration and Sequential Decision Making in Markovian Environments

Abstract

Concentration and Sequential Decision Making in Markovian Environments

by

Vrettos Moulos

Doctor of Philosophy in Computer Science

and the Designated Emphasis in

Communication, Computation, and Statistics

University of California, Berkeley

Professor Satish Rao, Chair

In this dissertation we study concentration properties of Markov chains, and sequential decision making problems which involve stochastic modeling with Markov chains.

We start by developing a simple yet powerful Hoeffding inequality for Markovian sums under only the irreducibility assumption. To illustrate its usefulness we provide two applications in multi-armed bandit problems. The first is about identifying an approximately best Markovian arm, while the second is concerned with regret minimization in the context of Markovian bandits, generalizing two well known algorithms from the i.i.d. case.

We proceed with the study of the concentration properties of a Lipschitz function applied to a Markov chain, which form a generalization of Hoeffding's inequality. In particular we investigate a transportation problem that arises naturally when the martingale method is applied. The so called bicausal optimal transport problem for Markov chains, is an optimal transport formulation suitable for stochastic processes which takes into consideration the accumulation of information as time evolves. Our analysis is based on a relation between the transport problem and the theory of Markov decision processes. This way we are able to derive necessary and sufficient conditions for optimality in the transport problem, as well as an iterative algorithm, namely the value iteration, for the calculation of the transportation cost. Additionally, we draw the connection with the classic theory on couplings for Markov chains, and in particular with the notion of faithful couplings.

Next we focus on a finite-sample analysis of large deviation results for Markov chains. First we study the exponential family of stochastic matrices, which serve as a change of measure, and we develop conditions under which the asymptotic Perron-Frobenius eigenvector stays

strictly positive. This leads to a Chernoff bound which attains a constant prefactor and an exponential decay with the optimal large deviations rate. Moreover, a finite-sample version of the law of the iterated logarithm is derived, and a uniform multiplicative ergodic theorem for the exponential family of tilted transition probability matrices is established.

On the applications side, 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 generalizing those of the i.i.d. setting, and we analyze the Track-and-Stop strategy, proving that asymptotically it is at most a factor of four apart from the lower bound.

We conclude with with an extension of the classic stochastic multi-armed bandit problem which involves multiple plays and Markovian rewards in the rested bandits setting. In order to tackle this problem we consider an adaptive allocation rule which at each stage combines the information from the sample means of all the arms, with the Kullback-Leibler upper confidence bound of a single arm which is selected in round-robin way. For rewards generated from a one-parameter exponential family of Markov chains, we provide a finite-time upper bound for the regret incurred from this adaptive allocation rule, which reveals the logarithmic dependence of the regret on the time horizon, and which is asymptotically optimal. As a byproduct of our analysis we also establish asymptotically optimal, finite-time guarantees for the case of multiple plays, and i.i.d. rewards drawn from a one-parameter exponential family of probability densities. Finally, we provide simulation results that illustrate that calculating Kullback-Leibler upper confidence bounds in a round-robin way, is significantly more efficient than calculating them for every arm at each round, and that the expected regrets of those two approaches behave similarly.

Dedicated to my father Vangelis, my mother Nomiki, and my sister Sevasti.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to extend my warmest thanks to my advisor Satish Rao. I met Satish during a time that I had lost my orientation in graduate school. With his endless help, support, motivation, and encouragement, I was able to get back on track and complete my thesis. Satish is a true academic, devoted to his work on teaching, researching, and influencing the next generations. He gave me the privilege of academic freedom, and most importantly he believed in me when I needed it.

I'm grateful to Venkat Anantharam who has been a source of inspiration and knowledge for me. I heard about Markov chains for the first time from a class that he taught, and then I ended up writing a thesis about them. In addition to teaching me great classes, he was a great mentor all the times I collaborated with him as a teaching assistant and as a researcher. He helped me to think clearly and structured, and to develop mathematical intuitions.

I'm thankful to Jim Pitman for introducing me to probability theory, for being so passionate and motivational, and for being approachable and open to guide me through the various topics that I was discussing with him.

I would like to thank my dissertation committee, Peter Bartlett, and Aditya Guntuboyina. Peter was very approachable and open to talk about my research and offer advice, as well as connect me with people in the field. Aditya inspired me with his clarity and deep knowledge of statistics. I'm thankful for the time he devoted to meet with me, and the research ideas he proposed. Part of my research was initialized during our meetings, and I'm also thankful to him for helping me with my next steps after graduate school.

I would like to thank Nikos Papaspyrou, and Stathis Zachos for helping me and encouraging me during the phase at which I was applying to graduate schools. I would like to thank Sanjit Seshia, and Stravros Tripakis for giving me the chance to study at the UC Berkeley, and for supporting me during difficult times. I would like to thank Michael Mahoney for including me in his group meetings, and for proposing research directions. I would like to thank Tomas Singliar for our collaboration during a summer internship at Amazon, where I was introduced to the world of industry for the first time.

I'm thankful to the amazing EECS academic staff: Susanne Kauer, Heather Levien, Jean Nguyen, Shirley Salanio, and Audrey Sillers, who do a tremendous work on supporting the EECS students.

Special thanks goes to all the professors with whom I collaborated as a teaching assistant: Venakt Anantharam, Sanjam Garg, Antonio Montalban, Prasad Raghavendra, Kannan Ramchandran, Satish Rao, Jonathan Shewchuk, and Theodore Slaman. I gained a lot of excellent teaching practises, as well as communication and organization skills from them. In addition to that I want to thank the numerous fellow teaching assistants with who I have collaborated, as well as the excellent Berkeley students from who I learnt a lot while teaching them. I also need to thank my co-instructors for the summer offering of CS 70: Hongling Lu, Allen Tang, and Sinho Chewi. In addition to a co-instructor, Sinho has been a friend, a collaborator, and one of the key people who motivated me to study probability. In Berkeley I took numerous classes, in probability, statistics, optimization, and mathematical

analysis, learning everything essentially from scratch, and I want to thank all the professors who taught me those amazing disciplines. I wish I can continue to learn more mathematics even after graduate school. For me the cycle of science consists of: learning, teaching, and researching.

At a personal level I want to thank all the people that I came close to in Berkeley: Christos Adamopoulos, Ligia Diana Amorim, Sinho Chewi, Kimon Fountoulakis, Perla Gamez, Kostas Giannopoulos, John Hector Haloulos, Fotis Iliopoulos, Stavros Karagianopoulos, Iraklis Koutrouvelis, Stergios Koutrouvelis, Nikos Liakakos, Vasilis Oikonomou, Thanos Panagopoulos, George Patsakis, Andrew Suh, Andy Theocharous, Dimitra Tsiaousi, Evangelos Vrettos, Cindy Wang, Natalie Yu, and Panagiotis Zarkos. Their friendship helped me a lot to cope with graduate school. I would like to thank my friends from Greece: Orestis Dimou, Nikitas Georgakis, Andreas Kaskaridis, Dimitris Konomis, Apostolis Kristallidis, Giannis Mauroeidis, Alexandros Ninos, Michalis Noltsis, Paulos Samaras, Georgios Sofronas, and Orestis Tzortzopoulos, for all their long distance support.

Finally I need to thank my extended family for their love, support, and encouragement. I want to thank my aunts, and uncles: Petros Maragos, Angeliki Maragou, Panagiotis Moulos, Giannis Moulos, as well as their families, my cousins, and especially thank my older cousin Vrettos Moulos for flying with me all the way from Greece to help me settle in Berkeley. Additionally I want to thank my grandparents: Vrettos Moulos, Titsa Moulou, Angelos Maragos, and Sevasti Maragou, for playing an important role to who I am today. To conclude I need to thank and express my love and gratitude to my father Vangelis, my mother Nomiki, and my sister Sevasti, to who this thesis is dedicated.

Chapter 1

Introduction

The goal of this dissertation is to develop tools that help the analysis of Markov chains in the finite sample regime, those usually take the form of a concentration inequality, and then utilize them in order to study sequential decision making problems in Markovian environments. In this introduction, we first briefly define what is a Markov chain, and we present three limit theorems that drive our finite sample developments. We then introduce the multi-armed bandits problem, and we describe the identification and the regret minimization objectives with which we will be concerned. The introduction is by no means a complete treatment of the theory of Markov chains, and multi-armed bandits - we merely present some results that will motivate the chapters that follow. We conclude this chapter by giving a general outline of the dissertation. Relevant literature can be found in the corresponding chapters. The target audience of the presentation has a background in probability and statistics at a graduate level.

1.1 Markov chains, limit theorems, and their finite-sample counterparts

The main object of study in this dissertation is homogeneous Markov chains on a finite state space S. That is a sequence of random variables $\{X_n\}_{n\in\mathbb{Z}_{\geq 0}}$ taking values on S, which is driven by an initial distribution, q, on S, and a transition probability matrix $P : S \times S \to [0, 1]$, so that the finite dimensional distributions are given, for every $n \in \mathbb{Z}_{\geq 0}$, by

$$\mathbb{P}_q(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = q(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n). \tag{1.1}$$

Through Kolmogorov's extension theorem those finite dimensional distributions will define a stochastic process. (1.1) implies the Markov property which roughly speaking states that the past and the future are conditionally independent given the present. More formally, for all $n \in \mathbb{Z}_{\geq 0}$, we have that

$$\mathbb{P}_q(X_{n+1} = x_{n+1} \mid X_n = x_n, \ldots, X_0 = x_0) = \mathbb{P}_q(X_{n+1} = x_{n+1} \mid X_n = x_n) = P(x_n, x_{n+1}),$$

for all $x_0, x_1, \ldots, x_{n+1} \in S$.

The Markov property extends to the strong Markov property which essentially allows us to substitute the fixed deterministic n, with a random stopping time T with $\mathbb{P}_q(T < \infty) = 1$ so that

$$\mathbb{P}_q(X_{T+1} = y \mid X_T = x, E) = \mathbb{P}_q(X_{T+1} = y \mid X_T = x) = P(x, y),$$

for all $x, y \in S$ and $E \in \mathcal{F}(X_0, \ldots, X_{T-1})$.

Using the strong Markov property we can decompose a Markov chain in a sequence of i.i.d. blocks plus some residuals, which allows us to port limit theorems from the i.i.d. case to the Markovian case. In particular we define recursively the k-th return time to the initial state as

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

Those return times partition the Markov chain in a sequence $\{v_k\}_{k \in \mathbb{Z}_{>0}}$ of i.i.d. random blocks given by

$$v_k = \left(X_{\tau_{k-1}}, \ldots, X_{\tau_k - 1}\right), \text{ for } k \geq 1.$$

Let P be an irreducible transition probability matrix. Then the Markov chain possess a unique stationary distribution π. Let $f : S \to \mathbb{R}$ be a real valued function on the state space, and define the partial sums

$$S_n = f(X_0) + f(X_1) + \cdots + f(X_n).$$

Let $\mu_0 = \sum_{x \in S} f(x)\pi(x)$ be the stationary mean of the chain. We will be concerned with the convergence properties of the centralized sums, $S_n - n\mu_0$, under various scalings.

Using the decomposition of a Markov chain in i.i.d. blocks, one can first establish a law of large numbers for Markov chains.

Theorem 1 (Law of Large Numbers, Theorem I.15.2 in [19])**.**

$$\frac{S_n - n\mu_0}{n} \overset{a.s.}{\to} 0, \text{ as } n \to \infty.$$

In this dissertation we utilize the theory of large deviations in order to give a Chernoff bound for the probability of a Markovian sample mean deviating from the stationary mean. Our bound captures an exponential decay with a tight rate as this is dictated by the asymptotic theory of large deviations, and prefactor which is constant with amount of deviation. This bound serves as finite-sample product of Theorem 1 and is presented in Theorem 10.

The next important limit theorem that we will be concerned with is the central limit theorem for Markov chains. In the Markovian case the variance that appears in the central limit theorem is the stationary variance plus a sum of decaying covariances, and is given by the following formula

$$\sigma_0^2 = \mathrm{var}_\pi(f(X_1)) + \sum_{k=1}^{\infty} \mathrm{cov}_\pi(f(X_1), f(X_{k+1})).$$

Theorem 2 (Central Limit Theorem, Theorem I.16.1 in [19])**.**

$$\frac{S_n - n\mu_0}{\sqrt{n}} \xrightarrow{d} N(0, \sigma_0^2).$$

In Theorem 4 we develop a Hoeffding inequality for Markov chains, while requiring only irreducibility, in contrast to other works which require aperiodicity as well. Our Hoeffding inequality describes the Gaussian tails that the centralized Markovian sums exhibit, and so it can be viewed as a finite-sample aspect of the central limit theorem for Markov chains. In particular it gives a variance proxy through the hitting time quantities of the chain.

A further extension of Hoeffding's inequality, and yet another manifestation of the central limit theorem, is the bounded differences inequality. Here we study an optimal transport problem which naturally arises when one applies the martingale method in order to derive the bounded differences inequality for Markov chains. We relate this optimal transport problem, with Markov decision processes, and using them we describe necessary and sufficient conditions for optimality as well as a fixed point iteration to solve it.

The central limit theorem, Theorem 2, implies that

$$\limsup_{n\to\infty} \frac{S_n - n\mu_0}{\sqrt{n}} \overset{a.s.}{=} \infty,$$

and if we compare this with law of large numbers, Theorem 1, it is natural to question what is the scaling under which the centralized sums lie almost surely in a compact interval. The answer to this question is given from the law of the iterated logarithm.

Theorem 3 (Law of the Iterated Logarithm, Theorem I.16.5 in [19])**.**

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n \log\log n}} \overset{a.s.}{=} \sqrt{2\sigma_0^2}.$$

In Theorem 11 we develop a finite sample deviation inequality in the spirit of Theorem 3. We use techniques from the theory of large deviations, an exponential martingale, and a peeling argument dividing time in exponential epochs which is typical in law of iterated logarithm type of proofs.

1.2 Multi armed bandits

From the application perspective in this dissertation we use Markov chains to model a sequential decision making problem in unknown random environments. More precisely, we consider the setting known under the conventional name of stochastic multi-armed bandit, in reference to the gambling game. In the multi-armed bandit model, the emphasis is put on focusing as quickly as possible on the best available options rather than on estimating precisely the efficiency of each option. These options are referred to as arms, and each of

them is associated with a stochastic process, with unknown statistics for the player. Arms are indexed by $a \in \{1, \dots, K\}$ and associated with a stochastic process paremetrized by θ_a and governed by the probability law $\mathbb{P}_{\theta_a}$. In our context the stochastic process are Markov chains, parametrized by their stationary mean, which is unknown to the player. At each time-step the player selects an arm and the corresponding Markov chain evolves by one time step, and she observes this evolution through a reward function, while the Markov chains for the rest of the arms stay frozen, i.e. we consider the rested bandits setting.

We study first the problem where the goal of the player is to identify, with some fixed confidence δ, the Markov chain with the largest stationary mean using as few samples as possible. Our contribution is a lower bound on the sampling complexity, the derivation of which involves the i.i.d. block structures that are inherent in Markov chains, as well as a sampling algorithm which together with the lower bound characterize the sampling complexity of the problem in the high confidence regime that $\delta \to 0$.

An alternative objective is the one of regret minimization. There a time horizon T is prescribed and the goal of the player is to select arms in such a way so as to make the cumulative reward over the whole time horizon T as large as possible. For this task the player is faced with an exploitation versus exploration dilemma. At each round she needs to decide whether she is going to exploit the best arm according to the information that we have gathered so far, or she is going to explore some other arms which do not seem to be so rewarding, just in case that the rewards she have observed so far deviate significantly from the expected rewards. The answer to this dilemma is usually coming by calculating indices for the arms and ranking them according to those indices, which should incorporate both information on how good an arm seems to be as well as on how many times it has been played so far. Here we take an alternative approach were instead of calculating the indices for all the arms at each round, we just calculate the index for a single arm in a round-robin way. We provide a finite time analysis of our algorithm which matches the known lower bound, as well as simulation results which illustrate that this round-robin scheme is computationally much more efficient than other well known algorithms. A practical example of this Markovian modeling involves a casino with slot-machines whose reward distribution is changing based on the reward just observed. The casino in attempt to make more money is allowed in this framework to change the reward distribution of an arm that just produced a high reward to a stingy one.

1.3 Organization

- In Chapter 2 we present Hoeffding's inequality for Markov chains, which reveals its Gaussian tails. Additionally, the bounded differences inequality for Markov chains gives rise to an optimal transport problem, which is related to coupling, and solved via the theory of Markov decision processes.

- In Chapter 3 we take a large deviations perspective on Markov chains, we study expo-

nential families of stochastic matrices, and using them we develop a Chernoff bound, as well as a maximal deviation inequality related to the law of the iterated logarithm.

- In Chapter 4 we investigate the problem of best Markovian arm identification with fixed confidence, where we develop a lower bound, as well as an algorithm, which combined characterize the sampling complexity of the problem.
- In Chapter 5 we study the problem of regret minimization for Markovian bandits with multiple plays. We give a finite-time analysis for a round-robin KL-UCB algorithm, which is asymptotically optimal, and much more efficient than other KL-UCB type of algorithms.

1.4 Bibliographic notes

The results in this dissertation ad based on the papers [73, 74, 70, 72, 71]. In particular the results in [74] are based on collaboration with Venkat Anantharam.

Chapter 2

Concentration inequalities and transportation problems for Markov chains

2.1 Introduction

Let $\{X_k\}_{k\in\mathbb{Z}_{\geq 0}}$ be a Markov chain on a finite state space S, with initial distribution q, and irreducible transition probability matrix P, governed by the probability law $\mathbb{P}_q$. Let π be its stationary distribution, and $f : S \to [a, b]$ be a real-valued function on the state space. Then the strong law of large numbers for Markov chains asserts that,

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \overset{\mathbb{P}_q - a.s.}{\to} \mathbb{E}_\pi[f(X_1)], \text{ as } n \to \infty.$$

Moreover, the central limit theorem for Markov chains provides a rate for this convergence,

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbb{E}_\pi[f(X_1)]\right) \overset{d}{\to} N(0, \sigma^2), \text{ as } n \to \infty,$$

where $\sigma^2 = \lim_{n\to\infty} \frac{1}{n} \operatorname{var}_q\left(\sum_{k=1}^{n} f(X_k)\right)$ is the limiting variance.

Those asymptotic results are insufficient in many applications which require finite-sample estimates. One of the most important such application is the convergence of Markov chain Monte Carlo (MCMC) approximation techniques [67], where a finite-sample estimate is needed to bound the approximation error. Further applications include theoretical computer science and the approximation of the permanent [47], as well as statistical learning theory and multi-armed bandit problems [73].

Motivated by this discussion we provide in Section 2.2 a finite-sample Hoeffding inequality for finite Markov chains. In the special case that the random variables $\{X_k\}_{k\in\mathbb{Z}_{\geq 0}}$ are

independent and identically distributed according to π, Hoeffding's classic inequality [44] states that,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}[f(X_1)]\right)\right|\geq t\right)\leq 2\exp\left\{-\frac{t^2}{2\nu^2}\right\},$$

where $\nu^2=\frac{1}{4}n(b-a)^2$. In Theorem 4 we develop a version of Hoeffding's inequality for finite state Markov chains. Our bound is very simple and easily computable, since it is based on martingale techniques and it only involves hitting times of Markov chains which are very well studied for many types of Markov chains [2]. It is worth mentioning that our bound is based solely on irreducibility, and it does not make any extra assumptions like aperiodicity or reversibility which prior works require.

There is a rich literature on finite-sample bounds for Markov chains. One of the earliest works [23] uses counting and a generalization of the method of types, in order to derive a Chernoff bound for Markov chains which are irreducible and aperiodic. An alternative approach [92, 74], uses the theory of large deviations to derive sharper Chernoff bounds. When reversibility is assumed, the transition probability matrix is symmetric with respect to the space $L^2(\pi)$, which enables the use of matrix perturbation theory. This idea leads to Hoeffding inequalities that involve the spectral gap of the Markov chain and was initiated in [39]. Refinements of this bound were given in a series of works [26, 48, 60, 59, 68]. In [77, 81, 34] a generalized spectral gap is introduced in order to obtain bounds even for a certain class of irreversible Markov chains as long as they posses a strictly positive generalized spectral gap. Information-theoretic ideas are used in [54] in order to derive a Hoeffding inequality for Markov chains with general state spaces that satisfy Doeblin's minorization condition, which in the case of a finite state space can be written as,

$$\exists m\in\mathbb{Z}_{>0}\ \exists y\in S\ \forall x\in S:\ P^m(x,y)>0. \tag{2.1}$$

Of course there are irreducible transition probability matrices P for which (2.1) fails, but if we further assume aperiodicity, then (2.1) is satisfied. Our approach uses Doob's martingale combined with Azuma's inequality, and is probably closest related to the work in [40], where a bound for Markov chains with general state spaces is established using Dynkin's martingale. But the result in [40] heavily relies on the Markov chains satisfying Doeblin's condition (2.1). A regeneration approach for uniformly ergodic Markov chains, where one splits the Markov chain in i.i.d. blocks, and reduces the problem to the concentration of an i.i.d. process, can be found in [28].

Another line of research is related to the concentration of a function of n random variables around its mean, under Markovian or other dependent structures. This was pioneered by the works of Marton [63, 64, 66] who used the transportation-information method, and further developed using the martingale method and coupling in [83, 65, 17, 53, 77, 52]. In Section 2.3 we study the optimal transport problem arising in the study of concentration of measure for Markov chains, from a causal/adaptive point of view.

We give some applications of our concentration results in Section 2.5, where we study two Markovian multi-armed bandit problems. The stochastic multi-armed bandits problem

is a prototypical statistical learning problem that exhibits an exploration-exploitation trade-off. One is given multiple options, referred to as arms, and each of them associated with a probability distribution. The emphasis is put on focusing as quickly as possible on the best available option, rather than estimating with high confidence the statistics of each option. The cornerstone of this field is the pioneering work of Lai and Robbins [56]. Here we study two variants of the multi-armed bandits problem where the probability distributions of the arms form Markov chains. First we consider the task of identifying with some fixed confidence an approximately best arm, and we use our bound to analyze the median elimination algorithm, originally proposed in [33] for the case of i.i.d. bandits. Then we turn into the problem of regret minimization for Markovian bandits, where we analyze the UCB algorithm that was introduced in [5] for i.i.d. bandits. For a thorough introduction to multi-armed bandits we refer the interested reader to the survey [15], and the books [58, 84].

2.2 A Hoeffding Inequality for Irreducible Finite State Markov Chains

The central quantity that shows up in our Hoeffding inequality, and makes it differ from the classical i.i.d. Hoeffding inequality, is the maximum hitting time of a Markov chain with an irreducible transition probability matrix P. This is defined as,

$$\mathrm{HitT}(P) = \max_{x,y \in S} \mathbb{E}[T_y \mid X_1 = x],$$

where $T_y = \inf\{n \geq 0 : X_{n+1} = y\}$ is the number of transitions taken in order to visit state y for the first time. $\mathrm{HitT}(P)$ is ensured to be finite due to irreducibility and the finiteness of the state space.

Theorem 4. *Let $\{X_k\}_{k \in \mathbb{Z}_{\geq 0}}$ be a Markov chain on a finite state space S, driven by an initial distribution q, and an irreducible transition probability matrix P. Let $f : S \to [a, b]$ be a real-valued function. Then, for any $t > 0$,*

$$\mathbb{P}_q \left(\left| \sum_{k=1}^{n} \left(f(X_k) - \mathbb{E}_q \left[f(X_k) \right] \right) \right| \geq t \right) \leq 2 \exp \left\{ -\frac{t^2}{2\nu^2} \right\},$$

where $\nu^2 = \frac{1}{4} n (b-a)^2 \mathrm{HitT}(P)^2$.

Proof. We define the sums $S_{l,m} = f(X_l) + \ldots + f(X_m)$, for $1 \leq l \leq m \leq n$, and the filtration $\mathcal{F}_0 = \sigma(\emptyset)$, $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$ for $k = 1, \ldots, n$. Then $\{\mathbb{E}(S_{1,n} \mid \mathcal{F}_k) - \mathbb{E}(S_{1,n} \mid \mathcal{F}_0)\}_{k=0}^{n}$, is a zero mean martingale with respect to $\{\mathcal{F}_k\}_{k=0}^{n}$, and let $\Delta_k = \mathbb{E}(S_{1,n} \mid \mathcal{F}_k) - \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1})$, for $k = 1, \ldots, n$, be the martingale differences.

We first note the following bounds on the martingale differences,

$$\min_{y \in S} \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}, X_k = y) - \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}) \leq \Delta_k,$$

and

$$\Delta_k \leq \max_{x\in S} \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}, X_k = x) - \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}).$$

Therefore, in order to bound the variation of Δ_k it suffices to control,

$$\begin{aligned}
&\max_{x\in S} \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}, X_k = x) - \min_{y\in S} \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}, X_k = y) \\
&\quad = \max_{x,y\in S} \{\mathbb{E}[S_{k,n} \mid X_k = x] - \mathbb{E}[S_{k,n} \mid X_k = y]\} \\
&\quad = \max_{x,y\in S} \{\mathbb{E}[S_{1,n-k+1} \mid X_1 = x] - \mathbb{E}[S_{1,n-k+1} \mid X_1 = y]\},
\end{aligned}$$

where in the first equality we used the Markov property, and in the second the time-homogeneity.

We now use a hitting time argument. Observe the following pointwise statements,

$$S_{1,n-k+1} \leq T_y b + S_{T_y+1,n-k+1}, \quad S_{T_y+1,n-k+1} + T_y a \leq S_{T_y+1,T_y+n-k+1},$$

from which we deduce that,

$$S_{1,n-k+1} \leq T_y(b-a) + S_{T_y+1,T_y+n-k+1}.$$

Taking $\mathbb{E}[\cdot \mid X_1 = x]$-expectations, and using the strong Markov property we obtain,

$$\mathbb{E}[S_{1,n-k+1} \mid X_1 = x] \leq (b-a)\,\mathbb{E}[T_y \mid X_1 = x] + \mathbb{E}[S_{1,n-k+1} \mid X_1 = y].$$

Therefore,

$$\max_{x,y\in S} \{\mathbb{E}[S_{1,n-k+1} \mid X_1 = x] - \mathbb{E}[S_{1,n-k+1} \mid X_1 = y]\} \leq (b-a)\mathrm{HitT}(P).$$

With this in our possession we apply Hoeffding's lemma, see for instance Lemma 2.3 in [25], in order to get,

$$\mathbb{E}\left(e^{\theta\Delta_k} \mid \mathcal{F}_{k-1}\right) \leq \exp\left\{\frac{\theta^2(b-a)^2\mathrm{HitT}(P)^2}{8}\right\} = \exp\left\{\frac{\theta^2\nu^2}{2n}\right\}, \text{ for all } \theta \in \mathbb{R}.$$

Using Markov's inequality, and successive conditioning we obtain that for $\theta > 0$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=1}^{n} (f(X_k) - \mathbb{E}_q[f(X_k)]) \geq t\right) &\leq e^{-\theta t}\,\mathbb{E}\left[e^{\theta\left(\sum_{k=1}^{n}\Delta_k\right)}\right] \\
&= e^{-\theta t}\,\mathbb{E}\left[\mathbb{E}\left(e^{\theta\Delta_n}\middle|\mathcal{F}_{n-1}\right) e^{\theta\left(\sum_{k=1}^{n-1}\Delta_k\right)}\right] \\
&\leq \exp\left\{-\theta t + \frac{\theta^2\nu^2}{2n}\right\}\mathbb{E}\left[e^{\theta\left(\sum_{k=1}^{n-1}\Delta_k\right)}\right] \\
&\leq \ldots \leq \exp\left\{-\theta t + \frac{\theta^2\nu^2}{2}\right\}.
\end{aligned}$$

Plugging in $\theta = t/\nu^2$, we see that,

$$\mathbb{P}\left(\sum_{k=1}^{n}\left(f(X_k) - \mathbb{E}_q\left[f(X_k)\right]\right) \geq t\right) \leq \exp\left\{-\frac{t^2}{2\nu^2}\right\}.$$

The conclusion follows by combining the inequality above for f and $-f$. □

Example 1. Consider a two-state Markov chain with $S = \{0, 1\}$ and $P(0, 1) = p,\ P(1, 0) = r$, with $p, r \in (0, 1]$. Then,

$$\mathrm{HitT}(P) = \max\{\mathbb{E}[\text{Geometric}(p)], \mathbb{E}[\text{Geometric}(r)]\} = 1/\min\{p, r\},$$

and Theorem 4 takes the form,

$$\mathbb{P}_q\left(\left|\sum_{k=1}^{n}\left(f(X_k) - \mathbb{E}_q\left[f(X_k)\right]\right)\right| \geq t\right) \leq 2\exp\left\{-\frac{2\min\{p^2, r^2\}t^2}{n(b-a)^2}\right\}.$$

Example 2. Consider the random walk on the m-cycle with state space $S = \{0, 1, \ldots, m-1\}$, and transition probability matrix $P(x, y) = (1\{y \equiv x + 1 \pmod m\} + 1\{y \equiv x - 1 \pmod m\})/2$. If m is odd, then the Markov chain is aperiodic, while if m is even, then the Markov chain has period 2. Then,

$$\mathrm{HitT}(P) = \max_{y \in S} \mathbb{E}[T_y \mid X_1 = 0] = \max_{y \in S} y(m - y) = \lfloor m^2/4 \rfloor,$$

and Theorem 4 takes the form,

$$\mathbb{P}_q\left(\left|\sum_{k=1}^{n}\left(f(X_k) - \mathbb{E}_q\left[f(X_k)\right]\right)\right| \geq t\right) \leq 2\exp\left\{-\frac{2t^2}{n(b-a)^2\lfloor m^2/4 \rfloor^2}\right\}.$$

Remark 1. Observe that the technique used to establish Theorem 4 is limited to Markov chains with a finite state space S. Indeed, if $\{X_k\}_{k \in \mathbb{Z}_{>0}}$ is a Markov chain on a countably infinite state space S with an irreducible and positive recurrent transition probability matrix P and a stationary distribution π, then we claim that,

$$\frac{1}{\pi(y)} \leq 1 + \sup_{x \in S} \mathbb{E}[T_y \mid X_1 = x], \text{ for all } y \in S,$$

from which it follows that $\sup_{x,y \in S} \mathbb{E}[T_y \mid X_1 = x] = \infty$, due to the fact that $\sum_{y \in S} \pi(y) = 1$ and S is countably infinite. The aforementioned inequality can be established as follows.

$$\begin{aligned}
\frac{1}{\pi(y)} &= \mathbb{E}[\inf\{n \geq 1 : X_{n+1} = y\} \mid X_1 = y] \\
&= \sum_{x \in S} \mathbb{E}[\inf\{n \geq 1 : X_{n+1} = y\} \mid X_2 = x]P(y, x) \\
&\leq \sup_{x \in S} \mathbb{E}[\inf\{n \geq 1 : X_{n+1} = y\} \mid X_2 = x] \\
&= 1 + \sup_{x \in S} \mathbb{E}[T_y \mid X_1 = x].
\end{aligned}$$

Moreover, through Theorem 4 we can obtain a concentration inequality for sums of a function evaluated on the transitions of a Markov chain. In particular, let

$$S^{(2)} = \{(x, y) \in S \times S : P(x, y) > 0\}.$$

On the state space $S^{(2)}$ define the transition probability matrix,

$$P^{(2)}\left((x, y), (z, w)\right) = I\{y = z\}P(y, w), \text{ for } (x, y), (z, w) \in S^{(2)}.$$

It is straightforward to verify that the fact that P is irreducible, implies that $P^{(2)}$ is irreducible as well. This readily gives the following theorem.

Theorem 5. *Let $\{X_k\}_{k\in\mathbb{Z}_{>0}}$ be a Markov chain on a finite state space S, driven by an initial distribution q, and an irreducible transition probability matrix P. Let $f^{(2)} : S^{(2)} \to [a, b]$ be a real-valued function evaluated on the transitions of the Markov chain. Then, for any $t > 0$,*

$$\mathbb{P}_q\left(\left|\sum_{k=1}^{n}\left(f^{(2)}(X_k, X_{k+1}) - \mathbb{E}_q\left[f^{(2)}(X_k, X_{k+1})\right]\right)\right| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{2\nu^2}\right\},$$

where $\nu^2 = \frac{1}{4}n(b-a)^2\mathrm{HitT}\left(P^{(2)}\right)^2$.

Corollary 1. *When the Markov chain is initialized with its stationary distribution, π, Theorem 4 and Theorem 5 give the following nonasymptotic versions of the weak law of large numbers for irreducible Markov chains. For any $\epsilon > 0$,*

$$\mathbb{P}_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbb{E}_\pi\left[f(X_1)\right]\right| \geq \epsilon\right) \leq 2\exp\left\{-\frac{2n\epsilon^2}{(b-a)^2\mathrm{HitT}(P)^2}\right\},$$

and,

$$\mathbb{P}_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n} f^{(2)}(X_k, X_{k+1}) - \mathbb{E}_\pi\left[f^{(2)}(X_1, X_2)\right]\right| \geq \epsilon\right) \leq 2\exp\left\{-\frac{2n\epsilon^2}{(b-a)^2\mathrm{HitT}\left(P^{(2)}\right)^2}\right\}.$$

2.3 Optimal transport for Markov chains

In this section we study the bicausal optimal transport problem for Markov chains, an optimal transport formulation suitable for stochastic processes which takes into consideration the accumulation of information as time evolves. Our analysis is based on a relation between the transport problem and the theory of Markov decision processes. This way we are able to derive necessary and sufficient conditions for optimality in the transport problem, as well as an iterative algorithm, namely the value iteration, for the calculation of the transportation cost. Additionally, we draw the connection with the classic theory on couplings for Markov chains, and in particular with the notion of faithful couplings. Finally, we illustrate how the transportation cost appears naturally in the study of concentration of measure for Markov chains.

2.3.1 Problem setting

Let S be a finite set equipped with the discrete topology, and let $\mathcal{S}$ be the corresponding Borel σ-field which in this case is the set of all subsets of S. Let S^∞ be the countable infinite product space, equipped with the product topology, and let $\mathcal{S}^\infty$ be the corresponding Borel σ-field. Let $\mathbf{X} = (X_k)_{k=0}^\infty$, $\mathbf{X}' = (X'_k)_{k=0}^\infty$ be two discrete time stochastic processes on the measurable space $(S^\infty, \mathcal{S}^\infty)$, governed by the probability laws μ, μ' respectively. By a coupling of $\mathbf{X}, \mathbf{X}'$ we mean a pair of stochastic processes $(\hat{\mathbf{X}}, \hat{\mathbf{X}}')$, on the measurable space $(S^\infty \times S^\infty, \mathcal{S}^\infty \otimes \mathcal{S}^\infty)$, governed by a probability law γ such that $\gamma(\cdot, S^\infty) = \mu$, and $\gamma(S^\infty, \cdot) = \mu'$. Denote the set of all couplings of μ, μ' by

$$\Gamma(\mu, \mu') = \{\gamma \in \mathcal{P}(S^\infty \times S^\infty) : \gamma(\cdot, S^\infty) = \mu,\ \gamma(S^\infty, \cdot) = \mu'\},$$

where $\mathcal{P}(S^\infty \times S^\infty)$ denotes the set of all probability laws on $(S^\infty \times S^\infty, \mathcal{S}^\infty \otimes \mathcal{S}^\infty)$.

Let $\mathbf{c} : S^\infty \times S^\infty \to [0, \infty]$ be a $\mathcal{S}^\infty \otimes \mathcal{S}^\infty$-measurable cost function, which has the following additive form

$$\mathbf{c}(\mathbf{x}, \mathbf{x}') = \sum_{k=0}^{\infty} \beta^k c(x_k, y_k), \tag{2.2}$$

for some $\beta \in (0, 1]$, and some $c : S \times S \to [0, \infty)$. In particular, it will be of special interest the case that the cost function $\mathbf{c}$ is the metric

$$\mathbf{d}_\beta(\mathbf{x}, \mathbf{x}') = \sum_{k=0}^{\infty} \beta^k I\{x_k \neq x'_k\}, \tag{2.3}$$

which for $\beta \in (0, 1)$ induces the product topology on S^∞. In a typical optimal transport problem, see for instance the book of [90], we are interested in finding a coupling γ which minimizes the cost function $\mathbf{c}$ according to the following cost criterion

$$W(\mu, \mu') = \inf_{\gamma \in \Gamma(\mu, \mu')} \int \mathbf{c}(\mathbf{x}, \mathbf{x}') \gamma(d\mathbf{x}, d\mathbf{x}'). \tag{2.4}$$

Such a formulation might be inadequate in the context of stochastic processes, as the evolution over time matters, and has to be accounted. In the context of finite horizon processes the works of [78, 79] recognize this and introduce the nested distance which takes into consideration the filtrations. They are motivated by applications of the nested distance for scenario reduction in the context of multistage stochastic optimization. See also the work of [6] for an application of the nested distance to stability in mathematical finance. The metric and topological properties of the nested distance has been recently studied in [7]. A generalization of the nested distance is the causal optimal transport problem introduced by [57] and further developed by [8] where a dynamic programming principle is developed as well. All those works deal in great generality with processes of finite horizon. In this thesis we study the bicausal optimal transport problem for Markov chains over an infinite horizon, drawing motivation by the classic theory of couplings for Markov chains where one might

naturally seek for an adapted coupling, i.e. one that cannot see into the future, that couples two chains as fast as possible.

To introduce the bicausal optimal transport problem we first note that any probability law γ can be factorized, for every n, as

$$\begin{aligned}&\gamma((dx_0,\dots,dx_n),(dx_0',\dots,dx_n'))\\&\quad=\gamma(dx_0,dx_0')\gamma(dx_1,dx_1'\mid x_0,x_0')\cdots\gamma(dx_n,dx_n'\mid x_0,\dots,x_{n-1},x_0',\dots,x_{n-1}'),\end{aligned}\tag{2.5}$$

where $\gamma(\cdot\mid x_0,\dots,x_{k-1},x_0',\dots,x_{k-1}')$ denotes the conditional probability law of $((\hat{X})_{i=k}^{\infty},(\hat{X}')_{i=k}^{\infty})$ given that $\hat{X}_0=x_0,\dots,\hat{X}_{k-1}=x_{k-1},\hat{X}_0'=x_0',\dots,\hat{X}_{k-1}'=x_{k-1}'$. In this work we are interested in couplings γ which are bicausal, in the sense that for every n, $\gamma(dx_n,S\mid x_0,\dots,x_{n-1},x_0',\dots,x_{n-1}')=\mu(dx_n\mid x_0,\dots,x_{n-1})$, and $\gamma(S,dx_n'\mid x_0,\dots,x_{n-1},x_0',\dots,x_{n-1}')=\mu'(dx_n'\mid x_0',\dots,x_{n-1}')$. We denote the set of all bicausal couplings of μ,μ' by

$$\Gamma_{bc}(\mu,\mu')=\left\{\gamma\in\mathcal{P}(S^{\infty}\times S^{\infty}):\ \begin{array}{l}\gamma(dx_n,S\mid x_0,\dots,x_{n-1},x_0',\dots,x_{n-1}')=\mu(dx_n\mid x_0,\dots,x_{n-1})\\ \gamma(S,dx_n'\mid x_0,\dots,x_{n-1},x_0',\dots,x_{n-1}')=\mu'(dx_n'\mid x_0',\dots,x_{n-1}')\end{array}\right\}.$$

Due to the factorization (2.5) it is clear that $\Gamma_{bc}(\mu,\mu')\subseteq\Gamma(\mu,\mu')$. Additionally, the product measure $\mu\otimes\mu'\in\Gamma_{bc}(\mu,\mu')$, hence none of those sets is empty. The corresponding bicausal optimal transport problem can be written as

$$W_{bc}(\mu,\mu')=\inf_{\gamma\in\Gamma_{bc}(\mu,\mu')}\int\mathbf{c}(\mathbf{x},\mathbf{x}')\gamma(d\mathbf{x},d\mathbf{x}').\tag{2.6}$$

The bicausal optimal transport problem (2.6) is particular interesting in the case that μ,μ' are Markovian laws, i.e. $(X_k)_{k=0}^{\infty},(X_k')_{k=0}^{\infty}$ are Markov chains. For the rest of this chapter we assume that there are initial states $x_0,x_0'\in S$, and transition probability kernels $P,P':S\times\mathcal{S}\to[0,1]$ such that for every n

$$\begin{aligned}\mu(\{x_0\},dx_1,\dots,dx_n)&=P(x_0,dx_1)P(x_1,dx_2)\cdots P(x_{n-1},dx_n),\\ \mu'(\{x_0'\},dx_1',\dots,dx_n')&=P'(x_0',dx_1')P'(x_1',dx_2')\cdots P'(x_{n-1}',dx_n'),\end{aligned}$$

and we write

$$\mu=\mathrm{Markov}(x_0,P),\quad \mu'=\mathrm{Markov}(x_0',P').$$

We note that using two fixed initial states x_0,x_0' is as general as considering arbitrary initial distributions, since x_0,x_0' can be thought of as auxiliary states inducing arbitrary initial distributions, $P(x_0,\cdot),P(x_0',\cdot)$ to X_1,X_1' respectively. For extra clarity we rewrite (2.6) as

$$W_{bc}\left(\mathrm{Markov}(x_0,P),\mathrm{Markov}(x_0',P')\right)=\inf_{\gamma\in\Gamma_{bc}\left(\mathrm{Markov}(x_0,P),\mathrm{Markov}(x_0',P')\right)}\int\mathbf{c}(\mathbf{x},\mathbf{x}')\gamma(d\mathbf{x},d\mathbf{x}').\tag{2.7}$$

We study the transportation problem (2.7) in Subsection 2.3.3 under the lens of dynamic programming, where we develop optimality conditions, as well as an iterative procedure, namely the value iteration, that solves the transportation problem.

Motivation In the special case that $P = P'$, and the cost function $\mathbf{c}$ is the metric $\mathbf{d}_\beta$ given in (2.3), the transportation problem (2.4) can be solved explicitly by works on maximal coupling from [42, 80, 41]. In particular

$$W\left(\text{Markov}(x_0, P), \text{Markov}(x_0', P)\right) = \sum_{k=0}^{\infty} \beta^k \left\| P^k(x_0, \cdot) - P^k(x_0', \cdot) \right\|_{TV},$$

where $\|\cdot\|_{TV}$ stands for half the total variation norm of a signed measure.

When $\beta = 1$ the transportation problem (2.4) reduces to finding a coupling of two different initializations of the same Markov chain that couples them in the smallest expected time. By introducing the coupling time

$$T = \inf\left\{ n \geq 0 : \hat{X}_n = \hat{X}_n' \right\},$$

the transportation problem (2.4) can be written as

$$W\left(\text{Markov}(x_0, P), \text{Markov}(x_0', P)\right) = \inf_{\gamma \in \Gamma\left(\text{Markov}(x_0, P), \text{Markov}(x_0', P)\right)} \mathbb{E}^\gamma\left[T\right]. \tag{2.8}$$

This transportation problem is particularly important because it directly leads to a bounded differences concentration inequality for Markov chains as we discuss in Section 2.4.

The maximal coupling of [80], works on the space-time plane by first simulating the meeting point, and then constructing the forward and backward chains. As such the coupling 'cheats' by looking into the future. [82] initiates the discussion of faithful couplings that do not look into the future, motivated by the fact that such couplings automatically posses the 'now equals forever' property which roughly speaking says that the two chains becoming equal at a single time is equivalent to having them remain equal for all future times. It is the bicausal version of the transportation problem

$$W_{bc}((\text{Markov}(x_0, P), \text{Markov}(x_0', P'))) = \inf_{\gamma \in \Gamma_{bc}\left(\left(\text{Markov}(x_0, P), \text{Markov}(x_0', P')\right)\right)} \mathbb{E}^\gamma\left[T\right], \tag{2.9}$$

that seeks for faithful couplings, that do not look into the future, and minimize the expected coupling time. In Subsection 2.3.4 we provide necessary and sufficient conditions for optimality at the transportation problem (2.9), as well as a discussion about properties of optimal couplings.

2.3.2 Markovian Couplings

Among the set of all bicausal couplings, $\Gamma_{bc}\left(\text{Markov}(x_0, P),\ \text{Markov}(x_0', P')\right)$, it suffices to turn our attention to Markovian couplings when considering the transportation problem (2.6). We will establish this in Subsection 2.3.3 as a consequence of the dynamic programming theory. A Markovian coupling of $\text{Markov}(x_0, P),\ \text{Markov}(x_0', P')$ is specified by

a transition kernel $Q : (S \times S) \times (\mathcal{S} \otimes \mathcal{S}) \to [0,1]$ such that $Q((x,x'),(\cdot,S)) = P(x,\cdot)$, and $Q((x,x'),(S,\cdot)) = P'(x',\cdot)$. We denote the set of all such transition kernels by

$$\Gamma_M(P,P') = \{Q : Q((x,x'),(\cdot,S)) = P(x,\cdot),\ Q((x,x'),(S,\cdot)) = P'(x',\cdot)\}$$

The corresponding Markovian coupling is given, for every n, by

$$\gamma((\{x_0\},\dots,dx_n),(\{x_0'\},\dots,dx_n')) = Q((x_0,x_0'),(dx_1,dx_1'))\cdots Q((x_{n-1},x_{n-1}'),(dx_n,dx_n')). \tag{2.10}$$

We note that any Markovian coupling is bicausal.

We now present some basic examples of Markovian couplings, for the case of a single Markov chain $P = P'$, which have been used extensively in the coupling literature, see for instance the book [89].

Example 3. The classic coupling, initially introduced by Doeblin in order to establish the convergence theorem for Markov chains, asserts that $\hat{X}_n$ and $\hat{X}_n'$ evolve independently until they reach a common state for the first time, and afterwards they move identically.

$$Q_{\text{classic}}((x,x'),(y,y')) = \begin{cases} P(x,y)P(x',y'), & \text{if } x \neq x', \\ P(x,y), & \text{if } x = x',\ \text{and } y = y', \\ 0, & \text{otherwise,} \end{cases} \tag{2.11}$$

Example 4. A variant of the classic coupling asserts that $\hat{X}_n$ and $\hat{X}_n'$ evolve independently at all times, and this independent coupling can be used as well to establish the convergence theorem for Markov chains.

$$Q_{\text{ind}}((x,x'),(y,y')) = P(x,y)P(x',y'), \tag{2.12}$$

In both the classic and the independent coupling it is apparent that if we seek for a Markovian coupling that minimizes some cost criterion, e.g. attaining coupling at the smallest expected time, then the independent movement can be wasteful. Instead, one should coordinate the movement of the two copies in a way that optimizes the objective under consideration. A first such attempt is the following coupling attributed to Wasserstein.

Example 5. Given that $\hat{X}_{n-1} = x$ and $\hat{X}_{n-1}' = x'$, the Wassertstein coupling makes $\hat{X}_n$ and $\hat{X}_n'$ agree with as great probability as possible (which is $1 - \|P(x,\cdot) - P(x',\cdot)\|_{TV}$), and then given that they differ they are made conditionally independent.

$$Q_{\text{W}}((x,x'),(y,y')) = \begin{cases} 0, & \text{if } x = x',\ \text{and } y \neq y', \\ P(x,y), & \text{if } x = x',\ \text{and } y = y', \\ P(x,y) \wedge P(x',y), & \text{if } x \neq x',\ \text{and } y = y', \\ 0, & \text{if } x \neq x',\ y \neq y',\ \text{and } \|P(x,\cdot) - P(x',\cdot)\|_{TV} = 0, \\ \frac{(P(x,y) - P(x',y))^+ (P(x',y') - P(x,y'))^+}{\|P(x,\cdot) - P(x',\cdot)\|_{TV}}, & \text{if } x \neq x',\ y \neq y',\ \text{and } \|P(x,\cdot) - P(x',\cdot)\|_{TV} \neq 0, \end{cases} \tag{2.13}$$

where $a \wedge b = \min(a,b)$, $a^+ = -((-a) \wedge 0)$, and $a^- = a \wedge 0$.

Making $\hat{X}_n$ and $\hat{X}'_n$ agree with as great probability as possible is a good first step towards a Markovian coupling with the smallest expected coupling time, although it might be too greedy of a choice and the conditional independence assertion surely leaves more room for improvements. It is the objective of this section to provide a characterization of optimal Markovian couplings using the theory of dynamic programming.

2.3.3 Bicausal optimal transport for Markov chains via dynamic programming

We start by illustrating that the bicasual transport problem for Markov chains (2.7) can be viewed as an instance of infinite horizon dynamic programming. When $\beta \in (0,1)$ we have an instance of discounted dynamic programming, initially studied by [12], while when $\beta = 1$ we have an instance of negative dynamic programming, initially studied by [85]. For two Markov chains $\text{Markov}(x_0, P)$, $\text{Markov}(x'_0, P')$ on the same state space $(S, \mathcal{S})$, and for the bicausal optimal transport problem (2.7) between them, the associated underlying Markov decision process is given by the tuple $((S \times S, \mathcal{S} \otimes \mathcal{S}), (A, \mathcal{A}), U, q, \beta, c)$ where:

- $(S \times S, \mathcal{S} \otimes \mathcal{S})$ stands for the state space of the Markov decision process.
- $(A, \mathcal{A})$ is the action space, where $A = \mathcal{P}(S \times S)$ is the set of all probability distributions on $S \times S$ equipped with the subspace topology induced from the standard topology on $\mathbb{R}^{|S \times S|}$, and $\mathcal{A}$ stands for the corresponding Borel σ-field.
- $U(x, x') = \{a \in A : a(\cdot, S) = P(x, \cdot),\ a(S, \cdot) = P'(x', \cdot)\}$ is the set of all allowable actions at state (x, x'), i.e. all the probability distributions on $S \times S$ which respect the coupling constraints.
- $q(\cdot \mid (x, x'), a) = a$ is the distribution of the state next visited by the Markov decision process if the system is currently in state (x, x') and action $a \in U(x, x')$ is taken.
- $\beta \in (0, 1]$ is the discount factor.
- $c : S \times S \to [0, \infty)$ is the cost function.

A policy is a bicausal coupling $\mu \in \Gamma_{bc}(\text{Markov}(x_0, P), \text{Markov}(x'_0, P'))$, and it can be decomposed as a sequence of conditional distributions as in (2.5) so that if the coupling μ is used, and up to time n we observe the trajectory $x_0, \ldots, x_n, x'_0, \ldots, x'_n$ then the action taken at time n is $\mu(\cdot \mid x_0, \ldots, x_n, x'_0, \ldots, x'_n)$ which is also the distribution of the state visited by the Markov decision process at time $n + 1$. As it turns out there exists an optimal coupling for which the conditional distributions do not depend on the whole trajectory but just on the current state, i.e. there exists an optimal coupling which is Markovian in the sense of (2.10).

We proceed with the definition of some typical operators from the dynamic programming literature. Let F be the set of all extended real valued functions $V : S \times S \to [0, \infty]$. For

$Q \in \Gamma_M(P, P')$ define the operator $T_Q : F \to F$ by

$$T_Q(V)(x, x') = c(x, x') + \beta \int Q\left((x, x'), (dy, dy')\right) V(y, y'),$$

which we may also write in functional notation as

$$T_Q(V) = c + \beta QV.$$

Additionally, define the Bellman operator $T : F \to F$ by

$$T(V)(x, x') = c(x, x') + \beta \inf_{a \in U(x,x')} \int a(dy, dy')V(y, y'),$$

or in functional notation as

$$T(V) = c + \beta \inf_{Q \in \Gamma_M(P,P')} QV.$$

We note that when $\beta \in (0, 1)$ the Bellman operator T is a β-contraction with respect to the sup-norm on F, and when $\beta = 1$ the Bellman operator T is an increasing mapping. In the following theorem we summarize the main results from this dynamic programming interpretation of the bicausal optimal transport problem between two Markov chains (2.7). As it is typical in dynamic programming we consider the transportation cost $W_{bc}\left(\text{Markov}(x_0, P), \text{Markov}(x_0', P')\right)$ as a function $W_{bc} : S \times S \to [0, \infty]$ of the initializations of the two Markov chains, and we write $W_{bc}(x_0, x_0')$ for the optimal cost.

Theorem 6.

1. *The transportation cost W_{bc} is a solution to the fixed point equation $V = T(V)$. When $\beta \in (0, 1)$ it is the unique solution, while when $\beta = 1$ if $V \geq 0$ and $V = T(V)$ then $V \geq W_{bc}$.*

2. *The transportation cost W_{bc} can be calculated via the fixed point iteration*

$$\begin{cases} V_0 &= 0, \\ V_k &= T(V_{k-1}), \quad k = 1, 2, \ldots, \end{cases} \tag{2.14}$$

   If $\beta \in (0, 1)$, then $\|V_k - W_{bc}\|_\infty \leq \beta^k \|W_{bc}\|_\infty$, and thus $V_k \to W_{bc}$ as $k \to \infty$ at a linear rate. If $\beta = 1$, then the convergence is monotonic, $V_k \uparrow W_{bc}$ as $k \to \infty$.

3. *There exists an optimal Markovian coupling.*

4. *Q is an optimal Markovian coupling if and only if $T_Q(W_{bc}) = W_{bc}$.*

Proof. When $\beta \in (0,1)$ parts 1, 2 and 4 follow from Proposition 1 in [11]. When $\beta = 1$ parts 1 and 3 follows from Propositions 5 and 7 in [11].

For the rest we need to note that for every $x, x' \in S,\ \lambda \in [0,\infty)$ and k, the set

$$U_k((x,x'),\lambda) = \left\{ a \in U(x,x') : c(x,x') + \beta \int a(dy,dy')V_k(y,y') \leq \lambda \right\},$$

is compact as the intersection of the compact set $U(x,x')$, with a closed half-space.

Then for $\beta \in (0,1)$ part 3 follows from Proposition 14 in [11], and for $\beta = 1$ parts 2 and 3 follow from Proposition 12 in [11]. □

We note that due to the special structure of $U(x,x')$, it is a convex polytope arising from the intersection of a probability simplex with an affine space, the value iteration (2.14) proceeds by solving at each iteration a linear program. Thus the value iteration (2.14) in this case can be thought as sequence of finite dimensional linear programs approximating the bicausal optimal transport cost W_{bc} which in (2.7) is formulated as an infinite dimensional linear program.

2.3.4 Coupling of Markov chains in minimum expected time

In this section we specialize the bicausal optimal transport for Markov chains to the case that we have a single irreducible and aperiodic transition kernel P, and the cost function $\mathbf{c}$ is the metric $\mathbf{d}_1$. So essentially we are solving for the bicausal coupling that couples two Markov chains with different initializations and the same transition kernel in the smallest expected time

$$\begin{aligned} W_{bc}(\mathrm{Markov}(x_0,P),\mathrm{Markov}(x_0',P)) &= \inf_{\gamma \in \Gamma_{bc}\left(\mathrm{Markov}(x_0,P),\mathrm{Markov}(x_0',P')\right)} \int \mathbf{d}_1(\mathbf{x},\mathbf{x}')\gamma(d\mathbf{x},d\mathbf{x}') \\ &= \inf_{\gamma \in \Gamma_{bc}\left(\left(\mathrm{Markov}(x_0,P),\mathrm{Markov}(x_0',P')\right)\right)} \mathbb{E}^{\gamma}\left[T\right]. \end{aligned}$$

Although in the general framework of negative dynamic programming the fixed point equation $V = T(V)$ is only a necessary condition for optimality, in our specialized setting we can establish that it is also sufficient, giving thus a complete set of necessary and sufficient conditions for both the optimal cost W_{bc}, and the optimal Markovian coupling Q.

Theorem 7. *W_{bc} is the unique solution of the equations*

$$0 \leq V < \infty, \quad V = T(V), \quad \textit{and} \quad V(x,x) = 0 \quad \textit{for } x \in S. \tag{2.15}$$

Proof. We already know from Theorem 6 that $W_{bc} = T(W_{bc})$. Using the classic Markovian coupling, (2.11) we see that $W_{bc}(x,x) = 0$ for all $x \in S$. Using the independent Markovian coupling, (2.12), which induces an irreducible Markov chain on $S \times S$ we see that

$$W_{bc}(x,x') \leq \min_{y \in S} \mathbb{E}^{Q_{\mathrm{ind}}}_{(x,x')}[T_{(y,y)}] < \infty, \quad \text{for all } x, x' \in S.$$

Let $V : S \to [0, \infty)$ be any function satisfying equations (2.15). Let Q be a Markovian coupling such that $W_{bc} = c + QW_{bc}$. Then

$$V = T(V) \leq c + QV = c + QT(V) \leq c + Qc + Q^2V \leq \cdots \leq \sum_{k=0}^{n-1} Q^k c + Q^n V. \tag{2.16}$$

By definition of Q we have that

$$W_{bc} = \sum_{k=0}^{\infty} Q^k c, \tag{2.17}$$

and since $W_{bc} < \infty$ we see that $\lim_{n\to\infty} Q^n c = 0$. We clearly have that $c \leq V$, and because $c(x, x') = 0 \;\Rightarrow\; V(x, x') = 0$, we obtain that $c \leq V \leq \|V\|_\infty c$. Since $V < \infty$ we deduce that

$$\lim_{n\to\infty} Q^n V = 0. \tag{2.18}$$

Combining (2.16), (2.17), and (2.18) we obtain that $V \leq W_{bc}$, and because from Theorem 6 W_{bc} is the minimal fixed point we deduce that $V = W_{bc}$. □

Next we dig in some properties of an optimal Markovian coupling. In particular, we show that an optimal Markovian coupling enjoys the 'sticky' property of the classic coupling (2.11), i.e. under an optimal Markovian coupling the two chains evolve in the same way as soon as they meet.

Lemma 1. *Any optimal Markovian coupling Q sticks to the diagonal as soon as it hits it, i.e.*

$$Q((x, x), (y, y')) = I\{y = y'\}P(x, y).$$

Proof. Fix an optimal Markovian coupling Q. From Theorem 6 it satisfies the equation

$$c + QW_{bc} = W_{bc},$$

and so in particular

$$\int Q((x, x), (dy, dy'))W_{bc}(y, y') = 0.$$

Since for $y \neq y'$, $W_{bc}(y, y') \geq 1$ we have that $Q((x, x), (y, y')) = 0$. Then it follows from the coupling constraint that $Q((x, x), (y, y)) = P(x, y)$. □

Additionally, we show that for two state chains the Wasserstein coupling (2.13) is the only optimal Markovian coupling.

Lemma 2. *When $|S| = 2$ there is a unique optimal Markovian coupling which is precisely the Wasserstein coupling* (2.13).

Proof. Let $S = \{x, x'\}$. Due to symmetry we have that that $W_{bc}(x, x') = W_{bc}(x', x)$, and in addition $W_{bc}(x, x) = W_{bc}(x', x') = 0$. So from Theorem 6 we get that

$$W_{bc}(x, x') = 1 + \min_{a \in U(x,x')} (1 - a(x, x) - a(x', x'))W_{bc}(x, x'). \tag{2.19}$$

It is clear that in the minimization in (2.19) we need to make $a(x, x)$, and $a(x', x')$ as large as possible. Due to the coupling constraint, $a \in U(x, x')$, those largest values are

$$a(x, x) = P(x, x) \wedge P(x', x), \quad \text{and} \quad a(x', x') = P(x, x') \wedge P(x', x').$$

Then from the coupling constraints there are unique corresponding values for $a(x, x')$, and $a(x', x)$. In particular

$$\begin{aligned} a(x, x') &= (P(x, x) - P(x', x))^+ = (P(x', x') - P(x, x'))^+, \\ a(x', x) &= (P(x', x) - P(x, x))^+ = (P(x, x') - P(x', x'))^+. \end{aligned}$$

We further note that

$$\|P(x, \cdot) - P(x', \cdot)\|_{TV} = |P(x, x) - P(x', x)| = |P(x', x') - P(x, x')|.$$

Hence in conjunction with Lemma 1 we conclude that there exists a unique optimal Markovian coupling and this is the Wassertstein coupling (2.13).

Moreover, we have closed form expressions for both the non-causal and the bicausal optimal transport costs

$$\begin{aligned} W(x, x') &= \frac{|P(x, x') - P(x', x)|}{P(x, x') + P(x', x)} \cdot \frac{1}{1 - \|P(x, \cdot) - P(x', \cdot)\|_{TV}} \\ &< W_{bc}(x, x') = \frac{1}{1 - \|P(x, \cdot) - P(x', \cdot)\|_{TV}}. \end{aligned}$$

□

Finally, we give an easy upper bound on the bicausal optimal transport cost for contractive Markov chains.

Lemma 3. *Let $\delta(P) = \max_{x,x' \in S} \|P(x, \cdot) - P(x', \cdot)\|_{TV}$ be the Doeblin-Dobrushin contraction coefficient, and assume that $\delta(P) < 1$. Then*

$$\|W_{bc}\|_\infty \leq \frac{1}{1 - \delta(P)}.$$

Proof. By definition $W_{bc}(x_0, x_0')$ is upper bounded by the cost incurred when the Wasserstein coupling (2.13) is used. Under the Wasserstein coupling, Q_W, the probability that we hit the diagonal in one step from state (x, x') is $1 - \|P(x, \cdot) - P(x', \cdot)\|_{TV}$. Thus $\mathbf{d}_1(\hat{\mathbf{X}}, \hat{\mathbf{X}}')$ under the Markovian coupling induced by Q_W is stochastically dominated by Geometric$(1 - \delta(P))$, and thus

$$W_{bc}(x_0, x_0') \leq \mathbb{E}^{Q_W}_{(x_0, x_0')}[\mathbf{d}_1(\hat{\mathbf{X}}, \hat{\mathbf{X}}')] \leq \frac{1}{1 - \delta(P)}, \quad \text{for any } x_0, x_0' \in S.$$

□

2.4 Concentration of measure for Markov chains

In this section we demonstrate how the transportation cost (2.8) shows up naturally when one studies concentration of measure for Markov chains. This was first observed in the works of [63, 64] about contracting Markov chains, and thereafter greatly generalized for classes of dependent random processes in terms of various mixing coefficients using the transportation-information method [66, 65, 83]. [17] uses the martingale method combined with maximal coupling to derive concentration for dependent processes, while [53] uses the martingale method and a linear programming inequality for the same task. For the Markovian case [77] using the martingale method establishes a concentration inequality that involves the mixing time of the chain.

Let $f : S^n \to \mathbb{R}$ be a functions which is 1-Lipschitz with respect to the Hamming distance

$$f(x_1, \ldots, x_n) - f(x'_1, \ldots, x'_n) \leq \sum_{k=1}^{n} I\{x_k \neq x'_k\}, \quad \text{for } x_1, \ldots, x_n, x'_1, \ldots, x'_n \in S.$$

Let $\mathbf{X} \sim \text{Markov}(x_0, P)$, where P is an irreducible and aperiodic transition kernel. We would like to study the deviation of the random variable $f(X_1, \ldots, X_n)$ from its mean $\mathbb{E}^P_{x_0}[f(X_1, \ldots, X_n)]$. The typical approach to do so is the martingale method. For $k = 0, \ldots, n$ we define the martingale

$$Z_k = \mathbb{E}^P_{x_0} \left(f(X_1, \ldots, X_n) | X_1, \ldots, X_k \right),$$

and for $k = 1, \ldots, n$ we define the martingale differences

$$\Delta_k = Z_k - Z_{k-1}.$$

Then the quantity of which we want to control the deviations can be written as a telescoping sum of the martingale differences

$$f(X_1, \ldots, X_n) - \mathbb{E}^P_{x_0} \left[f(X_1, \ldots, X_n) \right] = \sum_{k=1}^{n} \Delta_k,$$

and it suffices to control the range of the martingale differences. For this we note that

$$\min_{x' \in S} \left\{ \mathbb{E}^P_{x_0} \left(f(X_1, \ldots, X_n) | X_1, \ldots, X_k = x' \right) \right\} - \mathbb{E}^P_{x_0} \left(f(X_1, \ldots, X_n) | X_1, \ldots, X_{k-1} \right) \leq \Delta_k,$$

and that

$$\Delta_k \leq \max_{x \in S} \left\{ \mathbb{E}^P_{x_0} \left(f(X_1, \ldots, X_n) | X_1, \ldots, X_k = x \right) \right\} - \mathbb{E}^P_{x_0} \left(f(X_1, \ldots, X_n) | X_1, \ldots, X_{k-1} \right).$$

Thus in order to bound the length of the range of the martingale difference, we just need to bound

$$\max_{x, x' \in S} \left\{ \mathbb{E}^P_{x_0} \left(f(X_1, \ldots, X_n) | X_1, \ldots, X_k = x \right) - \mathbb{E}^P_{x_0} \left(f(X_1, \ldots, X_n) | X_1, \ldots, X_k = x' \right) \right\}. \tag{2.20}$$

Fix $x_1, \dots, x_{k-1}, x, x' \in S$, and a coupling $\gamma \in \Gamma(\mu(\cdot \mid x_1, \dots, x_{k-1}, x), \mu(\cdot \mid x_1, \dots, x_{k-1}, x'))$, where $\mu = \text{Markov}(x_0, P)$. Then

$$\begin{aligned}
&\mathbb{E}^P_{x_0}\left(f(X_1, \dots, X_n) | X_1 = x_1, \dots, X_{k-1} = x_{k-1}, X_k = x\right) - \\
&\qquad \mathbb{E}^P_{x_0}\left(f(X_1, \dots, X_n) | X_1 = x_1, \dots, X_{k-1} = x_{k-1}, X_k = x'\right) \\
&= \mathbb{E}^\gamma\left[f(x_1, \dots, x_{k-1}, \hat{X}_k, \dots, \hat{X}_n) - f(x_1, \dots, x_{k-1}, \hat{X}'_k, \dots, \hat{X}'_n)\right] \\
&\leq \mathbb{E}^\gamma\left[\sum_{i=k}^{n} I\{\hat{X}_i \neq \hat{X}'_i\}\right] \leq \mathbb{E}^\gamma\left[\sum_{i=k}^{\infty} I\{\hat{X}_i \neq \hat{X}'_i\}\right],
\end{aligned}$$

where we used the Lipschitz condition for f. Thus minimizing over the coupling γ we obtain that

$$\mathbb{E}^P_{x_0}\left(f(X_1, \dots, X_n) | X_1, \dots, X_k = x\right) - \mathbb{E}^P_{x_0}\left(f(X_1, \dots, X_n) | X_1, \dots, X_k = x'\right) \leq W(x, x').$$

All in all, we can bound the length of the range of the martingale difference Δ_k by $\|W\|_\infty$. Then using the standard martingale method one can obtain the concentration inequality

$$\mathbb{P}^P_{x_0}\left(\left|f(X_1, \dots, X_n) - \mathbb{E}^P_{x_0}[f(X_1, \dots, X_n)]\right| \geq t\right) \leq 2\exp\left\{-\frac{2t^2}{n\|W\|^2_\infty}\right\}. \tag{2.21}$$

For a full derivation of a concentration inequality which works in more general dependent settings, than just Markovian dependence, the interested reader is refereed to Theorem 1 in [17] which uses the martingale method together with maximal coupling, and to Theorem 1.1 of [53] which uses a linear programming inequality instead of a coupling argument. The linear programming inequality that appears in [53] actually also corresponds to maximal coupling, as it has been observed in [52]. Additionally, we note that if the Markov chain is periodic and thus coupling techniques are not any more applicable one can still bound (2.20), in the special case that the function f is additive, by using hitting time arguments as it is done in [70].

Clearly $\|W\|_\infty \leq \|W_{bc}\|_\infty$, and thus replacing $\|W\|_\infty$ with $\|W_{bc}\|_\infty$ in (2.21) results in a weaker inequality, although in this way the variance proxy $\|W_{bc}\|^2_\infty$ has the following interpretation: let Q be an optimal Markovian coupling, then $W_{bc}(x, x')$ corresponds to the expected time to hit diagonal when we start from (x, x') and we transition according Q, thus the variance proxy is the squared expected time required to hit the diagonal when the least favorite initialization is used. Additionally, when the transition kernel is contracting, $\delta(P) < 1$, we can apply Lemma 3 and further replace $\|W_{bc}\|_\infty$ by $1/(1-\delta(P))$, which results in a specialized version of Theorem 1.2 in [53].

2.5 Applications to Markovian multiarmed bandits

2.5.1 Setup

There are $K \geq 2$ arms, and each arm $a \in [K] = \{1, \ldots, K\}$ is associated with a parameter $\theta_a \in \mathbb{R}$ which uniquely encodes[1] an irreducible transition probability matrix P_{θ_a}. We will denote the overall parameter configuration of all K arms with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \mathbb{R}^K$. Arm a evolves according to the stationary Markov chain, $\{X_n^a\}_{n \in \mathbb{Z}_{>0}}$, driven by the irreducible transition probability matrix P_{θ_a} which has a unique stationary distribution π_{θ_a}, so that $X_1^a \sim \pi_{\theta_a}$. There is a common reward function $f : S \to [c, d]$ which generates the reward process $\{Y_n^a\}_{n \in \mathbb{Z}_{>0}} = \{f(X_n^a)\}_{n \in \mathbb{Z}_{>0}}$. The reward process, in general, is not going to be a Markov chain, unless f is injective, and it will have more complicated dependencies than the underlying Markov chain. Each time that we select arm a, this arm evolves by one transition and we observe the corresponding sample from the reward process $\{Y_n^a\}_{n \in \mathbb{Z}_{>0}}$, while all the other arms stay rested.

The stationary reward of arm a is $\mu(\theta_a) = \sum_{x \in S} f(x)\pi_{\theta_a}(x)$. Let $\mu^*(\boldsymbol{\theta}) = \max_{a \in [K]} \mu(\theta_a)$ be the maximum stationary mean, and for simplicity assume that there exists a unique arm, $a^*(\boldsymbol{\theta})$, attaining this maximum stationary mean, i.e. $\{a^*(\boldsymbol{\theta})\} = \arg\max_{a \in [K]} \mu(\theta_a)$. In the following sections we will consider two objectives: identifying an ϵ best arm with some fixed confidence level δ using as few samples as possible, and minimizing the expected regret given some fixed time horizon T.

2.5.2 Approximate Best Arm Identification

In the approximate best arm identification problem, we are given an approximation accuracy $\epsilon > 0$, and a confidence level $\delta \in (0, 1)$. Our goal is to come up with an adaptive algorithm $\mathcal{A}$ which collects a total of N samples, and returns an arm $\hat{a}$ that is within ϵ from the best arm, $a^*(\boldsymbol{\theta})$, with probability at least $1 - \delta$, i.e.

$$\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}}(\mu^*(\boldsymbol{\theta}) \geq \mu(\theta_{\hat{a}}) + \epsilon) \leq \delta.$$

Such an algorithm is called (ϵ, δ)-PAC (probably approximately correct).

In [62] a lower bound for the sample complexity of any (ϵ, δ)-PAC algorithm is derived. The lower bound states that no matter the (ϵ, δ)-PAC algorithm $\mathcal{A}$, there exists an instance $\boldsymbol{\theta}$ such that the sample complexity is at least,

$$\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}}[N] = \Omega\left(\frac{K}{\epsilon^2} \log \frac{1}{\delta}\right).$$

A matching upper bound is provided for i.i.d. bandits in [33] in the form of the median elimination algorithm. We demonstrate the usefulness of our Hoeffding inequality, by providing an analysis of the median elimination algorithm in the more general setting of Markovian bandits.

[1] $\mathbb{R}$ and the set of $|S| \times |S|$ irreducible transition probability matrices have the same cardinality, and hence there is a bijection between them.

Algorithm 1: The β-Median-Elimination algorithm.

Parameters: *number of arms* $K \geq 2$, *approximation accuracy* $\epsilon > 0$, *confidence level* $\delta \in (0,1)$, *parameter* β;

$r = 1,\ A_r = [K],\ \epsilon_r = \epsilon/4,\ \delta_r = \delta/2$;

while $|A_r| \geq 2$ **do**

$N_r = \left\lceil \frac{4\beta}{\epsilon_r^2} \log \frac{3}{\delta_r} \right\rceil$;

Sample each arm in A_r for N_r times;

For $a \in A_r$ calculate $\bar{Y}_a[r] = \frac{1}{N_r} \sum_{n=1}^{N_r} Y_n^a$;

$m_r = \textbf{median}\left((\bar{Y}_a[r])_{a \in A_r}\right)$;

Pick A_{r+1} such that:

- $A_{r+1} \subseteq \{a \in A_r : \bar{Y}_a[r] \geq m_r\}$;
- $|A_{r+1}| = \lfloor |A_r|/2 \rfloor$;

$\epsilon_{r+1} = 3\epsilon_r/4,\ \delta_{r+1} = \delta_r/2,\ r = r+1$;

end

return $\hat{a}$, *where* $A_r = \{\hat{a}\}$;

Theorem 8. *If* $\beta \geq \frac{1}{2}(d-c)^2 \max_{a \in [K]} \mathrm{HitT}(P_{\theta_a})^2$ *then, the* β*-Median-Elimination algorithm is* (ϵ, δ)*-PAC, and its sample complexity is upper bounded by* $O\left(\frac{K}{\epsilon^2} \log \frac{1}{\delta}\right)$.

Proof. The total number of sampling rounds is at most $\lceil \log_2 K \rceil$, and we can set them equal to $\lceil \log_2 K \rceil$ by setting $A_r = \{\hat{a}\}$, for $r \geq R_0$, where $A_{R_0} = \{\hat{a}\}$. Fix $r \in \{1, \ldots, \lceil \log_2 K \rceil\}$. We claim that,

$$\mathbb{P}_{\boldsymbol{\theta}}^{\beta-\mathrm{ME}} \left(\max_{a \in A_r} \mu(\theta_a) \geq \max_{a \in A_{r+1}} \mu(\theta_a) + \epsilon_r \right) \leq \delta_r. \tag{2.22}$$

We condition on the value of A_r. If $|A_r| = 1$, then the claim is trivially true, so we only consider the case $|A_r| \geq 2$. Let $\mu_r^* = \max_{a \in A_r} \mu(\theta_a)$, and $a_r^* \in \arg\max_{a \in A_r : \mu(\theta_a) = \mu_r^*} \bar{Y}_a[r]$. We consider the following set of bad arms,

$$B_r = \{b \in A_r : \bar{Y}_b[r] \geq \bar{Y}_{a_r^*}[r],\ \mu_r^* \geq \mu(\theta_b) + \epsilon_r\},$$

and observe that,

$$\mathbb{P}_{\boldsymbol{\theta}}^{\beta-\mathrm{ME}} \left(\mu_r^* \geq \mu_{r+1}^* + \epsilon_r \right) \leq \mathbb{P}_{\boldsymbol{\theta}}^{\beta-\mathrm{ME}}(|B_r| \geq |A_r|/2). \tag{2.23}$$

In order to upper bound the latter fix $b \in A_r$ and write,

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\theta}}^{\beta-\mathrm{ME}} \left(\bar{Y}_b[r] \geq \bar{Y}_{a_r^*}[r], \mu_r^* \geq \mu(\theta_b) + \epsilon_r \middle| \bar{Y}_{a_r^*}[r] > \mu_r^* - \epsilon_r/2 \right) \\ &\quad \leq \mathbb{P}_{\theta_b}(\bar{Y}_b[r] \geq \mu(\theta_b) + \epsilon_r/2) \leq \delta_r/3, \end{aligned}$$

where in the last inequality we used Corollary 1. Now via Markov's inequality this yields,

$$\mathbb{P}_{\boldsymbol{\theta}}^{\beta-\mathrm{ME}}\left(|B_r| \geq |A_r|/2 \middle| \bar{Y}_{a_r^*}[r] > \mu_r^* - \epsilon_r/2\right) \leq 2\delta_r/3. \tag{2.24}$$

Furthermore, Corollary 1 gives that for any $a \in A_r$,

$$\mathbb{P}_{\theta_a}(\bar{Y}_a[r] \leq \mu(\theta_a) - \epsilon_r/2) \leq \delta_r/3. \tag{2.25}$$

We obtain (2.22) by using (2.24) and (2.25) in (2.23).

With (2.22) in our possession, the fact that median elimination is (ϵ, δ)-PAC follows through a union bound,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}^{\beta-\mathrm{ME}}(\mu^*(\boldsymbol{\theta}) \geq \mu(\theta_{\hat{a}}) + \epsilon) &\leq \mathbb{P}_{\boldsymbol{\theta}}^{\beta-\mathrm{ME}}\left(\bigcup_{r=1}^{\lceil \log_2 K \rceil} \left\{\mu_r^* \geq \mu_{r+1}^* + \epsilon_r\right\}\right) \\ &\leq \sum_{r=1}^{\infty} \delta_r \leq \delta.\end{aligned}$$

Regarding the sample complexity, we have that the total number of samples is at most,

$$\begin{aligned}K \sum_{r=1}^{\lceil \log_2 K \rceil} N_r/2^{r-1} &\leq 2K + \frac{64\beta K}{\epsilon^2} \sum_{r=1}^{\infty} \left(\frac{8}{9}\right)^{r-1} \log \frac{2^r 3}{\delta} \\ &= O\left(\frac{K}{\epsilon^2} \log \frac{1}{\delta}\right).\end{aligned}$$

□

2.5.3 Regret Minimization

Our device to solve the regret minimization problem is an *adaptive allocation rule*, $\boldsymbol{\phi} = \{\phi_t\}_{t \in \mathbb{Z}_{>0}}$, which is a sequence of random variables where $\phi_t \in [K]$ is the arm that we select at time t. Let $N_a(t) = \sum_{s=1}^{t} I_{\{\phi_s = a\}}$, be the number of times we selected arm a up to time t. Our decision, ϕ_t, at time t is based on the information that we have accumulated so far. More precisely, the event $\{\phi_t = a\}$ is measurable with respect to the σ-field generated by the past decisions $\phi_1, \ldots, \phi_{t-1}$, and the past observations $\{X_n^1\}_{n=1}^{N_1(t-1)}, \ldots, \{X_n^K\}_{n=1}^{N_K(t-1)}$.

Given a time horizon T, and a parameter configuration $\boldsymbol{\theta}$, the expected regret incurred when the adaptive allocation rule $\boldsymbol{\phi}$ is used, is defined as,

$$R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T) = \sum_{b \notin a^*(\boldsymbol{\theta})} \mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}[N_b(T)] \Delta_b(\boldsymbol{\theta}),$$

where $\Delta_b(\boldsymbol{\theta}) = \mu^*(\boldsymbol{\theta}) - \mu(\theta_b)$. Our goal is to come up with an adaptive allocation rule that makes the expected regret as small as possible.

There is a known asymptotic lower bound on how much we can minimize the expected regret. Any adaptive allocation rule that is uniformly good across all parameter configurations should satisfy the following instance specific, asymptotic regret lower bound (see [4] for details),

$$\sum_{b \neq a^*(\boldsymbol{\theta})} \frac{\Delta_b(\boldsymbol{\theta})}{\overline{D}\left(\theta_b \parallel \theta_{a^*(\boldsymbol{\theta})}\right)} \leq \liminf_{T\to\infty} \frac{R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T)}{\log T},$$

where $\overline{D}\left(\theta \parallel \lambda\right)$ is the Kullback-Leibler divergence rate between the Markov chains with transition probability matrices P_θ and P_λ, given by,

$$\overline{D}\left(\theta \parallel \lambda\right) = \sum_{x,y \in S} \log \frac{P_\theta(x,y)}{P_\lambda(x,y)} \pi_\theta(x) P_\theta(x,y).$$

Here we utilize our Theorem 4 to provide a finite-time analysis of the β-UCB adaptive allocation rule for Markovian bandits, which is order optimal. The β-UCB adaptive allocation rule, is a simple and computationally efficient index policy based on upper confidence bounds which was initially proposed in [5] for i.i.d. bandits. It has already been studied in the context of Markovian bandits in [87], but in a more restrictive setting under the further assumptions of aperiodicity and reversibility due to the use of the bounds from [39, 60]. For adaptive allocation rules that asymptotically match the lower bound we refer the interested reader to [4, 72].

Algorithm 2: The β-UCB adaptive allocation rule.

Parameters: *number of arms* $K \geq 2$*, time horizon* $T \geq K$*, parameter* β;
Pull each arm in $[K]$ once;
for $t = K$ *to* $T-1$, **do**

$$\phi_{t+1} \in \arg\max_{a \in [K]} \left\{ \bar{Y}_a(t) + \sqrt{\frac{2\beta \log t}{N_a(t)}} \right\}$$

end

Theorem 9. *If* $\beta > \frac{1}{2}(d-c)^2 \max_{a\in[K]} \mathrm{HitT}(P_{\theta_a})^2$ *then,*

$$R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}_{\beta-\mathrm{UCB}}}(T) \leq 8\beta \left(\sum_{b \neq a^*(\boldsymbol{\theta})} \frac{1}{\Delta_b(\boldsymbol{\theta})} \right) \log T + \frac{\gamma}{\gamma - 2} \sum_{b \neq a^*(\boldsymbol{\theta})} \Delta_b(\boldsymbol{\theta}),$$

where $\gamma = \frac{4\beta}{(d-c)^2 \max_{a\in[K]} \mathrm{HitT}(P_{\theta_a})^2} > 2$.

Proof. Fix $b \neq a^*(\boldsymbol{\theta})$, and observe that,

$$N_b(T) \leq 1 + \frac{8\beta}{\Delta_b(\boldsymbol{\theta})^2} \log T + \sum_{t=2}^{T-1} I_{\left\{\phi_{t+1}=b,\ N_b(t) \geq \frac{8\beta}{\Delta_b(\boldsymbol{\theta})^2} \log T\right\}}.$$

On the event $\left\{\phi_{t+1} = b,\ N_b(t) \geq \frac{8\beta}{\Delta_b(\boldsymbol{\theta})^2}\log T\right\}$, we have that, either $\bar{Y}_b(t) \geq \mu(\theta_b) + \sqrt{\frac{2\beta\log t}{N_b(t)}}$, or $\bar{Y}_{a^*(\boldsymbol{\theta})}(t) \leq \mu^*(\boldsymbol{\theta}) - \sqrt{\frac{2\beta\log t}{N_{a^*(\boldsymbol{\theta})}(t)}}$, since otherwise the β-UCB index of $a^*(\boldsymbol{\theta})$ is larger than the β-UCB index of b which contradicts the assumption that $\phi_{t+1} = b$.

In addition, using Corollary 1, we obtain,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}_{\beta-\mathrm{UCB}}}\left(\bar{Y}_b(t) \geq \mu(\theta_b) + \sqrt{\frac{2\beta\log t}{N_b(t)}}\right) \\
&= \sum_{n=1}^{t} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}_{\beta-\mathrm{UCB}}}\left(\bar{Y}_b(t) \geq \mu(\theta_b) + \sqrt{\frac{2\beta\log t}{N_b(t)}}, N_b(t) = n\right) \\
&\leq \sum_{n=1}^{t} \mathbb{P}_{\theta_b}\left(\frac{1}{n}\sum_{k=1}^{n} Y_k^b \geq \mu(\theta_b) + \sqrt{\frac{2\beta\log t}{n}}\right) \\
&\leq \sum_{n=1}^{t} \frac{1}{t^\gamma} = \frac{1}{t^{\gamma-1}}.
\end{aligned}$$

Similarly we can see that,

$$\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}_{\beta-\mathrm{UCB}}}\left(\bar{Y}_{a^*(\boldsymbol{\theta})}(t) \leq \mu^*(\boldsymbol{\theta}) - \sqrt{\frac{2\beta\log t}{N_{a^*(\boldsymbol{\theta})}(t)}}\right) \leq \frac{1}{t^{\gamma-1}}.$$

The conclusion now follows by putting everything together and using the integral estimate,

$$\sum_{t=2}^{T-1} \frac{1}{t^{\gamma-1}} \leq \int_1^\infty \frac{1}{t^{\gamma-1}} dt = \frac{1}{\gamma-2}.$$

□

Chapter 3

Finite sample large deviations for Markov chains

3.1 Introduction

Let S be a finite set and $(X_k)_{k\in\mathbb{Z}_{\geq 0}}$ the coordinate process on $S^{\mathbb{Z}_{\geq 0}}$. Given an initial distribution q on S, and a stochastic matrix P, there exists a unique probability measure $\mathbb{P}_q$ on the sequence space such that the coordinate process $(X_k)_{k\in\mathbb{Z}_{\geq 0}}$ is a Markov chain with transition probability matrix P, with respect to the filtration of σ-fields $(\mathcal{F}_n := \sigma(X_0, \ldots, X_n), n \geq 0)$. If we assume further that P is irreducible, then there exists a unique stationary distribution π for the transition probability matrix P, and for any real-valued function $f : S \to \mathbb{R}$ the empirical mean $n^{-1}\sum_{k=1}^{n} f(X_k)$ converges $\mathbb{P}_q$-almost-surely to the stationary mean $\pi(f) := \sum_x f(x)\pi(x)$. The goal of this chapter is to quantify the rate of this convergence by developing finite sample upper bounds for the large deviations probability

$$\mathbb{P}_q\left(\frac{1}{n}\sum_{k=1}^{n} f(X_k) \geq \mu\right), \text{ for } \mu \geq \pi(f).$$

The significance of studying finite sample bounds for such tail probabilities is not only theoretical but also practical, since concentration inequalities for Markov dependent random variables have wide applicability in statistics, computer science and learning theory. Just to mention a few applications, first and foremost this convergence forms the backbone behind all Markov chain Monte Carlo (MCMC) integration techniques, see [67]. Moreover, tail bounds of this form have been used by [47] to develop an approximation algorithm for the permanent of a nonnegative matrix. In addition, in the stochastic multi-armed bandit literature the analysis of learning algorithms is based on tail bounds of this type, see the survey of [15]. More specifically the work of [73] uses such a bound to tackle a Markovian identification problem.

The classic large deviations theory for Markov chains due to [69, 27, 38, 31, 24] suggests that asymptotically the large deviations probability decays exponentially and the rate is

given by the convex conjugate $\Lambda^*(\mu)$ of the log-Perron-Frobenius eigenvalue $\Lambda(\theta)$ of the nonnegative irreducible matrix $\tilde{P}_\theta(x,y) := P(x,y)e^{\theta f(y)}$. In particular

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_q \left(\frac{1}{n} \sum_{k=1}^{n} f(X_k) \geq \mu \right) = -\Lambda^*(\mu), \text{ for } \mu \geq \pi(f).$$

Our objective is to develop a finite sample bound which captures this exponential decay and has a constant prefactor that does not depend on μ, and is thus useful in applications. A counting based approach by [23] is able to capture this exponential decay but with a suboptimal prefactor that depends polynomially on n. Through the development in the book of [24] (Theorem 3.1.2), which is also presented by [92], one is able to obtain a constant prefactor, which though depends on μ. This is unsatisfactory because exact large deviations for Markov chains, see [69, 55], yield that, at least when the supremum $\sup_{\theta\in\mathbb{R}}\{\theta\mu - \Lambda(\theta)\} = \Lambda^*(\mu)$ is attained at θ_μ, then

$$\mathbb{P}_q \left(\frac{1}{n} \sum_{k=1}^{n} f(X_k) \geq \mu \right) \sim \frac{\mathbb{E}_{X\sim q}[v_{\theta_\mu}(X)]}{\theta_\mu \sqrt{2\pi n \sigma^2_{\theta_\mu}}} e^{-n\Lambda^*(\mu)}, \text{ as } n \to \infty,$$

where $\sigma^2_{\theta_\mu} = \Lambda''(\theta_\mu)$ and v_{θ_μ} is a right Perron-Frobenius eigenvector of $\tilde{P}_{\theta_\mu}$. Here $\sim$ denotes that the ratio of the expressions on the left hand side and the right hand side converges to 1, and $\Lambda''(\theta_\mu)$ denotes the second derivative in θ of $\Lambda(\theta)$ at $\theta = \theta_\mu$. Thus, if we allow dependence on μ, then the prefactor should be able to capture a decay of the order $1/\sqrt{n}$. If we insist on no dependence on μ though, the best that we can hope for is a constant prefactor, because otherwise we will contradict the central limit theorem for Markov chains.

In Section 3.4 we establish a tail bound with the optimal rate of exponential decay and a constant prefactor which depends only on the function f and the stochastic matrix P, under the conditions of Section 3.3. The key technique to derive our Chernoff type bound is the old idea due to [32] of an exponential tilt, which lies at the heart of large deviations theory. In the world of statistics those exponential changes of measure go by the name exponential families and the standard reference is the book of [14]. Exponential tilts of stochastic matrices generalize those of finitely supported probability distributions, and were first introduced in the work of [69]. Subsequently they formed one of the main tools in the study of large deviations for Markov chains, see [27, 38, 31, 24, 9, 55]. Naturally they are also the key object when one conditions on the pair empirical distribution of a Markov chain and considers conditional limit theorems, as in [22, 13]. A more recent development by [75] gives an information geometry perspective to this concept, while [43] examine the problem of parameter estimation for exponential families of stochastic matrices. Here we build on exponential families of stochastic matrices and by studying the analyticity properties of the Perron-Frobenius eigenvalue and its associated eigenvector as we parametrically move the mean of f under exponential tilts, together with conjugate duality, we are able to establish our main Chernoff type bound.

In addition to that, in Section 3.5 we use an exponential martingale, coming from the exponential family of stochastic matrices, in order to derive a maximal inequality for Markov chains. In the literature there are several approaches that use martingale techniques either to derive Hoeffding inequalities for Markov chains [40, 70], or more generally to study concentration of measure for Markov chains [63, 64, 66, 83, 65, 17, 53, 77, 52]. Nonetheless, they're all based either on Dynkin's martingale or on Doob's martingale, combined with coupling ideas, and there is no evidence that they can lead to maximal inequalities. This maximal inequality constitutes a finite-sample version of the law of the iterated logarithm for Markov chains.

Finally, in Section 3.6 we establish a uniform multiplicative ergodic theorem. The classic linear ergodic theory for Markov chains, [19] suggests that

$$\frac{1}{n}\,\mathbb{E}_q\left[\sum_{k=1}^{n} f(X_k)\right] \to \pi(f),\ \text{as } n \to \infty.$$

[9] and [55] have proved a multiplicative version of this under appropriate assumptions, which state that the *scaled log-moment-generating-function* $\Lambda_n(\theta)$ converges pointwise to the log-Perron-Frobenius eigenvalue

$$\Lambda_n(\theta) \to \Lambda(\theta),\ \text{as } n \to \infty,\ \text{for any } \theta \in \mathbb{R},$$

where

$$\Lambda_n(\theta) := \frac{1}{n}\log \mathbb{E}_q\left[\exp\left\{\theta \sum_{k=1}^{n} f(X_k)\right\}\right].$$

For our class of finite Markov chains we are able to establish a uniform multiplicative ergodic theorem in the terminology of [9].

3.2 Exponential Family of Stochastic Matrices

3.2.1 Construction

Exponential tilting of stochastic matrices originates in the work of [69]. Following this, we define an exponential family of stochastic matrices which is able to produce Markov chains with shifted stationary means. The generator of the exponential family is an irreducible stochastic matrix P, and $\theta \in \mathbb{R}$ represents the canonical parameter of the family. Then we define

$$\tilde{P}_\theta(x, y) := P(x, y)e^{\theta f(y)}, \tag{3.1}$$

(or $\widetilde{(P)}_\theta(x, y)$, where $\widetilde{(\cdot)}_\theta$ is thought as an operator over matrices). $\tilde{P}_\theta$ has the same nonnegativity structure as P, hence it is irreducible and we can use the Perron-Frobenius theory in order to normalize it and turn it into a stochastic matrix. Let $\rho(\theta)$ (or $\rho(\tilde{P}_\theta)$) be the spectral radius of $\tilde{P}_\theta$, which from the Perron-Frobenius theory 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 both have all entries strictly 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 of [45]. Using $\tilde{P}_\theta$ we define a family of nonnegative irreducible matrices, parametrized by θ, in the following way

$$(P)_\theta(x,y) = P_\theta(x,y) := \frac{\tilde{P}_\theta(x,y)v_\theta(y)}{\rho(\theta)v_\theta(x)}, \tag{3.2}$$

which are stochastic, since

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

In addition the stationary distributions of the P_θ are given by

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

since

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

Note that the generator stochastic matrix, P, is the member of the family that corresponds to $\theta = 0$, i.e. $P_0 = P,\ \rho(0) = 1,\ u_0 = \pi,\ v_0 = \mathbf{1}$, and $\pi_0 = \pi$, where $\mathbf{1}$ is the all ones vector. In general it is possible that the family is degenerate as the following example suggests.

Example 6. Let $S = \{\pm 1\},\ P(x,y) = 1\{x \neq y\}$, and $f(x) = x$. Then $\rho(\theta) = 1,\ v_\theta(-1) = \frac{1+e^\theta}{2},\ v_\theta(1) = \frac{1+e^{-\theta}}{2}$, and $P_\theta = P$ for any $\theta \in \mathbb{R}$.

A basic property of the exponential family P_θ is that the composition of $(\cdot)_{\theta_1}$ with $(\cdot)_{\theta_2}$, is the transform $(\cdot)_{\theta_1+\theta_2}$, and so composition is commutative. Furthermore we can undo the transform $(\cdot)_\theta$ by applying $(\cdot)_{-\theta}$. We state this formally for convenience.

Lemma 4. *For any irreducible stochastic matrix P, and any $\theta_1, \theta_2 \in \mathbb{R}$*

$$\left((P)_{\theta_2}\right)_{\theta_1} = (P)_{\theta_1+\theta_2}.$$

Proof. It suffices to check that $\left(\frac{v_{\theta_1+\theta_2}(y)}{v_{\theta_2}(y)},\ y \in S\right)$ is a right eigenvector of the matrix with entries $\left(\frac{P(x,y)e^{\theta_2 f(y)}v_{\theta_2}(y)}{\rho(\theta)v_{\theta_2}(x)}\right)e^{\theta_1 f(y)}$, with the corresponding eigenvalue being $\frac{\rho(\theta_1+\theta_2)}{\rho(\theta_2)}$. This is a straightforward calculation. □

3.2.2 Mean Parametrization

The exponential family P_θ defined in (3.2) can be reparametrized using the mean parameters $\mu = \pi_\theta(f)$. The duality between the canonical parameters θ and the mean parameters μ is manifested through the log-Perron-Frobenius eigenvalue $\Lambda(\theta) := \log \rho(\theta)$. More specifically, from Lemma 5 it follows that there are two cases for the mapping $\theta \mapsto P_\theta$. In the nondegenerate case that this mapping is nonconstant, $\Lambda'(\theta)$ is a strictly increasing bijection between the set $\mathbb{R}$ of canonical parameters and the set

$$\mathcal{M} := \{\mu \in \mathbb{R} : \pi_\theta(f) = \mu, \text{ for some } \theta \in \mathbb{R}\} \tag{3.3}$$

of mean parameters, which is an open interval. Therefore, with some abuse of notation, for any $\mu \in \mathcal{M}$ we may write $u_\mu, v_\mu, P_\mu, \pi_\mu$ for $u_{\Lambda'^{-1}(\mu)}, v_{\Lambda'^{-1}(\mu)}, P_{\Lambda'^{-1}(\mu)}, \pi_{\Lambda'^{-1}(\mu)}$. In the degenerate case that the mapping is constant, $\Lambda'(\theta) = \pi(f)$, and the set $\mathcal{M}$ is the singleton $\{\pi(f)\}$. An illustration of the degenerate case is Example 6.

Lemma 5. *Let P be an irreducible stochastic matrix, and $f : S \to \mathbb{R}$ a real-valued function on the state space S. Then*

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

(b) $\Lambda'(\theta) = \pi_\theta(f)$.

(c) $\Lambda''(\theta) = \operatorname{var}_{(X,Y)\sim\pi_\theta\odot P_\theta}\left(f(Y) + \frac{v_\theta(X)}{v_\theta(Y)}\frac{d}{d\theta}\frac{v_\theta(Y)}{v_\theta(X)}\right)$, where $\pi_\theta \odot P_\theta$ denotes the bivariate distribution defined by $(\pi_\theta \odot P_\theta)(x, y) := \pi_\theta(x)P_\theta(x, y)$.

(d) Either $P_\theta = P_0 = P$ for all $\theta \in \mathbb{R}$ (degenerate case), or $\theta \mapsto P_\theta$ is an injection (nondegenerate case).

Moreover, in the degenerate case $\Lambda(\theta) = \pi_0(f)\theta$ is linear, while in the nondegenerate case $\Lambda(\theta)$ is strictly convex.

The proof of Lemma 5 can be found in Section 3.B.

3.2.3 Relative Entropy Rate and Conjugate Duality

For two probability distributions $\mathbb{Q}$ and $\mathbb{P}$ over the same measurable space we define the *relative entropy* between $\mathbb{Q}$ and $\mathbb{P}$ as

$$D(\mathbb{Q} \parallel \mathbb{P}) := \begin{cases} \mathbb{E}_\mathbb{Q}\left[\log \frac{d\mathbb{Q}}{d\mathbb{P}}\right], & \text{if } Q \text{ is absolutely continuous with respect to } P, \\ \infty, & \text{otherwise.} \end{cases}$$

Relative entropies of stochastic processes are most of the time trivial, and so we resort to the notion of relative entropy rate. Let Q, P be two stochastic matrices over the same state space S. We further assume that Q is irreducible with associated stationary distribution π_Q.

For any initial distribution q on S we define the *relative entropy rate* between the Markov chain $\mathbb{Q}_q$ induced by Q with initial distribution q, and the Markov chain $\mathbb{P}_q$ induced by P with initial distribution q as

$$D\left(Q \parallel P\right) := \lim_{n\to\infty} \frac{1}{n} D\left(\mathbb{Q}_q \mid_{\mathcal{F}_n} \parallel \mathbb{P}_q \mid_{\mathcal{F}_n}\right),$$

where $\mathbb{Q}_q \mid_{\mathcal{F}_n}$ and $\mathbb{P}_q \mid_{\mathcal{F}_n}$ denote the finite dimensional distributions of the probability measures restricted to the sigma algebra $\mathcal{F}_n$. Note that the definition is independent of the initial distribution q, since we can easily see using ergodic theory that

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

where $\pi_Q \odot Q$ denotes the bivariate 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 $\alpha > 0$, and $0 \log 0 = 0 \log \frac{0}{0} = 0$.

For stochastic matrices which are elements of the exponential family P_θ defined in (3.2) we simplify the relative entropy rate notation as follows. For $\theta_1, \theta_2 \in \mathbb{R}$ and $\mu_1 = \Lambda'(\theta_1)$, $\mu_2 = \Lambda'(\theta_2)$ we write

$$D\left(\theta_1 \parallel \theta_2\right), D\left(\mu_1 \parallel \mu_2\right) := D\left(\pi_{\theta_1} \odot P_{\theta_1} \parallel \pi_{\theta_1} \odot P_{\theta_2}\right).$$

For those relative entropy rates Lemma 6 suggests an alternative representation based on the parametrization. Its proof can be found in Section 3.B.

Lemma 6. *Let $\theta_1, \theta_2 \in \mathbb{R}$ and $\mu_1 = \Lambda'(\theta_1)$, $\mu_2 = \Lambda'(\theta_2)$. Then*

$$D\left(\theta_1 \parallel \theta_2\right) = \Lambda(\theta_2) - \Lambda(\theta_1) - \mu_1(\theta_2 - \theta_1).$$

We further define the convex conjugate of $\Lambda(\theta)$ as $\Lambda^*(\mu) := \sup_{\theta\in\mathbb{R}} \{\theta\mu - \Lambda(\theta)\}$. Moreover, since we saw in Lemma 5 that $\Lambda(\theta)$ is convex and analytic, we have that the biconjugate of $\Lambda(\theta)$ is $\Lambda(\theta)$ itself, i.e. $\Lambda(\theta) = \sup_{\mu\in\mathbb{R}} \{\mu\theta - \Lambda^*(\mu)\}$. The convex conjugate $\Lambda^*(\mu)$ represents the rate of exponential decay for large deviation events, and in the following Lemma 7, which is established in Section 3.B, we derive a closed form expression for it.

Lemma 7.

$$\Lambda^*(\mu) = \begin{cases} D\left(\mu \parallel \pi(f)\right), & \text{if } \mu \in \mathcal{M}, \\ \lim\limits_{\hat{\mu}\to\mu} D\left(\hat{\mu} \parallel \pi(f)\right), & \text{if } \mu \in \partial\mathcal{M}, \\ \infty, & \text{otherwise}, \end{cases}$$

where $\mathcal{M}$ is defined in (3.3).

An inspection of how the supremum was obtained in the previous Lemma 7 yields the following Corollary 2.

Corollary 2.

$$\Lambda^*(\mu) = \begin{cases} \sup_{\theta \geq 0} \{\theta\mu - \Lambda(\theta)\}, & \text{if } \mu \geq \pi(f), \\ \sup_{\theta \leq 0} \{\theta\mu - \Lambda(\theta)\}, & \text{if } \mu \leq \pi(f). \end{cases}$$

3.3 Conditions for the asymptotic positivity of the Perron-Frobenius eigenvector

In this section we would like to study under what conditions the Perron-Frobenius eigenvector v_θ remains strictly positive asymptotically as $\theta \to \pm\infty$. Recall that v_θ is the right Perron-Frobenius eigenvector of the matrix

$$\tilde{P}_\theta(x, y) = P(x, y)e^{\theta f(y)},$$

but since eigenvectors are invariant under scaling it will be more convenient to consider the matrix

$$\overline{P}_\theta(x, y) = P(x, y)e^{-\theta(b - f(y))},$$

where for the state reward function $f : S \to \mathbb{R}$, we define $b = \max_{x \in S} f(x)$, and $S_b = \arg\max_{x \in S} f(x)$. Additionally, since we only care about positivity, we will assume that $\sum_x v_\theta(x) = 1$.

By permuting rows and corresponding columns of P write

$$P = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right], \quad \overline{P}_\theta = \left[\begin{array}{c|c} A & B_\theta \\ \hline C & D_\theta \end{array}\right],$$

where A represents the transitions from S_b to S_b, B represents the transitions from S_b to $S - S_b$, C represents the transitions from $S - S_b$ to S_b, and D represents the transitions from $S - S_b$ to $S - S_b$. Additionally, $\lim_{\theta\to\infty} B_\theta = 0$, and $\lim_{\theta\to\infty} D_\theta = 0$, so by continuity

$$\lim_{\theta\to\infty} \rho(\overline{P}_\theta) = \rho(A).$$

Using standard terminology [10] we say that $x \in S_b$ has access to $y \in S_b$ if, $A^n(x, y) > 0$ for some $n \in \mathbb{Z}_{\geq 0}$. This way the nonnegative matrix A induces a partition of S_b in communicating classes. We call the class i final if it has access to no other class, and we call it basic if $\rho(A_i) = \rho(A)$, where A_i contains only the rows and columns of A that correspond to transitions within class i.

Let $1, \ldots, M$ be an enumeration of the final classes of A, and $M+1, \ldots, N$ an enumeration of the non-final classes of A. Write A_i for the transitions inside class i, and $A_{i,j}$ for the

transitions from class i to class j. By permuting rows and corresponding columns write

$$\left[\begin{array}{ccccccc|c} A_1 & 0 & & & & 0 & B_\theta^1 \\ 0 & A_2 & \ddots & & & & B_\theta^2 \\ \vdots & \vdots & \ddots & 0 & & & \vdots \\ 0 & 0 & \cdots & A_M & 0 & & B_\theta^M \\ A_{M+1,1} & A_{M+1,2} & \cdots & \cdots & A_{M+1} & \ddots & B_\theta^{M+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 & \vdots \\ A_{N,1} & A_{N,2} & \cdots & \cdots & \cdots & \cdots & A_N & B_\theta^N \\ \hline & & & C & & & & D_\theta \end{array}\right] \cdot \begin{bmatrix} v_\theta^1 \\ v_\theta^2 \\ \vdots \\ v_\theta^M \\ v_\theta^{M+1} \\ \vdots \\ v_\theta^N \\ v_\theta^{N+1} \end{bmatrix} = \rho(\overline{P}_\theta) \begin{bmatrix} v_\theta^1 \\ v_\theta^2 \\ \vdots \\ v_\theta^M \\ v_\theta^{M+1} \\ \vdots \\ v_\theta^N \\ v_\theta^{N+1} \end{bmatrix}. \quad (3.4)$$

We are now ready provide necessary conditions for the asymptotic positivity of v_θ as $\theta \to \infty$.

Lemma 8. *If* $\underline{\lim}_{\theta\to\infty} v_\theta > 0$, *then*

1. *All the rows of A and C are nonzero.*
2. *All the final classes of A are basic.*
3. *All the basic classes of A are final.*

Proof.

1. Assume, towards contradiction, that there exists $x_0 \in S$ such that $P(x_0, y) = 0$ for all $y \in S_b$. Then
$$\lim_{\theta\to\infty} \overline{P}_\theta(x_0, y) = 0, \text{ for all } y \in S,$$
and because $P_\theta(x_0, \cdot)$ is a probability mass function, there exists $y_0 \in S$ such that
$$\overline{\lim_{\theta\to\infty}} \frac{v_\theta(y_0)}{v_\theta(x_0)\rho(\overline{P}_\theta)} = \infty.$$
If $\rho(A) > 0$, then we already have a contradiction. Alternatively, assume that $\rho(A) = 0$, and pick $y_1 \in S_b$. Because P is irreducible there exists $x_1 \in S$ such that
$$\overline{P}_\theta(x_1, y_1) = P(x_1, y_1) > 0,$$
and because $P(x_1, \cdot)$ is a probability mass function
$$\overline{\lim_{\theta\to\infty}} \frac{v_\theta(y_1)}{v_\theta(x_1)\rho(\overline{P}_\theta)} < \infty,$$
which gives a contradiction.

2. From the previous part we already have that all the rows of A are non-zero and so $\rho(A) > 0$. For every final class $i = 1, \ldots, M$

$$A_i \varliminf_{\theta\to\infty} v_\theta^i = \rho(A) \varliminf_{\theta\to\infty} v_\theta^i,$$

and since $\varliminf_{\theta\to\infty} v_\theta^i > 0$, it follows that $\rho(A_i) = \rho(A)$, and hence i is a basic class.

3. Let $j = M+1, \ldots, N$ be a non-final class. Then

$$A_{j,1} \varliminf_{\theta\to\infty} v_\theta^1 + \cdots + A_{j,j-1} \varliminf_{\theta\to\infty} v_\theta^{j-1} + A_j \varliminf_{\theta\to\infty} v_\theta^j = \rho(A) \varliminf_{\theta\to\infty} v_\theta^j.$$

From this we obtain that

$$\max_x \frac{(A_j \varliminf_{\theta\to\infty} v_\theta^j)(x)}{\varliminf_{\theta\to\infty} v_\theta^j(x)} \leq \rho(A).$$

Additionally a standard upper bound on $\rho(A_j)$ (Theorem 8.1.26. in [45]) is

$$\rho(A_j) \leq \max_x \frac{(A_j \varliminf_{\theta\to\infty} v_\theta^j)(x)}{\varliminf_{\theta\to\infty} v_\theta^j(x)},$$

and so if j is a basic class, then

$$\rho(A_j) = \max_x \frac{(A_j \varliminf_{\theta\to\infty} v_\theta^j)(x)}{\varliminf_{\theta\to\infty} v_\theta^j(x)} = \rho(A).$$

We claim that $\varliminf_{\theta\to\infty} v_\theta^j$ is a right eigenvector of A_j corresponding to $\rho(A_j)$. Assume, towards contradiction, that there exists x_0 with

$$(A_j \varliminf_{\theta\to\infty} v_\theta^j)(x_0) < \rho(A_j) \varliminf_{\theta\to\infty} v_\theta^j(x_0).$$

Let u^j be a positive left eigenvector of A_j with corresponding eigenvalue $\rho(A_j)$. Then

$$\rho(A_j) \sum_x u^j(x) \varliminf_{\theta\to\infty} v_\theta^j(x) = \sum_x u^j(x)(A_j \varliminf_{\theta\to\infty} v_\theta^j)(x) < \rho(A_j) \sum_x u^j(x) \varliminf_{\theta\to\infty} v_\theta^j(x),$$

which is a contradiction, and so $\varliminf_{\theta\to\infty} v_\theta^j$ is a right eigenvector of A_j corresponding to $\rho(A_j)$. But this contradicts the fact that

$$\min_x \frac{(A_j \varliminf_{\theta\to\infty} v_\theta^j)(x)}{\varliminf_{\theta\to\infty} v_\theta^j(x)} < \rho(A),$$

since j is a non-final class and so at least one of the entries of $A_{j,1}, \ldots, A_{j,j-1}$ is positive.

□

Now we complement the necessary conditions from Lemma 8 with a set of sufficient conditions.

Lemma 9. *If*

1. *All the rows of A and C are nonzero.*
2. *All the final classes of A are basic.*
3. *All the basic classes of A are final.*
4. *A has exactly one final class.*

Then $\underline{\lim}_{\theta\to\infty} v_\theta > 0$.

Proof.
From (3.4) we get the following equations

- For a final class $i = 1, \ldots, M$

$$(\rho(\overline{P}_\theta)I - A_i)v_\theta^i = B_\theta^i v_\theta^{N+1}. \tag{3.5}$$

At the limit

$$(\rho(A)I - A_i)\ \underline{\lim}_{\theta\to\infty}\ v_\theta^i = 0, \tag{3.6}$$

and since i is a basic class, $\underline{\lim}_{\theta\to\infty} v_\theta^i$ is either positive or equal to zero.

- For a non-final class $j = M+1, \ldots, N$

$$v_\theta^j = (\rho(\overline{P}_\theta)I - A_j)^{-1}\left(\sum_{k=1}^{j-1} A_{j,k}v_\theta^k + B_\theta^j v_\theta^{N+1}\right), \tag{3.7}$$

since j is not a basic class, and so $\rho(A_j) < \rho(A) \leq \rho(\overline{P}_\theta)$. Additionally, we note that $(\rho(\overline{P}_\theta)I - A_j)^{-1}$, and $(\rho(A)I - A_j)^{-1}$ are positive matrices (Problem 8.3.P12 in [45]).

At the limit

$$\underline{\lim}_{\theta\to\infty}\ v_\theta^j = (\rho(A)I - A_j)^{-1}\left(\sum_{k=1}^{j-1} A_{j,k}\ \underline{\lim}_{\theta\to\infty}\ v_\theta^k\right), \tag{3.8}$$

and if $\underline{\lim}_{\theta\to\infty} v_\theta^i > 0$ for each final class $i = 1, \ldots, M$, then we see inductively that $\underline{\lim}_{\theta\to\infty} v_\theta^j > 0$ for each non-final class $j = M+1, \ldots, N$.

- For $S - S_b$, and for θ sufficiently large

$$v_\theta^{N+1} = (\rho(\overline{P}_\theta)I - D_\theta)^{-1}C \begin{bmatrix} v_\theta^1 \\ \vdots \\ v_\theta^N \end{bmatrix}. \tag{3.9}$$

  At the limit

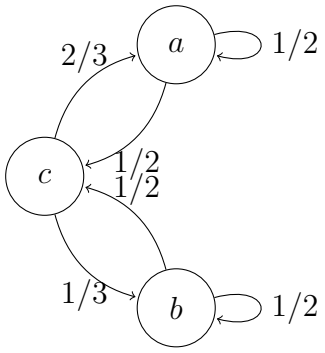
$$\underline{\lim}_{\theta\to\infty} v_\theta^{N+1} = \rho(A)^{-1}C \begin{bmatrix} \underline{\lim}_{\theta\to\infty} v_\theta^1 \\ \vdots \\ \underline{\lim}_{\theta\to\infty} v_\theta^N \end{bmatrix}, \tag{3.10}$$

  and because each row of C is nonzero, we see from the previous bullet that $\underline{\lim}_{\theta\to\infty} v_\theta^i > 0$ for each final class $i = 1, \ldots, M$, implies that $\underline{\lim}_{\theta\to\infty} v_\theta^{N+1} > 0$.

Now if there is just one final class ($M = 1$), then indeed $\underline{\lim}_{\theta\to\infty} v_\theta^1 > 0$ because, $\underline{\lim}_{\theta\to\infty} v_\theta^1 = 0$ implies that $\underline{\lim}_{\theta\to\infty} v_\theta = 0$, and this contradicts the normalization $\sum_{x\in S} v_\theta(x) = 1$. □

Note that the sufficient conditions from Lemma 9 require only that A has exactly one final class in addition to the necessary conditions from Lemma 8. As the following example suggest this requirement is not always necessary.

Example 7. Let P be the transition matrix for the chain



with $f(a) = f(b) = +1$, and $f(c) = -1$. Then

$$\overline{P}_\theta = \left[\begin{array}{cc|c} 1/2 & 0 & e^{-2\theta}/2 \\ 0 & 1/2 & e^{-2\theta}/2 \\ \hline 2/3 & 1/3 & 0 \end{array}\right],$$

with

$$\rho(\overline{P}_\theta) = \frac{1}{4}(1 + \sqrt{1 + 8e^{-4\theta}}) \downarrow \frac{1}{2}, \text{ as } \theta \uparrow \infty,$$

and v_θ (picked so that $v_\theta(a) + v_\theta(b) + v_\theta(c) = 1$) is

$$\frac{1}{2\rho(\overline{P}_\theta) + 1} \begin{bmatrix} \rho(\overline{P}_\theta) \\ \rho(\overline{P}_\theta) \\ 1 \end{bmatrix} \to \begin{bmatrix} 1/4 \\ 1/4 \\ 1/2 \end{bmatrix}, \text{ as } \theta \to \infty.$$

Note also that $1/2 (< \rho(\overline{P}_\theta))$ is another eigenvalue of $\overline{P}_\theta$ with corresponding eigenvector

$$\begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix}.$$

At the limit the two eigenvalues $1/2$ and $\rho(\overline{P}_\infty)$ coincide, and the two eigenvectors span the two dimensional nullspace of $I/2 - \overline{P}_\infty$.

If we further define $a = \min_{x \in S} f(x)$, and $S_a = \arg\min_{x \in S} f(x)$, then the sufficient conditions from Lemma 9 can be relaxed to the following simple conditions on the nonnegativity structure of P.

A 1. The submatrix of P with rows and columns in S_b is irreducible.

A 2. For every $x \in S - S_b$, there exists $y \in S_b$ such that $P(x, y) > 0$.

A 3. The submatrix of P with rows and columns in S_a is irreducible.

A 4. For every $x \in S - S_a$, there exists $y \in S_a$ such that $P(x, y) > 0$.

Lemma 10. *Under A 1, and A 2 we have that* $\varliminf_{\theta \to \infty} v_\theta > 0$, *while under A 3, and A 4 we have that* $\varliminf_{\theta \to -\infty} v_\theta > 0$.

A critical ingredient to obtain our tail bounds is the following Proposition 1 which states that under the assumptions A 1-A 2 the ratio of the entries of the right Perron-Frobenius eigenvector stays uniformly bounded.

Proposition 1. *Let P be an irreducible stochastic matrix on the finite state space S, which, combined with a real-valued function $f : S \to \mathbb{R}$, satisfies A 1-A 2. Then*

$$K_u := \sup_{\theta \in \mathbb{R}_{\geq 0}, x, y \in S} \frac{v_{\overline{P}_\theta}(x)}{v_{\overline{P}_\theta}(y)} < \infty,$$

where $K_u = K_u(P, f)$ is a constant depending on the stochastic matrix P, and the function f. In particular

- *if P induces an IID process, i.e. P has identical rows, then $K_u = 1$;*
- *if P is a positive stochastic matrix, then $K_u \leq \max_{x,y,z} \frac{P(x,z)}{P(y,z)}$.*

Proof. Lemma 5 yields that $\theta \mapsto v_\theta(x)/v_\theta(y)$ is continuous, and so in conjunction with Lemma 10 we have that the ratio of the entries of the right Perron-Frobenius eigenvector is uniformly bounded, hence $K_u < \infty$. □

3.4 Chernoff Bound

In this section we build on exponential families of stochastic matrices and by using the conditions from Lemma 10 that guarantee the positivity of the asymptotic Perron-Frobenius eigenvector we are able to establish our main Chernoff type bound, which we state below together with some remarks.

Theorem 10. *Let P be an irreducible stochastic matrix on the finite state space S, with stationary distribution π, which, combined with a real-valued function $f : S \to \mathbb{R}$, satisfies A 1-A 2. Then, for any initial distribution q, we have*

$$\mathbb{P}_q\left(\frac{1}{n}\sum_{k=1}^{n} f(X_k) \geq \mu\right) \leq K_u e^{-n\Lambda^*(\mu)}, \text{ for } \mu \geq \pi(f),$$

where $K_u = K_u(P, f)$ is the constant from Proposition 1, and depends only on the stochastic matrix P and the function f.

Remark 2. Since f is arbitrary and our assumptions A 1-A 2 and A 3-A 4 are symmetric, we can substitute f with $-f$, so that Theorem 10 yields a Chernoff type bound for the lower tail as well. In particular, assuming A 3-A 4 we have

$$\mathbb{P}_q\left(\frac{1}{n}\sum_{k=1}^{n} f(X_k) \leq \mu\right) \leq K_l e^{-n\Lambda^*(\mu)}, \text{ for } \mu \leq \pi(f),$$

where $K_l = K_u(P, -f)$.

Remark 3. Similarly assuming A 1-A 4 we have the following two-sided Chernoff type bound.

$$\mathbb{P}_q\left(\frac{1}{n}\sum_{k=1}^{n} f(X_k) \in F\right) \leq 2Ke^{-n\inf_{\mu\in F}\Lambda^*(\mu)}, \text{ for any } F \text{ closed in } \mathbb{R},$$

where $K = \max\{K_l, K_u\}$.

Remark 4. According to Proposition 1, when P is a positive stochastic matrix, i.e. all the transitions have positive probability, we can replace K with

$$K \leq \max_{x,y,z} \frac{P(x,z)}{P(y,z)}.$$

Remark 5. According to Proposition 1, when P induces an IID sequence, i.e. all the rows of P are identical, then $K = 1$. Thus Theorem 10 generalizes the classic bound of [18] for finitely supported IID sequences.

Proof of Theorem 10. In order to derive our bounds we use a change of measure argument, an idea due to [32]. We denote by $\mathbb{P}_q^{(\theta)}$ the probability distribution of the Markov chain with

initial distribution q and stochastic matrix P_θ, while for $\theta = 0$ we just write $\mathbb{P}_q$ for $\mathbb{P}_q^{(0)}$. The finite dimensional distributions $\mathbb{P}_q \mid_{\mathcal{F}_n}$ and $\mathbb{P}_q^{(\theta)} \mid_{\mathcal{F}_n}$ are absolutely continuous with each other and their Radon-Nikodym derivative is given by

$$\frac{d\,\mathbb{P}_q \mid_{\mathcal{F}_n}}{d\,\mathbb{P}_q^{(\theta)} \mid_{\mathcal{F}_n}} = \frac{v_\theta(X_0)}{v_\theta(X_n)} \exp\left\{-\theta S_n + n\Lambda(\theta)\right\},$$

where we denote the sums by $S_n := \sum_{k=1}^n f(X_k)$.

Fix $\theta \in \mathbb{R}_{\geq 0}$. Then

$$\begin{aligned}
\mathbb{P}_q(S_n \geq n\mu) &= \mathbb{E}_q\left[1\{S_n \geq n\mu\}\right] \\
&= \mathbb{E}_q^{(\theta)}\left[\frac{v_\theta(X_0)}{v_\theta(X_n)} e^{-\theta S_n + n\Lambda(\theta)} 1\{S_n \geq n\mu\}\right] \\
&\leq K_u\, \mathbb{E}_q^{(\theta)}\left[e^{-\theta(S_n - n\mu)} 1\{S_n \geq n\mu\}\right] e^{-n(\theta\mu - \Lambda(\theta))} \\
&\leq K_u e^{-n(\theta\mu - \Lambda(\theta))},
\end{aligned}$$

where in the first inequality we used Proposition 1.

When $\mu \in [\pi(f), b)$, we can set $\theta = \Lambda'^{-1}(\mu) \geq \Lambda'^{-1}(\pi(f)) = 0$ and then from Lemma 6 we have that $D\left(\mu \parallel \pi(f)\right) = \theta\mu - \Lambda(\theta)$. When $\mu = b$, we let θ go to ∞. The conclusion follows from Corollary 2. □

In this bound we cannot hope for something more than a constant prefactor. First of all, by differentiating twice the formula proved in Lemma 6 we obtain

$$\lim_{\mu \to \pi(f)} \frac{1}{(\mu - \pi(f))^2} D\left(\mu \parallel \pi(f)\right) = \frac{1}{2}\frac{1}{\Lambda''(0)}.$$

In addition, if we fix $z \geq 0$ and set $\mu = \pi(f) + cz/\sqrt{n}$, where $c^2 = \pi\left(\hat{f}^2 - (P\hat{f})^2\right)$ and $\hat{f}$ is a solution of the *Poisson equation* $(I - P)\hat{f} = f - \pi(f)$, then due to the central limit theorem for Markov chains, see for instance [19], we have that

$$\lim_{n \to \infty} \mathbb{P}_q(S_n \geq n\mu) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Therefore if we want the optimal rate of exponential decay and a prefactor which does not depend on μ, then the best we can attain is a constant prefactor.

3.5 A maximal inequality for Markov chains

In order to derive a maximal inequality for Markov chains we will utilize an exponential martingale which is essentially coming from the Radon-Nikodym derivative $\frac{d\,\mathbb{P}_q^{(\theta)}|_{\mathcal{F}_n}}{d\,\mathbb{P}_q|_{\mathcal{F}_n}}$.

Lemma 11. *Let $\{X_n\}_{n\in\mathbb{Z}_{\geq 0}}$ be a Markov chain over the finite state space S with an irreducible transition matrix P and initial distribution q. Let $f : S \to \mathbb{R}$ be a nonconstant real-valued function on the state space. Fix $\theta \in \mathbb{R}$ and define,*

$$M_n^\theta = \frac{v_\theta(X_n)}{v_\theta(X_0)} \exp\left\{\theta(f(X_1) + \ldots + f(X_n)) - n\Lambda(\theta)\right\}. \tag{3.11}$$

Then $\{M_n^\theta\}_{n\in\mathbb{Z}_{>0}}$ is a martingale with respect to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{Z}_{>0}}$, where $\mathcal{F}_n$ is the σ-field generated by $X_0, \ldots, X_n$.

Proof.

$$\begin{aligned}
\mathbb{E}(M_{n+1}^\theta \mid \mathcal{F}_n) &= M_n^\theta \frac{e^{-\Lambda(\theta)}}{v_\theta(X_n)} \mathbb{E}(v_\theta(X_{n+1}) e^{\theta f(X_{n+1})} \mid \mathcal{F}_n) \\
&= M_n^\theta \frac{e^{-\Lambda(\theta)}}{v_\theta(X_n)} \sum_{x\in S} v_\theta(x) e^{\theta f(x)} P(X_n, y) \\
&= M_n^\theta \frac{e^{-\Lambda(\theta)}}{v_\theta(X_n)} \sum_{x\in S} \tilde{P}_\theta(X_n, x) v_\theta(x) \\
&= M_n^\theta,
\end{aligned}$$

where in the last equality we used the fact that v_θ is a right Perron-Frobenius eigenvector of $\tilde{P}_\theta$. □

Theorem 11. *Let $\{X_n\}_{n\in\mathbb{Z}_{\geq 0}}$ be an irreducible Markov chain over the finite state space S with transition matrix P, initial distribution q, and stationary distribution π. Let $f : S \to \mathbb{R}$ be a non-constant function on the state space. Denote by $\mu(0) = \sum_{x\in S} f(x)\pi(x)$ the stationary mean when f is applied, and by $\bar{Y}_n = \frac{1}{n}\sum_{k=1}^n Y_k$ the empirical mean, where $Y_k = f(X_k)$. Assume that the pair P, f satisfied conditions A 1, A 2, A 3, and A 4. Then for all $t, c > 1$ we have*

$$\mathbb{P}\left(\exists k \geq 1 : kD\left(\bar{Y}_k \,\middle\|\, \mu(0)\right) \geq \frac{tc}{t-1}\log\log k + t\right) \leq \frac{2Kect^c}{c-1} e^{-t},$$

where $K = K(P, f)$ is a positive constant depending only on the transition probability matrix P and the function f.

Proof. In order to handle the Markovian dependence we need to use the exponential martingale for Markov chains from Lemma 11, as well as continuity results for the right Perron-Frobenius eigenvector.

Following the proof strategy used to establish the law of the iterated logarithm, we split the positive integers into chunks of exponentially increasing sizes. Denote by $\alpha > 1$ the

growth factor, to be specified later, and let $n_m = \lfloor \alpha^m \rfloor$ be the end point of the m-th chunk, with $n_0 = 0$. For the m-th chunk we have that

$$\bigcup_{k=n_{m-1}+1}^{n_m} \left\{ \mu(0) \geq \bar{Y}_k, kD\left(\bar{Y}_k \,\middle\|\, \mu(0)\right) \geq \frac{tc}{t-1} \log\log k + t \right\} \subseteq$$
$$\bigcup_{k=n_{m-1}+1}^{n_m} \left\{ \mu(0) \geq \bar{Y}_k, D\left(\bar{Y}_k \,\middle\|\, \mu(0)\right) \geq \frac{tc}{(t-1)n_m} \log\log(n_{m-1}+1) + \frac{t}{n_m} \right\}.$$

Let $\mu_m = \inf\left\{ \mu < \mu(0) : D\left(\mu \,\|\, \mu(0)\right) \leq \frac{tc}{(t-1)n_m} \log\log(n_{m-1}+1) + \frac{t}{n_m} \right\}$, and $\theta_m = \dot{\Lambda}^{-1}(\mu_m) < \dot{\Lambda}^{-1}(\mu(0)) = 0$ so that $\theta_m \mu_m - \Lambda(\theta_m) = D\left(\mu_m \,\|\, \mu(0)\right)$. Then,

$$\begin{aligned}
\left\{ \mu(0) \geq \bar{Y}_k,\ D\left(\bar{Y}_k \,\middle\|\, \mu(0)\right) \geq \frac{\epsilon}{n_m} \right\} &\subseteq \left\{ \bar{Y}_k \leq \mu_m \right\} \\
&= \left\{ e^{\theta_m k \bar{Y}_k - k\Lambda(\theta_m)} \geq e^{k(\theta_m \mu_m - \Lambda(\theta_m))} \right\} \\
&= \left\{ M_k^{\theta_m} \geq \frac{v_{\theta_m}(X_k)}{v_{\theta_m}(X_0)} e^{kD(\mu_m \,\|\, \mu(0))} \right\} \\
&\subseteq \left\{ M_k^{\theta_m} \geq \frac{v_{\theta_m}(X_k)}{v_{\theta_m}(X_0)} e^{(n_{m-1}+1)D(\mu_m \,\|\, \mu(0))} \right\}.
\end{aligned}$$

At this point we use assumptions A 3, and A 4 in order to invoke Proposition 1, and get that there exists a constant $K_l = K_l(P, f) \geq 1$ such that

$$\frac{1}{K_l} \leq \inf_{\theta \in \mathbb{R}_{\leq 0}, x, y \in S} \frac{v_\theta(y)}{v_\theta(x)}.$$

This gives us the inclusion,

$$\left\{ M_k^{\theta_m} \geq \frac{v_{\theta_m}(X_k)}{v_{\theta_m}(X_0)} e^{(n_{m-1}+1)D(\mu_m \,\|\, \mu(0))} \right\} \subseteq \left\{ M_k^{\theta_m} \geq \frac{e^{(n_{m-1}+1)D(\mu_m \,\|\, \mu(0))}}{K_l} \right\}.$$

In Lemma 11 we have established that $M_k^{\theta_m}$ is a positive martingale, which combined with a maximal inequality for martingales due to [91] (see Exercise 4.8.2 in [29] for a modern reference), yields that,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{k=n_{m-1}+1}^{n_m} \left\{ M_k^{\theta_m} \geq \frac{e^{(n_{m-1}+1)D(\mu_m \,\|\, \mu(0))}}{K_l} \right\} \right) \\
&\quad \leq K_l e^{-(n_{m-1}+1)D(\mu_m \,\|\, \mu(0))} \\
&\quad \leq K_l \exp\left\{ -\frac{n_{m-1}+1}{n_m} \cdot \frac{tc}{(t-1)} \log\log(n_{m-1}+1) - \frac{n_{m-1}+1}{n_m} t \right\} \\
&\quad \leq K_l \exp\left\{ -\frac{tc}{\alpha(t-1)} \log\log(\lfloor \alpha^{m-1} \rfloor + 1) - \frac{t}{\alpha} \right\}.
\end{aligned}$$

We pick the growth factor $\alpha = t/(t-1)$, and we union bound over the chunks, to deduce that

$$\mathbb{P}\left(\exists k \geq 1 : \mu(0) \geq \bar{Y}_k, kD\left(\bar{Y}_k \,\middle\|\, \mu(0)\right) \geq \frac{tc}{t-1}\log\log k + t\right) \leq \frac{K_l e c t^c}{c-1} e^{-t}.$$

In the same way we can also deduce, by consuming assumptions A 1, and A 2, that

$$\mathbb{P}\left(\exists k \geq 1 : \mu(0) \leq \bar{Y}_k, kD\left(\bar{Y}_k \,\middle\|\, \mu(0)\right) \geq \frac{tc}{t-1}\log\log k + t\right) \leq \frac{K_u e c t^c}{c-1} e^{-t}.$$

From this the conclusion follows with $K = \max\{K_l, K_u\}$. □

3.6 A Uniform Multiplicative Ergodic Theorem

Theorem 12. *Let P be an irreducible stochastic matrix on the finite state space S, which combined with a real-valued function $f : S \to \mathbb{R}$ satisfies A 1-A 4. Then*

$$\sup_{\theta \in \mathbb{R}} |\Lambda_n(\theta) - \Lambda(\theta)| \leq \frac{\log K}{n}.$$

where K is the constant from Proposition 1.

Therefore $\Lambda_n(\theta)$ converges uniformly on $\mathbb{R}$ to $\Lambda(\theta)$ as $n \to \infty$.

Proof. We start with the calculation

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

From this, using the fact that v_θ is a right Perron-Frobenius eigenvector of $\tilde{P}_\theta$, we obtain

$$\min_{x,y} \frac{v_\theta(y)}{v_\theta(x)} \leq \exp\left\{n\Lambda_n(\theta) - n\Lambda(\theta)\right\} \leq \max_{x,y} \frac{v_\theta(y)}{v_\theta(x)}.$$

The conclusion now follows by applying Proposition 1. □

Appendix

3.A Analyticity of Perron-Frobenius Eigenvalues and Eigenvectors

Here we use the implicit function theorem in order to deduce in Lemma 12 that the Perron-Frobenius eigenvalue and eigenvectors are analytic functions of the entries of the matrix, at a level of generality adequate for our purposes.

Lemma 12. *Let $M \in \mathbb{R}^{s\times s}_{\geq 0}$ be a nonnegative irreducible matrix. Let W range over $\mathbb{R}^{s\times s}$ in an open neighborhood of M. Then u_W, $\rho(W)$ and v_W are analytic as functions of the entries of the matrix W in an open neighborhood of M where W is irreducible.*

Proof. We define the vector-valued function $F : \mathbb{R}^{(s+1)^2} \to \mathbb{R}^{2(s+1)}$

$$F(W, u, \rho, v) := \begin{bmatrix} (W^\top - \rho I)u \\ \mathbf{1}^\top u - 1 \\ (W - \rho I)v \\ u^\top v - 1 \end{bmatrix},$$

where we use column vectors, and $\mathbf{1}$ denotes the all ones vector. At this point no assumptions are made about the structure of W. Note that each coordinate of the vector $F(W, u, \rho, v)$ is a multivariate polynomial of degree at most two, and hence each coordinate is an analytic function of W, u, ρ and v.

In addition $F(M, u_M, \rho(M), v_M) = 0$, and the Jacobian of F with respect to u, ρ, v evaluated at $W = M, u = u_M, \rho = \rho(M), v = v_M$ is

$$J_{F,u,\rho,v}(M, u_M, \rho(M), v_M) = \begin{bmatrix} M^\top - \rho(M)I & -u_M & 0 \\ \mathbf{1}^\top & 0 & 0 \\ 0 & -v_M & M - \rho(M)I \\ v_M^\top & 0 & u_M^\top \end{bmatrix}.$$

We can easily verify that this Jacobian is left invertible. If $\begin{bmatrix} u^\top & \rho & v^\top \end{bmatrix}^\top$ is in the kernel of $J_{F,u,\rho,v}(M, u_M, \rho(M), v_M)$, then $M^\top u = \rho(M)u + \rho u_M$, so if we multiply from the left with $v_M^\top$, we get that $\rho = 0$. In the same fashion we can deduce that $u = v = 0$, and thus the kernel of the Jacobian is trivial.

Then the analytic implicit function theorem guarantees that there exists a unique vector-valued function $g : \mathbb{R}^{s^2} \to \mathbb{R}^{2s+1}$ with each coordinate analytic, such that

$$g(M) = \begin{bmatrix} u_M \\ \rho(M) \\ v_M \end{bmatrix}, \text{ and } F(W, g(W)) = 0, \text{ for all } W \text{ in a neighborhood of } M.$$

Finally, due to the Perron-Frobenius theorem $g(W)$ has to equal $\begin{bmatrix} u_W^\top & \rho(W) & v_W^\top \end{bmatrix}^\top$ for irreducible matrices W in this neighborhood of M. □

3.B Proofs from Section 2

Proof of Lemma 5.

(a) Each entry of $\tilde{P}_\theta$ is an analytic function of θ, and the conclusion follows from Lemma 12 in Section 3.A.

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

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

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

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

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

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

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

$$\begin{aligned} \Lambda''(\theta) &= -\mathbb{E}_{(X,Y)\sim\pi_\theta\odot P_\theta} \frac{d^2}{d\theta^2} \log P_\theta(X, Y) \\ &= \mathbb{E}_{(X,Y)\sim\pi_\theta\odot P_\theta} \left(\frac{d}{d\theta} \log P_\theta(X, Y) \right)^2 \\ &= \mathbb{E}_{(X,Y)\sim\pi_\theta\odot P_\theta} \left(f(Y) - \pi_\theta(f) + \frac{v_\theta(X)}{v_\theta(Y)} \frac{d}{d\theta} \frac{v_\theta(Y)}{v_\theta(X)} \right)^2. \end{aligned}$$

(d) Part (c) already ensures that $\Lambda(\theta)$ is convex. Moreover we see that

$$\Lambda''(\theta) = 0 \text{ for all } \theta \in (\theta_1, \theta_2), \quad \text{iff} \quad P_\theta = P_{\frac{\theta_1+\theta_2}{2}} \text{ for all } \theta \in (\theta_1, \theta_2).$$

If such an interval (θ_1, θ_2) exists, then we claim that we can enlarge it to the whole real line. To see this fix any $0 < \epsilon < \frac{\theta_2-\theta_1}{2}$. Then using Lemma 4 twice we obtain that for any $\theta \in (\theta_1, \theta_2)$

$$P_{\theta\pm\epsilon} = (P_\theta)_{\pm\epsilon} = \left(P_{\frac{\theta_1+\theta_2}{2}}\right)_{\pm\epsilon} = P_{\frac{\theta_1+\theta_2}{2}\pm\epsilon} = P_{\frac{\theta_1+\theta_2}{2}}.$$

By repeating this process we see that $P_\theta = P_0 = P$ for all $\theta \in \mathbb{R}$.

Alternatively, if no such interval exists, then $\Lambda'(\theta)$ is strictly increasing and $\Lambda(\theta)$ is strictly convex. Moreover, for $\theta_1 < \theta_2$ we have that $\pi_{\theta_1}(f) = \Lambda'(\theta_1) < \Lambda'(\theta_2) = \pi_{\theta_2}(f)$, and so $P_{\theta_1} \neq P_{\theta_2}$, establishing that in this case $\theta \mapsto P_\theta$ is an injection.

□

Proof of Lemma 6.

$$\begin{aligned} D(\theta_1 \parallel \theta_2) &= \mathbb{E}_{(X,Y)\sim\pi_{\theta_1}\odot P_{\theta_1}} \log \frac{P_{\theta_1}(X,Y)}{P_{\theta_2}(X,Y)} \\ &= \Lambda(\theta_2) - \Lambda(\theta_1) - \Lambda'(\theta_1)(\theta_2 - \theta_1) \\ &\quad + \mathbb{E}_{(X,Y)\sim\pi_{\theta_1}\odot P_{\theta_1}} \log \frac{v_{\theta_1}(Y)}{v_{\theta_1}(X)} - \mathbb{E}_{(X,Y)\sim\pi_{\theta_1}\odot P_{\theta_1}} \log \frac{v_{\theta_2}(Y)}{v_{\theta_2}(X)} \\ &= \Lambda(\theta_2) - \Lambda(\theta_1) - \mu_1(\theta_2 - \theta_1), \end{aligned}$$

where the second equality is using the calculations from the proof of Lemma 5 (b). □

Proof of Lemma 7. From Lemma 5 we have that $\theta \mapsto \theta\mu - \Lambda(\theta)$ is either the linear function $\theta \mapsto (\mu - \pi(f))\theta$, in which case the conclusion follows right away, or otherwise it is strictly concave.

In the latter case $\mathcal{M} = (\mu_-, \mu_+)$ for some $\mu_- < \mu_+$. If $\mu \in \mathcal{M}$, then $\theta = \Lambda'^{-1}(\mu)$ is the unique maximizer and the conclusion follows from Lemma 6. If $\mu = \mu_+$, then the function keeps on growing as $\theta \to \infty$, or equivalently as $\hat{\mu} \to \mu$, which in conjunction with the representation of the relative entropy rate from Lemma 6 establishes this case. If $\mu > \mu_+$, then $\lim_{\theta\to\infty}(\theta\mu - \Lambda(\theta)) = \lim_{\theta\to\infty}\theta(\mu - \mu_+) + \lim_{\hat{\mu}\to\mu_+} D(\hat{\mu} \parallel \pi(f)) = \infty$. The arguments are the same for the other two cases. □

Chapter 4

Best Markovian Arm Identification with Fixed Confidence

4.1 Introduction

In this chapter we study a problem about best arm identification. There are K independent options which are referred to as arms. Each arm a is associated with a discrete time stochastic process, which is characterized by a parameter θ_a and it's governed by the probability law $\mathbb{P}_{\theta_a}$. At each round we select one arm, without any prior knowledge of the statistics of the stochastic processes. The stochastic process that corresponds to the selected arm evolves by one time step, and we observe this evolution through a reward function, while the stochastic processes for the rest of the arms stay still. A confidence level $\delta \in (0, 1)$ is prescribed, and our goal is to identify the arm that corresponds to the process with the highest stationary mean with probability at least $1 - \delta$, and using as few samples as possible.

In the work of [37] 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 $\mathbb{P}_{\theta_a}$ is the law of an irreducible finite state Markov chain associated with a stationary mean $\mu(\theta_a)$. We establish a lower bound (Theorem 13) for the expected sample complexity, as well as an analysis of the Track-and-Stop strategy, proposed for the IID setting in [37], which shows (Theorem 14) 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 [37] 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 [15]. 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. This way one can model for instance the evolution of a rigged slot machine, which as soon as it generates a big reward for the gambler, it changes the reward distribution to a distribution which is skewed towards smaller rewards.

The cornerstone of stochastic multi-armed bandits is the seminal work of [56]. They considered K IID process with the objective being to maximize the expected value of the sum of the observed rewards, or equivalently to minimize *regret*. In the same spirit [3, 4] examine the generalization where one is allowed to collect multiple rewards at each time step, first in the case that processes are IID [3], and then in the case that the processes are irreducible and aperiodic Markov chains [4]. A survey of the regret minimization literature is contained in [15].

An alternative objective is the one of identifying the process with the highest stationary mean as fast as and as accurately as possible, notions which are made precise in Subsection 4.2.1. In the IID setting, [33] establish an elimination based algorithm in order to find an approximate best arm, and [62] provide a matching lower bound. [46] 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 [37], 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 [50].

4.2 Problem Formulation

4.2.1 One-parameter family of Markov Chains

In order to model the problem we will use a one-parameter family of Markov chains on a finite state space S. Each Markov chain in the family corresponds to a parameter $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}$ is the parameter space, and is completely characterized by an initial distribution $q_\theta = [q_\theta(x)]_{x\in S}$, and a stochastic transition matrix $P_\theta = [P_\theta(x,y)]_{x,y\in S}$, which satisfy the following conditions.

$$P_\theta \text{ is irreducible for all } \theta \in \Theta. \tag{4.1}$$

$$P_\theta(x,y) > 0 \;\Rightarrow\; P_\lambda(x,y) > 0, \text{ for all } \theta,\lambda \in \Theta,\; x,y \in S. \tag{4.2}$$

$$q_\theta(x) > 0 \;\Rightarrow\; q_\lambda(x) > 0, \text{ for all } \theta,\lambda \in \Theta,\; x \in S. \tag{4.3}$$

There are K Markovian arms with parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \Theta^K$, and each arm $a \in [K] = \{1, \ldots, K\}$ evolves as a Markov chain with parameter θ_a which we denote by $\{X_n^a\}_{x\in\mathbb{Z}_{\geq 0}}$. A non-constant real valued reward function $f : S \to \mathbb{R}$ is applied at each state and produces the reward process $\{Y_n^a\}_{n\in\mathbb{Z}_{\geq 0}}$ given by $Y_n^a = f(X_n^a)$. We can only observe the reward process but not the internal Markov chain. Note that the reward process is a function of the Markov chain and so in general it will have more complicated dependencies than the

Markov chain. The reward process is a Markov chain if and only if f is injective. For each $\theta \in \Theta$ there is a unique stationary distribution $\pi_\theta = [\pi_\theta(x)]_{x\in S}$ associated with the stochastic matrix P_θ, due to (4.1). This allows us to define the stationary reward of the Markov chain corresponding to the parameter θ as $\mu(\theta) = \sum_x f(x)\pi_\theta(x)$. We will assume that among the K Markovian arms there exists precisely one that possess the highest stationary mean, and we will denote this arm by $a^*(\boldsymbol{\theta})$, so in particular

$$\{a^*(\boldsymbol{\theta})\} = \arg\max_{a\in[K]} \mu(\theta_a).$$

The set of all parameter configurations that possess a unique highest mean is denoted by

$$\boldsymbol{\Theta} = \left\{ \boldsymbol{\theta} \in \Theta^K : \left| \arg\max_{a\in[K]} \mu(\theta_a) \right| = 1 \right\}.$$

The *Kullback-Leibler divergence rate* characterizes the sample complexity of the Markovian identification problem that we are about to study. For two Markov chains of the one-parameter family that are indexed by θ and λ respectively it is given by,

$$D(\theta \parallel \lambda) = \sum_{x,y\in S} \log \frac{P_\theta(x,y)}{P_\lambda(x,y)} \pi_\theta(x) P_\theta(x,y),$$

where we use the standard notational conventions $\log 0 = \infty$, $\log \frac{\alpha}{0} = \infty$ if $\alpha > 0$, and $0 \log 0 = 0 \ln \frac{0}{0} = 0$. It is always nonnegative, $D(\theta \parallel \lambda) \geq 0$, with equality occurring if and only if $P_\theta = P_\lambda$, and so $\mu(\theta) \neq \mu(\lambda)$ yields that $D(\theta \parallel \lambda) > 0$. Furthermore, $D(\theta \parallel \lambda) < \infty$ due to (4.2).

With some abuse of notation we will also write $D(\mathbb{P} \parallel \mathbb{Q})$ for the Kullback-Leibler divergence between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same measurable space, which is defined as

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

where $\mathbb{P} \ll \mathbb{Q}$ means that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$, and in that case $\frac{d\mathbb{P}}{d\mathbb{Q}}$ denotes the Radon-Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{Q}$.

4.2.2 Best Markovian Arm Identification with Fixed Confidence

Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be an unknown parameter configuration for the K Markovian arms. Let $\delta \in (0,1)$ be a given confidence level. Our goal is to identify $a^*(\boldsymbol{\theta})$ with probability at least $1-\delta$ using as few samples as possible. At each time t we select a single arm A_t and we observe the next sample from the reward process $\{Y_n^{A_t}\}_{n\in\mathbb{Z}_{\geq 0}}$, while all the other reward processes stay still. Let $N_a(t) = \sum_{s=1}^t I_{\{A_s=a\}} - 1$ be the number of transitions of the Markovian arm a up to time t. Let $\mathcal{F}_t$ be the σ-field generated by our choices $A_1, \ldots, A_t$ and the observations $\{Y_n^1\}_{n=0}^{N_1(t)}, \ldots, \{Y_n^a\}_{n=0}^{N_K(t)}$. A sampling strategy, $\mathcal{A}_\delta$, 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 decisions and 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 a stopping time with respect to the filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}_{>0}}$, such that $\mathbb{E}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}[\tau_\delta] < \infty$ for all $\boldsymbol{\lambda} \in \boldsymbol{\Theta}$;
- 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.

Sampling strategies need to perform well across all possible parameter configurations in $\boldsymbol{\Theta}$, 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 confidence level $\delta \in (0,1)$, a sampling strategy $\mathcal{A}_\delta = ((A_t)_{t\in\mathbb{Z}_{>0}}, \tau_\delta, \hat{a}_{\tau_\delta})$ is said to be δ-PC (Probably Correct) if,

$$\mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}(\hat{a}_{\tau_\delta} \neq a^*(\boldsymbol{\lambda})) \leq \delta, \text{ for all } \boldsymbol{\lambda} \in \boldsymbol{\Theta}.$$

Therefore our goal is to study the quantity,

$$\inf_{\mathcal{A}_\delta:\delta-PC} \mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[\tau_\delta],$$

both in terms of finding a lower bound, i.e. establishing that no δ-PC strategy can have expected sample complexity less than our lower bound, and also in terms of finding an upper bound, i.e. a δ-PC strategy with very small expected sample complexity. We will do so in the high confidence regime of $\delta \to 0$, by establishing instance specific lower and upper bounds which differ just by a factor of four.

4.3 Lower Bound on the Sample Complexity

Deriving lower bounds in the multi-armed bandits setting is a task performed by change of measure arguments initial introduced by [56]. Those change of measure arguments capture the simple idea 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. Fix $\theta \in \boldsymbol{\Theta}$, and define the set of parameter configurations that exhibit as best arm an arm different than $a^*(\boldsymbol{\theta})$ by

$$\text{Alt}(\boldsymbol{\theta}) = \{\boldsymbol{\lambda} \in \boldsymbol{\Theta} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\theta})\}.$$

Then we consider an alternative parametrization $\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\theta})$ and we write their log-likelihood ratio up to time t

$$\begin{aligned}\log\left(\frac{d\,\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}\,|_{\mathcal{F}_t}}{d\,\mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}\,|_{\mathcal{F}_t}}\right) &= \sum_{a=1}^{K} I_{\{N_a(t)\geq 0\}} \log\frac{q_{\theta_a}(X_0^a)}{q_{\lambda_a}(X_0^a)} \\ &+ \sum_{a=1}^{K}\sum_{x,y} N_a(x,y,0,t)\log\frac{P_{\theta_a}(x,y)}{P_{\lambda_a}(x,y)},\end{aligned} \tag{4.4}$$

where $N_a(x,y,0,t) = \sum_{s=0}^{t-1} 1\{X_s^a = x, X_{s+1}^a = y\}$. The log-likelihood ratio enables us to perform changes of measure for fixed times t, and more generally for stopping times τ with respect to $(\mathcal{F}_t)_{t\in\mathbb{Z}_{>0}}$, which are $\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}$ and $\mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}$-a.s. finite, through the following change of measure formula,

$$\mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}(\mathcal{E}) = \mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}\left[I_{\mathcal{E}} \frac{d\,\mathbb{P}_{\boldsymbol{\lambda}}\,|_{\mathcal{F}_\tau}}{d\,\mathbb{P}_{\boldsymbol{\theta}}\,|_{\mathcal{F}_\tau}}\right], \text{ for any } \mathcal{E}\in\mathcal{F}_\tau. \tag{4.5}$$

In order to derive our lower bound we use a technique developed for the IID case by [37] which combines several changes of measure at once. To make this technique work in the Markovian setting we need the following inequality which we derive in Section 4.A using a renewal argument for Markov chains.

Lemma 13. *Let $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ and $\boldsymbol{\lambda}\in Alt(\boldsymbol{\theta})$ be two parameter configurations. Let τ be a stopping time with respect to $(\mathcal{F}_t)_{t\in\mathbb{Z}_{>0}}$, with $\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[\tau],\ \mathbb{E}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}[\tau]<\infty$. Then*

$$\begin{aligned} D\left(\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}\,|_{\mathcal{F}_\tau} \,\middle\|\, \mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}\,|_{\mathcal{F}_\tau}\right) &\le \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_a(\tau)] D\left(\theta_a \,\|\, \lambda_a\right) \\ &+ \sum_{a=1}^{K} D\left(q_{\theta_a} \,\|\, q_{\lambda_a}\right) + \sum_{a=1}^{K} R_{\theta_a} \sum_{x,y} \pi_{\theta_a}(x) P_{\theta_a}(x,y) \left|\log \frac{P_{\theta_a}(x,y)}{P_{\lambda_a}(x,y)}\right|, \end{aligned}$$

where $R_{\theta_a} = \mathbb{E}_{\theta_a}\left[\inf\{n>0 : X_n^a = X_0^a\}\right] < \infty$, the first summand is finite due to (4.3), and the second summand is finite due to (4.2).

Combining those ingredients with the data processing inequality we derive our instance specific lower bound for the Markovian bandit identification problem in Section 4.A.

Theorem 13. *Assume that the one-parameter family of Markov chains on the finite state space S satisfies conditions (4.1), (4.2), and (4.3). Fix $\delta\in(0,1)$, let $f: S\to\mathbb{R}$ be a nonconstant reward function, let $\mathcal{A}_\delta$ be a δ-PC sampling strategy, and fix a parameter configuration $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Then*

$$T^*(\boldsymbol{\theta}) \le \liminf_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[\tau_\delta]}{\log\frac{1}{\delta}},$$

where

$$T^*(\boldsymbol{\theta})^{-1} = \sup_{\boldsymbol{w}\in\mathcal{M}_1([K])} \inf_{\boldsymbol{\lambda}\in Alt(\boldsymbol{\theta})} \sum_{a=1}^{K} w_a D\left(\theta_a \,\|\, \lambda_a\right),$$

and $\mathcal{M}_1([K])$ denotes the set of all probability distributions on $[K]$.

As noted in [37] the sup in the definition of $T^*(\boldsymbol{\theta})$ is actually attained uniquely, and therefore we can define $\boldsymbol{w}^*(\boldsymbol{\theta})$ as the unique maximizer,

$$\{\boldsymbol{w}^*(\boldsymbol{\theta})\} = \operatorname*{arg\,max}_{\boldsymbol{w}\in\mathcal{M}_1([K])} \inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\theta})} \sum_{a=1}^{K} w_a D\left(\theta_a \,\|\, \lambda_a\right).$$

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

The (α, δ)-Track-and-Stop strategy, which was proposed in [37] in order to tackle the IID setting, tries to track the optimal weights $w_a^*(\boldsymbol{\theta})$. In the sequel we will also write $\boldsymbol{w}^*(\boldsymbol{\mu})$, with $\boldsymbol{\mu} = (\mu(\theta_1), \dots, \mu(\theta_K))$, to denote $\boldsymbol{w}^*(\boldsymbol{\theta})$. Not having access to $\boldsymbol{\mu}$, the (α, δ)-Track-and-Stop strategy tries to approximate $\boldsymbol{\mu}$ using sample means. Let $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(N_1(t)), \dots, \hat{\mu}_K(N_K(t)))$ 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.

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

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.

4.4.1 Sampling Rule: Tracking the Optimal Proportions

For initialization reasons the first $2K$ samples that we are going to observe are $Y_0^1, Y_1^1, \dots, Y_0^K, Y_1^K$. After that, for $t \geq 2K$ we let $U_t = \{a : N_a(t) < \sqrt{t} - K/2\}$ and we follow the tracking rule:

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

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 $\boldsymbol{\mu} \mapsto \boldsymbol{w}^*(\boldsymbol{\mu})$, 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 $a = 1, \dots, K$.

4.4.2 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 [37]

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

where $\beta_{\alpha,\delta}(t) = 2\log\frac{Dt^\alpha}{\delta}$, $D = \dfrac{2\alpha K C^2}{\alpha - 1}$, and $C = C(P,f)$ is the constant from Proposition 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 $C = C(P,f)$ does not depend on the bandit instance $\boldsymbol{\theta}$ or the confidence level δ, but only on the generator stochastic matrix P and the reward function f. In other words it is a characteristic of the exponential family of Markov chains and not of the particular bandit instance, $\boldsymbol{\theta}$, under consideration.

4.4.3 Decision Rule: Best Sample Mean

For a fixed arm a it is clear that, $\min_{b\neq a} Z_{a,b}(t) > 0$ if and only if $\hat{\mu}_a(N_a(t)) > \hat{\mu}_b(N_b(t))$ 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:

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

4.4.4 Sample Complexity Analysis

In this section we establish that the (α,δ)-Track-and-Stop strategy is δ-PC, and we upper bound its expected sample complexity. In order to do this we use our Markovian concentration bound Theorem 10 from Section 3.4.

We first use it in order to establish the following uniform deviation bound.

Lemma 14. *Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\delta \in (0,1)$, and $\alpha > 1$. Let $\mathcal{A}_\delta$ be a sampling strategy that uses an arbitrary sampling rule, the (α,δ)-Chernoff's stopping rule and the best sample mean decision rule. Then, for any arm a,*

$$\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}\left(\exists t \in \mathbb{Z}_{>0} : N_a(t) D\left(\hat{\mu}_a(N_a(t)) \,\|\, \mu_a\right) \geq \beta_{\alpha,\delta}(t)/2\right) \leq \frac{\delta}{K}.$$

With this in our possession we are able to prove in Section 4.B that the (α,δ)-Track-and-Stop strategy is δ-PC.

Proposition 2. *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, $\delta \to 0$, 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 Theorem 13.

Theorem 14. *Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and $\alpha \in (1, e/4]$. The (α, δ)-Track-and-Stop strategy, denoted here by $\mathcal{A}_\delta$, has its asymptotic expected sample complexity upper bounded by,*

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

Appendix

4.A Lower Bound on the Sample Complexity

We first prove Lemma 13, for which we will apply a renewal argument. Using the *strong Markov property* we can derive the following standard, see [30], decomposition of a Markov chain in IID blocks.

Fact 1. Let $\{X_n\}_{n\in\mathbb{Z}_{\geq 0}}$ be an irreducible Markov chain with initial distribution q, and transition matrix P. 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 $\{v_k\}_{k\in\mathbb{Z}_{>0}}$ of IID random blocks given by

$$v_k = (r_k, X_{\tau_{k-1}}, \ldots, X_{\tau_k - 1}), \text{ for } k \geq 1.$$

Let $N(x, n, m)$ be the number of visits to x that occurred from time n up to 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, see [30], that the stationary distribution π of the Markov chain is given by

$$\pi(x) = \frac{\mathbb{E}_{(q,P)}\, N(x, 0, \tau_1)}{\mathbb{E}_{(q,P)}\, \tau_1}, \text{ for any } x \in S. \tag{4.6}$$

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

Lemma 15.

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

Proof. Using (4.6) it is enough to show that for any initial state x_0,

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

or equivalently 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

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

where the second to last equality holds true due to the reversed Markov property

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

□

The following Lemma, which is a variant of Lemma 2.1 in [4], is the place where we use the IID block structure of the Markov chain.

Lemma 16. *Define the mean return time of the Markov chain with initial distribution q and irreducible transition matrix P by*

$$R = \mathbb{E}_{(q,P)}[\inf\{n > 0 : X_n = X_0\}] < \infty.$$

Let $\mathcal{F}_n$ be the σ-field generated by $X_0, X_1, \ldots, X_n$. Let τ be a stopping time with respect to $(\mathcal{F}_n)_{n\in\mathbb{Z}_{\geq 0}}$, with $\mathbb{E}_{(q,P)}\,\tau < \infty$. Then

$$\left|\mathbb{E}_{(q,P)}\,N(x,y,0,\tau) - \pi(x)P(x,y)\,\mathbb{E}_{(q,P)}\,\tau\right| \leq \pi(x)P(x,y)R, \text{ for all } x,y \in S.$$

Proof. Using the k-th return times from Fact 1 we decompose $N(x,y,0,\tau_k)$ in k IID summands

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

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

$$\tau_\kappa - \tau \leq \tau_\kappa - \tau_{\kappa-1}.$$

Taking expectations we obtain

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

which also gives that

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

This allows us to use Wald's identity, followed by Lemma 15, followed by Wald's identity again, in order to get

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

Therefore,

$$\begin{aligned}
\mathbb{E}_{(q,P)}\,N(x,y,0,\tau) &\leq \mathbb{E}_{(q,P)}\,N(x,y,0,\tau_\kappa) \\
&= \pi(x)P(x,y)\,\mathbb{E}_{(q,P)}[\tau_\kappa] \\
&\leq \pi(x)P(x,y)(\mathbb{E}_{(q,P)}[\tau] + R).
\end{aligned}$$

For the other direction we use the pointwise inequality

$$\tau - \tau_{\kappa-1} \leq \tau_\kappa - \tau_{\kappa-1},$$

to deduce that

$$\begin{aligned}
\mathbb{E}_{(q,P)}\,N(x,y,0,\tau) &\geq \mathbb{E}_{(q,P)}\,N(x,y,0,\tau_{\kappa-1}) \\
&= \pi(x)P(x,y)\,\mathbb{E}_{(q,P)}[\tau_{\kappa-1}] \\
&\geq \pi(x)P(x,y)(\mathbb{E}_{(q,P)}[\tau] - R).
\end{aligned}$$

□

Proof of Lemma 13.
Follows by taking $\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}$ of the log-likelihood ratio, $\log\left(\frac{\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}|_{\mathcal{F}_\tau}}{\mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}|_{\mathcal{F}_\tau}}\right)$, given by (4.4), and applying Lemma 16 K times for the stopping times $N_a(\tau)+1,\ a=1,\dots,K$. □

The last part of Section 4.A involves the proof of Theorem 13.

Proof of Theorem 13.
Consider an alternative parametrization $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\theta})$. The data processing inequality, see [21], gives us as a way to lower bound the Kullback-Leibler divergence between the two probability measures $\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}|_{\mathcal{F}_{\tau_\delta}}$ and $\mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}|_{\mathcal{F}_{\tau_\delta}}$. In particular,

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

where for $p, q \in [0,1]$, $D_2(p \,\|\, q)$ denotes the binary Kullback-Leibler divergence,

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

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

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

hence

$$D_2(\delta \,\|\, 1-\delta) \le D\left(\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}|_{\mathcal{F}_{\tau_\delta}} \,\middle\|\, \mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}|_{\mathcal{F}_{\tau_\delta}}\right).$$

Combining this with Lemma 13 we get that

$$D_2(\delta \,\|\, 1-\delta) \le \sum_{a=1}^{K} D(q_{\theta_a} \,\|\, q_{\lambda_a}) + \sum_{a=1}^{K}\left(\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_a(\tau_\delta)] + R_a\right) D(\theta_a \,\|\, \lambda_a).$$

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

$$\begin{aligned} &D_2(\delta \,\|\, 1-\delta) - \sum_{a=1}^{K} D(q_{\theta_a} \,\|\, q_{\lambda_a}) \\ &\quad\le \left(\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[\tau_\delta] + \sum_{a=1}^{K} R_a\right) \sum_{a=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_a(\tau_\delta)] + R_a}{\sum_{b=1}^{K}\left(\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_b(\tau_\delta)] + R_b\right)} D(\theta_a \,\|\, \lambda_a), \end{aligned}$$

and now we follow the technique of [37] which combines multiple alternative models $\boldsymbol{\lambda}$,

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

The conclusion follows by letting δ go to 0, and using the fact that

$$\lim_{\delta\to 0} \frac{D_2\left(\delta \parallel 1-\delta\right)}{\log\frac{1}{\delta}} = 1.$$

□

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

The proof of Lemma 14 uses the concentration bound Theorem 10, combined with the monotonicity of the Kullback-Leibler divergence rate.

Proof of Lemma 14.
We first note the following inclusion of events

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

where the last equality follows because, by the monotonicity of $t \mapsto \beta_{\alpha,\delta}(t)/2$ we have that for each $n \in \mathbb{Z}_{>0}$ and for each $t = n, n+1, \ldots$

$$\left\{n D\left(\hat{\mu}_a(n) \parallel \mu_a\right) \geq \beta_{\alpha,\delta}(t)/2\right\} \subseteq \left\{n D\left(\hat{\mu}_a(n) \parallel \mu_a\right) \geq \beta_{\alpha,\delta}(n)/2\right\}.$$

Combining this with a union bound we obtain

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}\left(\exists t \in \mathbb{Z}_{>0} : N_a(t) D\left(\hat{\mu}_a(N_a(t)) \parallel \mu_a\right) \geq \beta_{\alpha,\delta}(t)/2\right) \\
&\quad \leq \mathbb{P}_{\theta_a}\left(\exists t \in \mathbb{Z}_{>0} : t D\left(\hat{\mu}_a(t) \parallel \mu_a\right) \geq \beta_{\alpha,\delta}(t)/2\right) \\
&\quad \leq \sum_{t=1}^{\infty} \mathbb{P}_{\theta_a}\left(D\left(\hat{\mu}_a(t) \parallel \mu_a\right) \geq \frac{\beta_{\alpha,\delta}(t)}{2t}\right).
\end{aligned}$$

We focus on upper bounding

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

Let $\mu_{a,t}$ be the unique (due to the monotonicity of the Kullback-Leibler divergence rate) solution (if no solution exists then the probability is already zero) of the equations

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

Then the combination of the monotonicity of the Kullback-Leibler divergence rate, and Theorem 10 gives

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

We further upper bound the constant $c(P_{\mu_a})$ by $c(P)^2$ using Lemma 4, in order to obtain a uniform upper bound for any Markovian arm coming from the family.

A similar bound holds true for

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

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 3. *Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\delta \in (0,1)$, and $\alpha > 1$. Let $\mathcal{A}_\delta$ be a sampling strategy that uses an arbitrary sampling rule, the (α,δ)-Chernoff's stopping rule and the best sample mean decision rule. Then,*

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

Proof. The following lemma which is easy to check, and its proof is omitted, will be useful in our proof of Proposition 3.

Lemma 17. *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 17, and the last inequality follows from the fact that $\mu_a < \mu_{a^*(\boldsymbol{\mu})}$.

This in turn implies that

$$\beta_{\alpha,\delta}(t)/2 < N_a(t) D\left(\hat{\mu}_a(N_a(t)) \parallel \mu_a\right),$$

or

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

Therefore by union bounding over the K arms we obtain

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

The conclusion now follows by applying Lemma 14. □

Proof of Proposition 2.
Following the proof of Proposition 13 in [37], and observing that in their proof they show that $\tau_{\alpha,\delta}$ is essentially bounded we obtain that

$$\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[\tau_{\alpha,\delta}] < \infty.$$

This combined with Proposition 3 establishes that the (α, δ)-Track-and-Stop strategy is δ-PC. □

Proof of Theorem 14.
Finally for the proof the sample complexity of the (α, δ)-Track-and-Stop strategy in Theorem 14 we follow the proof of Theorem 14 in [37], 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 10. □

Chapter 5

Regret minimization for Markovian bandits

5.1 Introduction

In this chapter we study a generalization of the stochastic multi-armed bandit problem, where there are K independent arms, and each arm $a \in [K] = \{1, \dots, K\}$ is associated with a parameter $\theta_a \in \mathbb{R}$, and modeled as a discrete time stochastic process governed by the probability law $\mathbb{P}_{\theta_a}$. A time horizon T is prescribed, and at each round $t \in [T] = \{1, \dots, T\}$ we select M arms, where $1 \leq M \leq K$, without any prior knowledge of the statistics of the underlying stochastic processes. The M stochastic processes that correspond to the selected arms evolve by one time step, and we observe this evolution through a reward function, while the stochastic processes for the rest of the arms stay frozen, i.e. we consider the rested bandits setting. Our goal is to select arms in such a way so as to make the cumulative reward over the whole time horizon T as large as possible. For this task we are faced with an exploitation versus exploration dilemma. At each round we need to decide whether we are going to exploit the best M arms according to the information that we have gathered so far, or we are going to explore some other arms which do not seem to be so rewarding, just in case that the rewards we have observed so far deviate significantly from the expected rewards. The answer to this dilemma is usually coming by calculating indices for the arms and ranking them according to those indices, which should incorporate both information on how good an arm seems to be as well as on how many times it has been played so far. Here we take an alternative approach where instead of calculating the indices for all the arms at each round, we just calculate the index for a single arm in a round-robin way.

5.1.1 Contributions

1. We first consider the case that the K stochastic processes are irreducible Markov chains, coming from a one-parameter exponential family of Markov chains. The objective is to play as much as possible the M arms with the largest stationary means, although

we have no prior information about the statistics of the K Markov chains. The difference of the best possible expected rewards coming from those M best arms and the expected reward coming from the arms that we played is the regret that we incur. To minimize the regret we consider an index based adaptive allocation rule, Algorithm 3, which is based on sample means and Kullback-Leibler upper confidence bounds for the stationary expected rewards using the Kullback-Leibler divergence rate. We provide a finite-time analysis, Theorem 15, for this KL-UCB adaptive allocation rule which shows that the regret depends logarithmically on the time horizon T, and matches exactly the asymptotic lower bound, Corollary 3.

2. In order to make the finite-time guarantee possible we devise several deviation lemmata for Markov chains. An exponential martingale for Markov chains is proven, Lemma 11, which leads to a maximal inequality for Markov chains, Lemma 18. In the literature there are several approaches that use martingale techniques either to derive Hoeffding inequalities for Markov chains [40, 70], or more generally to study concentration of measure for Markov chains [63, 64, 66, 83, 65, 17, 53, 77]. Nonetheless, they're all based either on Dynkin's martingale or on Doob's martingale, combined with coupling ideas, and there is no evidence that they can lead to maximal inequalities. Moreover, a Chernoff bound for Markov chains is devised, Lemma 19, and its relation with the work of [74] is discussed in Remark 10.

3. We then consider the case that the K stochastic processes are i.i.d. processes, each corresponding to a density coming from a one-parameter exponential family of densities. We establish, Theorem 16, that Algorithm 3 still enjoys the same finite-time regret guarantees, which are asymptotically optimal. The case where Theorem 16 follows directly from Theorem 15 is discussed in Remark 8. The setting of single plays is studied in [16], but with a much more computationally intense adaptive allocation rule.

4. In Section 5.6 we provide simulation results illustrating the fact that round-robin KL-UCB adaptive allocation rules are much more computationally efficient than KL-UCB adaptive allocation rules, and similarly round-robin UCB adaptive allocation rules are more computationally efficient than UCB adaptive allocation rules, while the expected regrets, in each family of algorithms, behave in a similar way. This brings to light round-robin schemes as an appealing practical alternative to the mainstream schemes that calculate indices for all the arms at each round.

5.1.2 Motivation

Multi-armed bandits provide a simple abstract statistical model that can be applied to study real world problems such as clinical trials, ad placement, gambling, adaptive routing, resource allocation in computer systems etc. We refer the interested reader to the survey of [15] for more context, and to the recent books of [58, 84]. The need for multiple plays

can be understood in the setting of resource allocation. Scheduling jobs to a single CPU is an instance of the multi-armed bandit problem with a single play at each round, where the arms correspond to the jobs. If there are multiple CPUs we get an instance of the multi-armed bandit problem with multiple plays. The need of a richer model which allows the presence of Markovian dependence is illustrated in the context of gambling, where the arms correspond to slot-machines. It is reasonable to try to model the assertion that if a slot-machine produced a high reward the n-th time played, then it is very likely that it will produce a much lower reward the $(n+1)$-th time played, simply because the casino may decide to change the reward distribution to a much stingier one if a big reward was just produced. This assertion requires, the reward distributions to depend on the previous outcome, which is precisely captured by the Markovian reward model. Moreover, we anticipate this to be an important problem attempting to bridge classical stochastic bandits, controlled Markov chains (MDPs), and non-stationary bandits.

5.1.3 Related Work

The cornerstone of the multi-armed bandits literature is the pioneering work of [56], which studies the problem for the case of i.i.d. rewards and single plays. [56] introduce the change of measure argument to derive a lower bound for the problem, as well as round robin adaptive allocation rules based on upper confidence bounds which are proven to be asymptotically optimal. [3] extend the results of [56] to the case of i.i.d. rewards and multiple plays, while [1] considers index based allocation rules which are only based on sample means and are computationally simpler, although they may not be asymptotically optimal. The work of [1] inspired the first finite-time analysis for the adaptive allocation rule called UCB by [5], which is though asymptotically suboptimal. The works of [16, 36, 61] bridge this gap by providing the KL-UCB adaptive allocation rule, with finite-time guarantees which are asymptotically optimal. Additionally, [51] study a Thompson sampling algorithm for multiple plays and binary rewards, and they establish a finite-time analysis which is asymptotically optimal. Here we close the problem of multiple plays and rewards coming from an exponential family of probability densities by showing finite-time guarantees which are asymptotically optimal, via adaptive allocation rules which are much more efficiently computable than their precursors.

The study of Markovian rewards and multiple plays in the rested setting, is initiated in the work of [4]. They report an asymptotic lower bound, as well as a round robin upper confidence bound adaptive allocation rule which is proven to be asymptotically optimal. However, it is unclear if the statistics that they use in order to derive the upper confidence bounds, in their Theorem 4.1, can be recursively computed, and the practical applicability of their results is therefore questionable. In addition, they don't provide any finite-time analysis, and they use a different type of assumption on their one-parameter family of Markov chains. In particular, they assume that their one-parameter family of transition probability matrices is log-concave in the parameter, equation (4.1) in [4], while we assume that it is a one-parameter exponential family of transition probability matrices. [87, 88] extend the UCB adaptive allocation rule of [5], to the case of Markovian rewards and multiple plays.

They provide a finite-time analysis, but their regret bounds are suboptimal. Moreover they impose a different type of assumption on their configuration of Markov chains. They assume that the transition probability matrices are reversible, so that they can apply Hoeffding bounds for Markov chains from [39, 60]. In a recent work [70] developed a Hoeffding bound for Markov chains, which does not assume any conditions other than irreducibility, and using this he extended the analysis of UCB to an even broader class of Markov chains. One of our main contributions is to bridge this gap and provide a KL-UCB adaptive allocation rule, with a finite-time guarantee which is asymptotically optimal. In a different line of work [76, 88] consider the restless bandits Markovian reward model, in which the state of each arm evolves according to a Markov chain independently of the player's action. Thus in the restless setting the state that we next observe is now dependent on the amount of time that elapses between two plays of the same arm.

5.2 Problem Formulation

5.2.1 One-Parameter Family of Markov Chains

We consider a one-parameter family of irreducible Markov chains on a finite state space S. Each member of the family is indexed by a parameter $\theta \in \mathbb{R}$, and is characterized by an initial distribution $q_\theta = [q_\theta(x)]_{x\in S}$, and an irreducible transition probability matrix $P_\theta = [P_\theta(x,y)]_{x,y\in S}$, which give rise to a probability law $\mathbb{P}_\theta$. There are $K \geq 2$ arms, with overall parameter configuration $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \mathbb{R}^K$, and each arm $a \in [K] = \{1, \ldots, K\}$ evolves internally as the Markov chain with parameter θ_a which we denote by $\{X_n^a\}_{n\in\mathbb{Z}_{\geq 0}}$. There is a common noncostant real-valued reward function on the state space $f : S \to \mathbb{R}$, and successive plays of arm a result in observing samples from the stochastic process $\{Y_n^a\}_{n\in\mathbb{Z}_{\geq 0}}$, where $Y_n^a = f(X_n^a)$. In other words, the distribution of the rewards coming from arm a is a function of the Markov chain with parameter θ_a, and thus it can have more complicated dependencies. As a special case, if we pick the reward function f to be injective, then the distribution of the rewards is Markovian.

For $\theta \in \mathbb{R}$, due to irreducibility, there exists a unique stationary distribution for the transition probability matrix P_θ which we denote with $\pi_\theta = [\pi_\theta(x)]_{x\in S}$. Furthermore, let $\mu(\theta) = \sum_{x\in S} f(x)\pi_\theta(x)$ be the stationary mean reward corresponding to the Markov chain parametrized by θ. Without loss of generality we may assume that the K arms are ordered so that,

$$\mu(\theta_1) \geq \ldots \geq \mu(\theta_N) > \mu(\theta_{N+1}) \ldots = \mu(\theta_M) = \ldots = \mu(\theta_L) > \mu(\theta_{L+1}) \geq \ldots \geq \mu(\theta_K),$$

for some $N \in \{0, \ldots, M-1\}$ and $L \in \{M, \ldots, K\}$, where $N = 0$ means that $\mu(\theta_1) = \ldots = \mu(\theta_M)$, $L = K$ means that $\mu(\theta_M) = \ldots = \mu(\theta_K)$, and we set $\mu(\theta_0) = \infty$ and $\mu(\theta_{K+1}) = -\infty$.

5.2.2 Regret Minimization

We fix a time horizon T, and at each round $t \in [T] = \{1, \dots, T\}$ we play a set ϕ_t of M distinct arms, where $1 \leq M \leq K$ is the same through out the rounds, and we observe rewards $\{Z_t^a\}_{a\in[K]}$ given by,

$$Z_t^a = \begin{cases} Y^a_{N_a(t)}, & \text{if } a \in \phi_t \\ 0, & \text{if } a \notin \phi_t, \end{cases}$$

where $N^a(t) = \sum_{s=1}^t I\{a \in \phi_s\}$ is the number of times we played arm a up to time t. Using the stopping times $\tau_n^a = \inf\{t \geq 1 : N^a(t) = n\}$, we can also reconstruct the $\{Y_n^a\}_{n\in\mathbb{Z}_{>0}}$ process, from the observed $\{Z_t^a\}_{t\in\mathbb{Z}_{>0}}$ process, via the identity $Y_n^a = Z^a_{\tau_n^a}$. Our play ϕ_t is based on the information that we have accumulated so far. In other words, the event $\{\phi_t = A\}$, for $A \subseteq [K]$ with $|A| = M$, belongs to the σ-field generated by $\phi_1, \{Z_1^a\}_{a\in[K]}, \dots, \phi_{t-1}, \{Z_{t-1}^a\}_{a\in[K]}$. We call the sequence $\boldsymbol{\phi} = \{\phi_t\}_{t\in\mathbb{Z}_{>0}}$ of our plays an *adaptive allocation rule*. Our goal is to come up with an adaptive allocation rule $\boldsymbol{\phi}$, that achieves the greatest possible expected value for the sum of the rewards,

$$S_T = \sum_{t=1}^{T} \sum_{a\in[K]} Z_t^a = \sum_{a\in[K]} \sum_{n=1}^{N^a(T)} Y_n^a,$$

which is equivalent to minimizing the expected regret,

$$R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T) = T \sum_{a=1}^{M} \mu(\theta_a) - \mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}[S_T]. \tag{5.1}$$

5.2.3 Asymptotic Lower Bound

A quantity that naturally arises in the study of regret minimization for Markovian bandits is the *Kullback-Leibler divergence rate* between two Markov chains, which is a generalization of the usual Kullback-Leibler divergence between two probability distributions. We denote by $\overline{D}\left(\theta \parallel \lambda\right)$ the Kullback-Leibler divergence rate between the Markov chain with parameter θ and the Markov chain with parameter λ, which is given by,

$$\overline{D}\left(\theta \parallel \lambda\right) = \sum_{x,y\in S} \log \frac{P_\theta(x,y)}{P_\lambda(x,y)} \pi_\theta(x) P_\theta(x,y), \tag{5.2}$$

where we use the standard notational conventions $\log 0 = \infty$, $\log \frac{\alpha}{0} = \infty$ if $\alpha > 0$, and $0 \log 0 = 0 \log \frac{0}{0} = 0$. Indeed note that, if $P_\theta(x,\cdot) = p_\theta(\cdot)$ and $P_\lambda(x,\cdot) = p_\lambda(\cdot)$, for all $x \in S$, i.e. in the special case that the Markov chains correspond to IID processes, then the Kullback-Leibler divergence rate $\overline{D}\left(\theta \parallel \lambda\right)$ is equal to the Kullback-Leibler divergence $D\left(p_\theta \parallel p_\lambda\right)$ between p_θ and p_λ,

$$\overline{D}\left(\theta \parallel \lambda\right) = \sum_{x,y\in S} \log \frac{p_\theta(y)}{p_\lambda(y)} p_\theta(x) p_\theta(y) = \sum_{y\in S} \log \frac{p_\theta(y)}{p_\theta(y)} p_\theta(y) = D\left(p_\theta \parallel p_\lambda\right).$$

Under some regularity assumptions on the one-parameter family of Markov chains, [4] in their Theorem 3.1 are able to establish the following asymptotic lower bound on the expected regret for any adaptive allocation rule $\boldsymbol{\phi}$ which is uniformly good across all parameter configurations,

$$\liminf_{T\to\infty} \frac{R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T)}{\log T} \geq \sum_{b=L+1}^{K} \frac{\mu(\theta_M) - \mu(\theta_b)}{\overline{D}\left(\theta_b \parallel \theta_M\right)}. \tag{5.3}$$

A further discussion of this lower bound, as well as an alternative derivation can be found in Section 5.D,

The main goal of this work is to derive a finite time analysis for an adaptive allocation rule which is based on Kullback-Leibler divergence rate indices, that is asymptotically optimal. We do so for the one-parameter exponential family of Markov chains, which forms a generalization of the classic one-parameter exponential family generated by a probability distribution with finite support.

5.2.4 One-Parameter Exponential Family Of Markov Chains

Let S be a finite state space, $f : S \to \mathbb{R}$ be a nonconstant reward function on the state space, and P an irreducible transition probability matrix on S, with associated stationary distribution π. P will serve as the generator stochastic matrix of the family. Let $\mu(0) = \sum_{x\in S} f(x)\pi(x)$ be the stationary mean of the Markov chain induced by P when f is applied. By tilting exponentially the transitions of P we are able to construct new transition matrices that realize a whole range of stationary means around $\mu(0)$ and form the exponential family of stochastic matrices. Let $\theta \in \mathbb{R}$, and consider the matrix $\tilde{P}_\theta(x, y) = P(x, y)e^{\theta f(y)}$. Denote by $\rho(\theta)$ its spectral radius. According to the Perron-Frobenius theory, see Theorem 8.4.4 in the book of [45], $\rho(\theta)$ is a simple eigenvalue of $\tilde{P}_\theta$, called the Perron-Frobenius eigenvalue, and we can associate to it unique left u_θ and right v_θ eigenvectors such that they are both positive, $\sum_{x\in S} u_\theta(x) = 1$ and $\sum_{x\in S} u_\theta(x)v_\theta(x) = 1$. Using them we define the member of the exponential family which corresponds to the natural parameter θ as,

$$P_\theta(x, y) = \frac{v_\theta(y)}{v_\theta(x)} \exp\left\{\theta f(y) - \Lambda(\theta)\right\} P(x, y), \tag{5.4}$$

where $\Lambda(\theta) = \log \rho(\theta)$ is the log-Perron-Frobenius eigenvalue. It can be easily seen that $P_\theta(x, y)$ is indeed a stochastic matrix, and its stationary distribution is given by $\pi_\theta(x) = u_\theta(x)v_\theta(x)$. The initial distribution q_θ associated to the parameter θ, can be any distribution on S, since the KL-UCB adaptive allocation rule that we devise, and its guarantees, will be valid no matter the initial distributions.

Example 8 (Two-state chain). Let $S = \{0, 1\}$, and consider the transition probability matrix, P, representing two coin-flips, Bernoulli(p) when we're in state 0, and Bernoulli(q) when we're in state 1. We require that P is irreducible, so $p \in (0, 1]$ and $q \in [0, 1)$.

$$P = \begin{bmatrix} 1-p & p \\ 1-q & q \end{bmatrix}$$

The exponential family of transition probability matrices generated by P and $f(x) = 2x - 1$ is given by,

$$P_\theta = \frac{1}{\rho(\theta)} \begin{bmatrix} (1-p)e^{-\theta} & \rho(\theta) - (1-p)e^{-\theta} \\ \rho(\theta) - qe^{\theta} & qe^{\theta} \end{bmatrix},$$

where,

$$\rho(\theta) = \frac{(1-p)e^{-\theta} + qe^{\theta} + \sqrt{((1-p)e^{-\theta} - qe^{\theta})^2 + 4p(1-q)}}{2}.$$

In the special case that $p = q$, we get back the typical exponential family of Bernoulli(p_θ) coin-flips, with

$$1 - p_\theta = \frac{(1-p)e^{-\theta}}{(1-p)e^{-\theta} + pe^{\theta}}.$$

Exponential families of Markov chains date back to the work of [69]. For a short overview of one-parameter exponential families of Markov chains, as well as proofs of the following properties, we refer the reader to Section 2 in [74]. The log-Perron-Frobenius eigenvalue $\Lambda(\theta)$ is a convex analytic function on the real numbers, and through its derivative, $\dot{\Lambda}(\theta)$, we obtain the stationary mean $\mu(\theta)$ of the Markov chain with transition matrix P_θ when f is applied, i.e. $\dot{\Lambda}(\theta) = \mu(\theta) = \sum_{x \in S} f(x)\pi_\theta(x)$. When $\Lambda(\theta)$ is not the linear function $\theta \mapsto \mu(0)\theta$, the log-Perron-Frobenius eigenvalue, $\Lambda(\theta)$, is strictly convex and thus its derivative $\dot{\Lambda}(\theta)$ is strictly increasing, and it forms a bijection between the natural parameter space, $\mathbb{R}$, and the mean parameter space, $\mathcal{M} = \dot{\Lambda}(\mathbb{R})$, which is a bounded open interval.

The Kullback-Leibler divergence rate from (5.2), when instantiated for the exponential family of Markov chains, can be expressed as,

$$\overline{D}\left(\theta \parallel \lambda\right) = \Lambda(\lambda) - \Lambda(\theta) - \dot{\Lambda}(\theta)(\lambda - \theta),$$

which is convex and differentiable over $\mathbb{R} \times \mathbb{R}$. Since $\dot{\Lambda} : \mathbb{R} \to \mathcal{M}$ forms a bijection from the natural parameter space, $\mathbb{R}$, to the mean parameter space, $\mathcal{M}$, with some abuse of notation we will write $\overline{D}\left(\mu \parallel \nu\right)$ for $\overline{D}\left(\dot{\Lambda}^{-1}(\mu) \,\middle\|\, \dot{\Lambda}^{-1}(\nu)\right)$, where $\mu, \nu \in \mathcal{M}$. Furthermore, $\overline{D}\left(\cdot \parallel \cdot\right) : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$ can be extended continuously, to a function $\overline{D}\left(\cdot \parallel \cdot\right) : \bar{\mathcal{M}} \times \bar{\mathcal{M}} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, where $\bar{\mathcal{M}}$ denotes the closure of $\mathcal{M}$. This can even further be extended to a convex function on $\mathbb{R} \times \mathbb{R}$, by setting $\overline{D}\left(\mu \parallel \nu\right) = \infty$ if $\mu \notin \bar{\mathcal{M}}$ or $\nu \notin \bar{\mathcal{M}}$. For fixed $\nu \in \mathbb{R}$, the function $\mu \mapsto \overline{D}\left(\mu \parallel \nu\right)$ is decreasing for $\mu \leq \nu$ and increasing for $\mu \geq \nu$. Similarly, for fixed $\mu \in \mathbb{R}$, the function $\nu \mapsto \overline{D}\left(\mu \parallel \nu\right)$ is decreasing for $\nu \leq \mu$ and increasing for $\nu \geq \mu$.

5.3 A Maximal Inequality for Markov Chains

The following definition is the technical condition that we will require for our maximal inequality.

Definition 2 (Doeblin's type of condition). Let P be a transition probability matrix on the finite state space S. For a nonempty set of states $A \subset S$, we say that P is A-Doeblin if, the

submatrix of P with rows and columns in A is irreducible, and for every $x \in S - A$ there exists $y \in A$ such that $P(x, y) > 0$.

Example 8 (continued). For this example P being $\{0\}$-Doeblin means that $p, q \in [0, 1)$, but already irreducibility imposed the constraints $p \in (0, 1]$ and $q \in [0, 1)$, hence the only additional constraint is $p \neq 1$.

Remark 6. Our Definition 2 is inspired by the classic Doeblin's Theorem, see Theorem 2.2.1 in [86]. Doeblin's Theorem states that, if the transition probability matrix P satisfies Doeblin's condition (namely there exists $\epsilon > 0$, and a state $y \in S$ such that for all $x \in S$ we have $P(x, y) \geq \epsilon$), then P has a unique stationary distribution π, and for all initial distributions q we have geometric convergence to stationarity, i.e. $\|qP^n - \pi\|_1 \leq 2(1 - \epsilon)^n$. Doeblin's condition, according to our Definition 2, corresponds to P being $\{y\}$-Doeblin for some $y \in S$.

Lemma 18 (Maximal inequality for irreducible Markov chains satisfying Doeblin's condition)**.** *Let $\{X_n\}_{n \in \mathbb{Z}_{\geq 0}}$ be an irreducible Markov chain over the finite state space S with transition matrix P, initial distribution q, and stationary distribution π. Let $f : S \to \mathbb{R}$ be a non-constant function on the state space. Denote by $\mu(0) = \sum_{x \in S} f(x)\pi(x)$ the stationary mean when f is applied, and by $\bar{Y}_n = \frac{1}{n}\sum_{k=1}^{n} Y_k$ the empirical mean, where $Y_k = f(X_k)$. Assume that P is $(\arg\min_{x \in S} f(x))$-Doeblin. Then for all $\epsilon > 1$ we have*

$$\mathbb{P}\left(\bigcup_{k=1}^{n} \left\{\mu(0) \geq \bar{Y}_k \text{ and } k\overline{D}\left(\bar{Y}_k \,\middle\|\, \mu(0)\right) \geq \epsilon\right\}\right) \leq C_{-} e^{\lceil \epsilon \log n \rceil} e^{-\epsilon},$$

where $C_{-} = C_{-}(P, f)$ is a positive constant depending only on the transition probability matrix P and the function f.

Remark 7. If we only consider values of ϵ from a bounded subset of $(1, \infty)$, then we don't need to assume that P is $(\arg\min_{x \in S} f(x))$-Doeblin, and the constant C_{-} will further depend on this bounded subset. But in the analysis of the KL-UCB adaptive allocation rule we will need to consider values of ϵ that increase with the time horizon T, therefore we have to impose the assumption that P is $(\arg\min_{x \in S} f(x))$-Doeblin, so that C_{-} has no dependencies on ϵ.

i.i.d. versions of this maximal inequality have found applicability not only in multi-armed bandit problems, but also in the case of context tree estimation, [35], indicating that our Lemma 18 may be of interest for other applications as well.

5.4 The Round-Robin KL-UCB Adaptive Allocation Rule for Multiple Plays and Markovian Rewards

For each arm $a \in [K]$ we define the empirical mean at the global time t as,

$$\bar{Y}_a(t) = (Y_1^a + \ldots + Y_{N_a(t)}^a)/N_a(t), \tag{5.5}$$

and its local time counterpart as,

$$\bar{Y}_n^a = (Y_1^a + \ldots + Y_n^a)/n,$$

with their link being $\bar{Y}_n^a = \bar{Y}_a(\tau_n^a)$, where $\tau_n^a = \inf\{t \geq 1 : N_a(t) = n\}$. At each round t we calculate a single upper confidence bound index,

$$U_a(t) = \sup\left\{\mu \in \mathcal{M} : D\left(\bar{Y}_a(t) \,\middle\|\, \mu\right) \leq \frac{g(t)}{N_a(t)}\right\}, \tag{5.6}$$

where $g(t)$ is an increasing function, and we denote its local time version by,

$$U_n^a(t) = \sup\left\{\mu \in \mathcal{M} : D\left(\bar{Y}_n^a \,\middle\|\, \mu\right) \leq \frac{g(t)}{n}\right\}.$$

Note that $U_a(t)$ is efficiently computable via a bisection method due to the monotonicity of $D\left(\bar{Y}_a(t) \,\middle\|\, \cdot\right)$. It is straightforward to check, using the definition of $U_n^a(t)$, the following two relations,

$$\bar{Y}_n^a \leq U_n^a(t) \text{ for all } n \leq t, \tag{5.7}$$

$$U_n^a(t) \text{ is increasing in } t \geq n \text{ for fixed } n. \tag{5.8}$$

Furthermore, in Section 5.B we study the concentration properties of those upper confidence indices and of the sample means, using the concentration results for Markov chains from Section 5.3. The idea of calculating indices in a round robin way, dates back to the seminal work of [56]. Here we exploit this idea, which seems to have been forgotten over time in favor of algorithms that calculate indices for all the arms at each round, and we augment it with the usage of the upper confidence bounds in (5.6), which are efficiently computable, see Section 5.6 for simulation results, as opposed to the statistics in Theorem 4.1 from [4]. Moreover, this combination of a round-robin scheme and the indices in (5.6) is amenable to

a finite-time analysis, see Section 5.C.

Algorithm 3: The round-robin KL-UCB adaptive allocation rule.

Parameters: *number of arms* $K \geq 2$*, time horizon* $T \geq K$*, number of plays* $1 \leq M \leq K$*,*
KL divergence rate function $\overline{D}(\cdot \,\|\, \cdot) : \bar{\mathcal{M}} \times \bar{\mathcal{M}} \to \mathbb{R}_{\geq 0}$*, increasing function* $g : \mathbb{Z}_{>0} \to \mathbb{R}$*, parameter* $\delta \in (0, 1/K)$*;*
Initializaton: *In the first* K *rounds pull each arm* M *times and set* $\bar{Y}_a(K) = (Y_1^a + \ldots + Y_M^a)/M$*, for* $a = 1, \ldots, K$*;*
for $t = K, \ldots, T-1$ **do**
  Let $W_t = \{a \in [K] : N_a(t) \geq \lceil \delta t \rceil\}$;
  Pick any subset of arms $L_t \subseteq W_t$ such that:
  - $|L_t| = M$;
  - and $\min_{a \in L_t} \bar{Y}_a(t) \geq \sup_{b \in W_t - L_t} \bar{Y}_b(t)$;

  Let $b \equiv t+1 \pmod{K}$, with $b \in [K]$;
  Let $U_b(t) = \sup\left\{\mu \in \mathcal{M} : \overline{D}\left(\bar{Y}_b(t) \,\middle\|\, \mu\right) \leq \frac{g(t)}{N_b(t)}\right\}$;
  if $b \in L_t$ ***or*** $\min_{a \in L_t} \bar{Y}_a(t) \geq U_b(t)$ **then**
    Pull the M arms in $\phi_{t+1} = L_t$;
  else
    Pick any $a \in \arg\min_{a \in L_t} \bar{Y}_a(t)$;
    Pull the M arms in $\phi_{t+1} = (L_t \setminus \{a\}) \cup \{b\}$;
  end
end

Proposition 4. *For each* $t \geq K$ *we have that* $|W_t| \geq M$*, and so Algorithm 3 is well defined.*

Theorem 15 (Markovian rewards and multiple plays: finite-time guarantees)**.** *Let* P *be an irreducible transition probability matrix on the finite state space* S*, and* $f : S \to \mathbb{R}$ *be a real-valued reward function, such that* P *is* $(\arg\min_{x \in S} f(x))$*-Doeblin. Assume that the* K *arms correspond to the parameter configuration* $\boldsymbol{\theta} \in \mathbb{R}^K$ *of the exponential family of Markov chains, as described in Equation 5.4. Without loss of generality assume that the* K *arms are ordered so that,*

$$\mu(\theta_1) \geq \ldots \geq \mu(\theta_N) > \mu(\theta_{N+1}) \ldots = \mu(\theta_M) = \ldots = \mu(\theta_L) > \mu(\theta_{L+1}) \geq \ldots \geq \mu(\theta_K).$$

Fix $\epsilon \in (0, \min(\mu(\theta_N) - \mu(\theta_M), \mu(\theta_M) - \mu(\theta_{L+1})))$*. The KL-UCB adaptive allocation rule for Markovian rewards and multiple plays, Algorithm 3, with the choice* $g(t) = \log t + 3 \log\log t$*,*

enjoys the following finite-time upper bound on the regret,

$$R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T) \leq \sum_{b=L+1}^{K} \frac{\mu(\theta_M) - \mu(\theta_b)}{D\left(\mu(\theta_b) \parallel \mu(\theta_M) - \epsilon\right)} \log T + c_1 \sqrt{\log T} + c_2 \log\log T + c_3 \sqrt{\log\log T} + c_4,$$

where c_1, c_2, c_3, c_4 are constants with respect to T, which are given more explicitly in the analysis.

Corollary 3 (Asymptotic optimality)**.** *In the context of Theorem 15 the KL-UCB adaptive allocation rule, Algorithm 3, is asymptotically optimal, and,*

$$\lim_{T\to\infty} \frac{R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T)}{\log T} = \sum_{b=L+1}^{K} \frac{\mu(\theta_M) - \mu(\theta_b)}{D\left(\mu(\theta_b) \parallel \mu(\theta_M)\right)}.$$

5.5 The Round-Robin KL-UCB Adaptive Allocation Rule for Multiple Plays and i.i.d. Rewards

As a byproduct of our work in Section 5.4 we further obtain a finite-time regret bound, which is asymptotically optimal, for the case of multiple plays and i.i.d. rewards, from an exponential family of probability densities.

We first review the notion of an exponential family of probability densities, for which the standard reference is [14]. Let $(X, \mathcal{X}, \rho)$ be a probability space. A one-parameter exponential family is a family of probability densities $\{p_\theta : \theta \in \Theta\}$ with respect to the measure ρ on X, of the form,

$$p_\theta(x) = \exp\{\theta f(x) - \Lambda(\theta)\} h(x), \tag{5.9}$$

where $f : X \to \mathbb{R}$ is called the sufficient statistic, is $\mathcal{X}$-measurable, and there is no $c \in \mathbb{R}$ such that $f(x) \overset{\rho-a.s.}{=} c$, $h : X \to \mathbb{R}_{\geq 0}$ is called the carrier density, and is a density with respect to ρ, and Λ is the log-Moment-Generating-Function and is given by $\Lambda(\theta) = \log \int_X e^{\theta f(x)} h(x) \rho(dx)$, which is finite for θ in the natural parameter space $\Theta = \{\theta \in \mathbb{R} : \int_X e^{\theta f(x)} h(x) \rho(dx) < \infty\}$. The log-MGF, $\Lambda(\theta)$, is strictly convex and its derivative forms a bijection between the natural parameters, θ, and the mean parameters, $\mu(\theta) = \int_X f(x) p_\theta(x) \rho(dx)$. The Kullback-Leibler divergence between p_θ and p_λ, for $\theta, \lambda \in \Theta$, can be written as $D\left(\theta \parallel \lambda\right) = \Lambda(\lambda) - \Lambda(\theta) - \dot{\Lambda}(\theta)(\lambda - \theta)$.

For this section, each arm $a \in [K]$ with parameter θ_a corresponds to the i.i.d. process $\{X_n^a\}_{n \in \mathbb{Z}_{>0}}$, where each X_n^a has density p_{θ_a} with respect to ρ, which gives rise to the i.i.d. reward process $\{Y_n^a\}_{n \in \mathbb{Z}_{>0}}$, with $Y_n^a = f(X_n^a)$.

Remark 8. When there is a finite set $S \in \mathcal{X}$ such that $\rho(S) = 1$, then the exponential family of probability densities in Equation 5.9, is just a special case of the exponential family of Markov chains in Equation 5.4, as can be seen by setting $P(x, \cdot) = h(\cdot)$, for all $x \in S$. Then $v_\theta(x) = 1$ for all $x \in S$, the log-Perron-Frobenius eigenvalue coincides with the log-MGF,

and $\Theta = \mathbb{R}$. Therefore, Theorem 15 already resolves the case of multiple plays and i.i.d. rewards from an exponential family of finitely supported densities.

Theorem 16 (i.i.d. rewards and multiple plays: finite-time guarantees)**.** *Let $(X, \mathcal{X}, \rho)$ be a probability space, $f : X \to \mathbb{R}$ a $\mathcal{X}$-measurable function, and $h : X \to \mathbb{R}_{\geq 0}$ a density with respect to ρ. Assume that the K arms correspond to the parameter configuration $\boldsymbol{\theta} \in \Theta^K$ of the exponential family of probability densities, as described in Equation 5.9. Without loss of generality assume that the K arms are ordered so that,*

$$\mu(\theta_1) \geq \ldots \geq \mu(\theta_N) > \mu(\theta_{N+1}) \ldots = \mu(\theta_M) = \ldots = \mu(\theta_L) > \mu(\theta_{L+1}) \geq \ldots \geq \mu(\theta_K).$$

Fix $\epsilon \in (0, \min(\mu(\theta_N) - \mu(\theta_M), \mu(\theta_M) - \mu(\theta_{L+1})))$. The KL-UCB adaptive allocation rule for i.i.d. rewards and multiple plays, Algorithm 3, with the choice $g(t) = \log t + 3 \log \log t$, enjoys the following finite-time upper bound on the regret,

$$R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T) \leq \sum_{b=L+1}^{K} \frac{\mu(\theta_M) - \mu(\theta_b)}{D\left(\mu(\theta_b) \parallel \mu(\theta_M) - \epsilon\right)} \log T + c_1 \sqrt{\log T} + c_2 \log \log T + c_3 \sqrt{\log \log T} + c_4,$$

where c_1, c_2, c_3, c_4 are constants with respect to T.

Consequently, the KL-UCB adaptive allocation rule, Algorithm 3, is asymptotically optimal, and,

$$\lim_{T \to \infty} \frac{R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T)}{\log T} = \sum_{b=L+1}^{K} \frac{\mu(\theta_M) - \mu(\theta_b)}{D\left(\mu(\theta_b) \parallel \mu(\theta_M)\right)}.$$

Remark 9. For the special case of single plays, $M = 1$, such a finite-time regret bound is derived in [16], and here we generalize it for multiple plays, $1 \leq M \leq K$. One striking difference is that we consider calculations of KL upper confidence bounds in a round-robin way, as opposed to calculating them for all the arms at each round. But computing KL-UCB indices adds an extra computational overhead, as it entails inverting an increasing function via the bisection method. Thus, our approach has important practical implications as it leads to significantly more efficient algorithms. We verify this via simulations in Section 5.6.

5.6 Simulation Results

In the context of Example 8, we set $p = 0.49, q = 0.45, K = 14$, and $T = 10^6$. We generated the bandit instance $\theta_1, \ldots, \theta_K$ by drawing i.i.d. $N(0, 1/16)$ samples. Four adaptive allocation rules were taken into consideration:

1. **UCB**: at reach round calculate all UCB indices,

$$U_a^{\mathrm{UCB}}(t) = \bar{Y}_a(t) + \beta \sqrt{\frac{2 \log t}{N_a(t)}}, \text{ for } a = 1, \ldots, K.$$

2. **Round-Robin UCB**: at reach round calculate a single UCB index,

$$U_b^{\mathrm{UCB}}(t) = \bar{Y}_b(t) + \beta\sqrt{\frac{2\log t}{N_b(t)}}, \text{ only for } b \equiv t+1 \pmod{K}.$$

3. **KL-UCB**: at reach round calculate all KL-UCB indices,

$$U_a^{\mathrm{KL-UCB}}(t) = \sup\left\{\mu \in \mathcal{M} : D\left(\bar{Y}_a(t) \,\|\, \mu\right) \leq \frac{\log t + 3\log\log t}{N_a(t)}\right\}, \text{ for } a = 1, \ldots, K.$$

4. **Round-Robin KL-UCB**: at reach round calculate a single KL-UCB index,

$$U_b^{\mathrm{KL-UCB}}(t) = \sup\left\{\mu \in \mathcal{M} : D\left(\bar{Y}_b(t) \,\|\, \mu\right) \leq \frac{\log t + 3\log\log t}{N_b(t)}\right\}, \text{ only for } b \equiv t+1 \pmod{K}.$$

For the UCB indices, after some tuning, we picked $\beta = 1$ which is significantly smaller than the theoretical values of β from [87, 88, 70]. For each of those adaptive allocation rules 10^4 Monte Carlo iterations were performed in order to estimate the expected regret, and the simulation results are presented in the following plots.

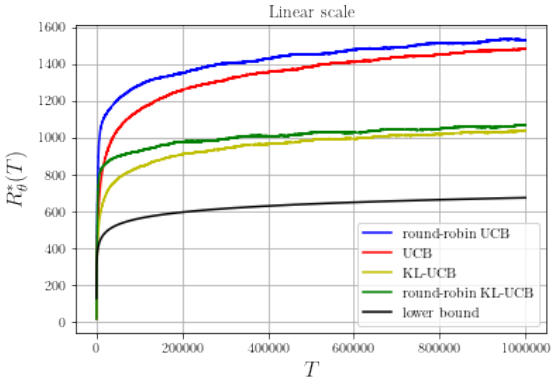


Figure 5.6.1: Regret of the various algorithms as a function of time in linear scale.

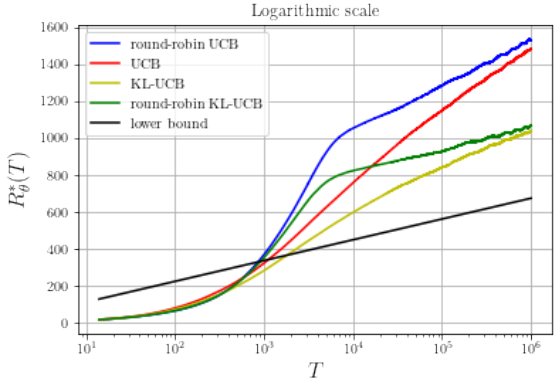


Figure 5.6.2: Regret of the various algorithms as a function of time in logarithmic scale.

For our simulations we used the programming language C, to produce highly efficient code, and a personal computer with a 2.6GHz processor and 16GB of memory. We report that the simulation for the Round-Robin KL-UCB adaptive allocation rule was 14.48 times faster than the simulation for the KL-UCB adaptive allocation rule. This behavior is expected since each calculation of a KL-UCB index induces a significant computation cost as it involves

finding the inverse of an increasing function using the bisection method. Additionally, the simulation for the Round-Robin UCB adaptive allocation rule was 3.15 times faster than the simulation for the KL-UCB adaptive allocation rule, and this is justified from the fact that calculating mathematical functions such as $\log(\cdot)$ and $\sqrt{\cdot}$, is more costly than calculating averages which only involve a division. Our simulation results yield that in practice round-robin schemes are significantly faster than schemes that calculate the indices of all the arms at each round, and the computational gap is increasing with the number of arms K, while the behavior of the expected regrets is very similar.

Appendix

5.A Concentration Lemmata for Markov Chains

We first develop a Chernoff bound, which remarkably does not impose any conditions on the Markov chain other than irreducibility, which is though a mandatory requirement for the stationary mean to be well-defined.

Lemma 19 (Chernoff bound for irreducible Markov chains)**.** *Let $\{X_n\}_{n\in\mathbb{Z}_{\geq 0}}$ be an irreducible Markov chain over the finite state space S with transition probability matrix P, initial distribution q, and stationary distribution π. Let $f : S \to \mathbb{R}$ be a nonconstant function on the state space. Denote by $\mu(0) = \sum_{x\in S} f(x)\pi(x)$ the stationary mean when f is applied, and by $\bar{Y}_n = \frac{1}{n}\sum_{k=1}^{n} Y_k$ the empirical mean, where $Y_k = f(X_k)$. Let F be a closed subset of $\mathcal{M} \cap [\mu(0), \infty)$. Then,*

$$\mathbb{P}\left(\bar{Y}_n \geq \mu\right) \leq C_+ e^{-n\overline{D}(\mu \,\|\, \mu(0))}, \text{ for } \mu \in F,$$

where $\overline{D}\left(\cdot \,\|\, \cdot\right)$ stands for the Kullback-Leibler divergence rate in the exponential family of stochastic matrices generated by P and f, and $C_+ = C_+(P, f, F)$ is a positive constant depending only on the transition probability matrix P, the function f and the closed set F.

Proof of Lemma 19.
Using the standard exponential transform followed by Markov's inequality we obtain that for any $\theta \geq 0$,

$$\mathbb{P}(\bar{Y}_n \geq \mu) \leq \mathbb{P}(e^{n\theta\bar{Y}_n} \geq e^{n\theta\mu}) \leq \exp\left\{-n\left(\theta\mu - \frac{1}{n}\log\mathbb{E}\left[e^{\theta(f(X_1)+\ldots+f(X_n))}\right]\right)\right\}.$$

We can upper bound the expectation from above in the following way,

$$\begin{aligned}
\mathbb{E}\left[e^{\theta(f(X_1)+\ldots+f(X_n))}\right] &= \sum_{x_0,\ldots,x_n\in S} q(x_0)P(x_0,x_1)e^{\theta f(x_1)}\ldots P(x_{n-1},x_n)e^{\theta f(x_n)} \\
&= \sum_{x_0,x_n\in S} q(x_0)\tilde{P}_\theta^n(x_0,x_n) \\
&\leq \frac{1}{\min_{x\in S} v_\theta(x)} \sum_{x_0,x_n\in S} q(x_0)\tilde{P}_\theta^n(x_0,x_n)v_\theta(x_n) \\
&= \frac{\rho(\theta)^n}{\min_{x\in S} v_\theta(x)} \sum_{x_0\in S} q(x_0)v_\theta(x_0) \\
&\leq \max_{x,y\in S} \frac{v_\theta(y)}{v_\theta(x)}\rho(\theta)^n,
\end{aligned}$$

where in the last equality we used the fact that v_θ is a right Perron-Frobenius eigenvector of $\tilde{P}_\theta$.

From those two we obtain,

$$\mathbb{P}(\bar{Y}_n \geq \mu) \leq \max_{x,y\in S} \frac{v_\theta(y)}{v_\theta(x)} \exp\left\{-n(\theta\mu - \Lambda(\theta))\right\},$$

and if we plug in $\theta_\mu = \dot{\Lambda}^{-1}(\mu)$, which is a nonnegative real number since $\mu \in F \subseteq \mathcal{M} \cap [\mu(0),\infty)$, we obtain,

$$\mathbb{P}(\bar{Y}_n \geq \mu) \leq \max_{x,y\in S} \frac{v_{\theta_\mu}(y)}{v_{\theta_\mu}(x)} \exp\left\{-n\overline{D}\left(\mu \parallel \mu(0)\right)\right\},$$

We assumed that F is closed, and moreover F is bounded since it is a subset of the bounded open interval $\mathcal{M}$. Therefore, F is compact, and so $\dot{\Lambda}^{-1}(F)$ is compact as well. Then due to the fact that $\theta \mapsto v_\theta(x)/v_\theta(y)$ is continuous, from Lemma 2 in [74], we deduce that,

$$\sup_{\theta\in\dot{\Lambda}^{-1}(F)} \max_{x,y\in S} \frac{v_\theta(y)}{v_\theta(x)} < \infty,$$

which we define to be the finite constant C_+ of Lemma 19, and which may only depend on P, f and F. □

Remark 10. This bound is a variant of Theorem 1 in [74], where the authors derive a Chernoff bound under some structural assumptions on the transition probability matrix P and the function f. In our Lemma 19, following their techniques, we derive a Chernoff bound without any assumptions, relying though on the fact that μ lies in a closed subset of the mean parameter space.

Next, we proceed with the proof of the maximal inequality in Section 5.3.

Proof of Lemma 18.
Our proof extends the argument from Lemma 11 in [16], which deals with IID random variables. In order to handle the Markovian dependence we need to use the exponential martingale for Markov chains from Lemma 11, as well as continuity results for the right Perron-Frobenius eigenvector.

Following the proof strategy used to establish the law of the iterated logarithm, we split the range of the union $[n]$ into chunks of exponentially increasing sizes. Denote by $\alpha > 1$ the growth factor, to be specified later, and let $n_m = \lfloor \alpha^m \rfloor$ be the end point of the m-th chunk, with $n_0 = 0$. An upper bound on the number of chunks is $M = \lceil \log n / \log \alpha \rceil$, and so we have that

$$\begin{aligned}\bigcup_{k=1}^{n} \left\{\mu(0) \geq \bar{Y}_k,\ k\overline{D}\left(\bar{Y}_k \,\middle\|\, \mu(0)\right) \geq \epsilon\right\} &\subseteq \bigcup_{m=1}^{M} \bigcup_{k=n_{m-1}+1}^{n_m} \left\{\mu(0) \geq \bar{Y}_k, k\overline{D}\left(\bar{Y}_k \,\middle\|\, \mu(0)\right) \geq \epsilon\right\} \\ &\subseteq \bigcup_{m=1}^{M} \bigcup_{k=n_{m-1}+1}^{n_m} \left\{\mu(0) \geq \bar{Y}_k, \overline{D}\left(\bar{Y}_k \,\middle\|\, \mu(0)\right) \geq \frac{\epsilon}{n_m}\right\}.\end{aligned}$$

Let $\mu_m = \inf\{\mu < \mu(0) : D(\mu \,\|\, \mu(0)) \leq \epsilon/n_m\}$, and $\theta_m = \dot{\Lambda}^{-1}(\mu_m) < \dot{\Lambda}^{-1}(\mu(0)) = 0$ so that $\theta_m \mu_m - \Lambda(\theta_m) = D(\mu_m \,\|\, \mu(0))$. Then,

$$\begin{aligned}\left\{\mu(0) \geq \bar{Y}_k,\ D\left(\bar{Y}_k \,\middle\|\, \mu(0)\right) \geq \frac{\epsilon}{n_m}\right\} &\subseteq \left\{\bar{Y}_k \leq \mu_m\right\} \\ &= \left\{e^{\theta_m k\bar{Y}_k - k\Lambda(\theta_m)} \geq e^{k(\theta_m\mu_m - \Lambda(\theta_m))}\right\} \\ &= \left\{M_k^{\theta_m} \geq \frac{v_{\theta_m}(X_k)}{v_{\theta_m}(X_0)} e^{kD(\mu_m \,\|\, \mu(0))}\right\} \\ &\subseteq \left\{M_k^{\theta_m} \geq \frac{v_{\theta_m}(X_k)}{v_{\theta_m}(X_0)} e^{(n_{m-1}+1)D(\mu_m \,\|\, \mu(0))}\right\}.\end{aligned}$$

At this point we use the assumption that P is $(\arg\min_{x\in S} f(x))$-Doeblin in order to invoke Proposition 1 from [74], which in our setting states that there exists a constant $C_- = C_-(P, f) \geq 1$ such that,

$$\frac{1}{C_-} \leq \inf_{\theta \in \mathbb{R}_{\leq 0}, x, y \in S} \frac{v_\theta(y)}{v_\theta(x)}.$$

This gives us the inclusion,

$$\left\{M_k^{\theta_m} \geq \frac{v_{\theta_m}(X_k)}{v_{\theta_m}(X_0)} e^{(n_{m-1}+1)D(\mu_m \,\|\, \mu(0))}\right\} \subseteq \left\{M_k^{\theta_m} \geq \frac{e^{(n_{m-1}+1)D(\mu_m \,\|\, \mu(0))}}{C_-}\right\}.$$

In Lemma 11 we have established that $M_k^{\theta_m}$ is a positive martingale, which combined with a maximal inequality for martingales due to [91] (see Exercise 4.8.2 in [29] for a modern

reference), yields that,

$$\mathbb{P}\left(\bigcup_{k=n_{m-1}+1}^{n_m}\left\{M_k^{\theta_m} \geq \frac{e^{(n_{m-1}+1)D(\mu_m \parallel \mu(0))}}{C_-}\right\}\right) \leq C_- e^{-(n_{m-1}+1)D(\mu_m \parallel \mu(0))}$$
$$\leq C_- e^{-\epsilon\frac{n_{m-1}+1}{n_m}} \leq C_- e^{-\frac{\epsilon}{\alpha}}.$$

To conclude, we pick the growth factor $\alpha = \epsilon/(\epsilon-1)$, and we upper bound the number of chunks by $M \leq \lceil \epsilon \log n \rceil$. □

5.B Concentration Properties of Upper Confidence Bounds and Sample Means

Lemma 20. *For every arm $a = 1, \ldots, K$, and $t \geq 3$, we have that,*

$$\mathbb{P}_{\theta_a}\left(\min_{n=1,\ldots,t} U_n^a(t) \leq \mu(\theta_a)\right) \leq \frac{4eC_-^a}{t\log t}, \tag{5.10}$$

where C_-^a is the constant prescribed in Lemma 18, when the maximal inequality is applied to the Markov chain with parameter θ_a.

Proof.

$$\mathbb{P}_{\theta_a}\left(\min_{n=1,\ldots,t} U_n^a(t) \leq \mu(\theta_a)\right) \leq \mathbb{P}_{\theta_a}\left(\bigcup_{n=1}^{t}\{\mu(\theta_a) > \bar{Y}_n^a \text{ and } nD\left(\bar{Y}_n^a \parallel \mu(\theta_a)\right) \geq g(t)\}\right)$$
$$\leq C_-^a e\lceil g(t)\log t\rceil e^{-g(t)} \leq 4C_-^a e(\log t)^2 e^{-g(t)} = \frac{4eC_-^a}{t\log t},$$

where for the first inequality we used Equation 5.7 and the definition of $U_n^a(t)$, while for the second inequality we used Lemma 18. □

Lemma 21. *For every arm $a = 1, \ldots, K$, and for $\mu(\lambda) > \mu(\theta_a)$,*

$$\sum_{n=1}^{\infty}\mathbb{P}_{\theta_a}(\mu(\lambda) \leq U_n^a(T)) \leq \frac{g(T)}{D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)} + 1 + 8\sigma_{\theta_a,\lambda}^2\left(\frac{\dot{D}\left(\mu(\theta_a) \parallel \mu(\lambda)\right)}{\overline{D}\left(\mu(\theta_a) \parallel \mu(\lambda)\right)}\right)^2 \tag{5.11}$$
$$+ 2\sqrt{2\pi\sigma_{\theta_a,\lambda}^2}\sqrt{\frac{\dot{D}\left(\mu(\theta_a) \parallel \mu(\lambda)\right)^2}{\overline{D}\left(\mu(\theta_a) \parallel \mu(\lambda)\right)^3}}\sqrt{g(T)},$$

where $\sigma_{\theta,\lambda}^2 = \sup_{\theta\in[\theta_a,\lambda]}\ddot{\Lambda}(\theta) \in (0,\infty)$, and $\dot{D}\left(\mu(\theta_a) \parallel \mu(\lambda)\right) = \frac{dD(\mu \parallel \mu(\lambda))}{d\mu}\,|_{\mu=\mu(\theta_a)}$.

Proof. The proof is based on the argument given in Appendix A.2 of [16], adapted though for the case of Markov chains. If $\mu(\lambda) \leq U_n^a(T)$, and $\bar{Y}_n^a \leq \mu(\lambda)$, then $D\left(\bar{Y}_n^a \,\middle\|\, \mu(\lambda)\right) \leq g(T)/n$. Let $\mu_x = \inf\{\mu \leq \mu(\lambda) : D\left(\mu \,\|\, \mu(\lambda)\right) \leq x\}$. This in turn implies that $D\left(\bar{Y}_n^a \,\middle\|\, \mu(\lambda)\right) \leq D\left(\mu_{g(T)/n} \,\middle\|\, \mu(\lambda)\right)$, and using the monotonicity of $\mu \mapsto D\left(\mu \,\|\, \mu(\lambda)\right)$ for $\mu \leq \mu(\lambda)$, we further have that $\bar{Y}_n^a \geq \mu_{g(T)/n}$. This argument shows that,

$$\mathbb{P}_{\theta_a}(\mu(\lambda) \leq U_n^a(T)) \leq \mathbb{P}_{\theta_a}(\mu_{g(T)/n} \leq \bar{Y}_n^a).$$

Therefore,

$$\sum_{n=1}^{\infty} \mathbb{P}_{\theta_a}(\mu(\lambda) \leq U_n^a(T)) \leq \frac{g(T)}{D\left(\mu(\theta_a) \,\|\, \mu(\lambda)\right)} + 1 + \sum_{n=n_0+1}^{\infty} \mathbb{P}_{\theta_a}(\mu_{g(T)/n} \leq \bar{Y}_n^a),$$

where $n_0 = \left\lceil \frac{g(T)}{D(\mu(\theta_a) \,\|\, \mu(\lambda))} \right\rceil$.

Fix $n \geq n_0 + 1$. Then $D\left(\mu(\theta_a) \,\|\, \mu(\lambda)\right) > g(T)/n$, and therefore $\mu_{g(T)/n} > \mu(\theta_a)$. Furthermore note that $\mu_{g(T)/n}$ is increasing to $\mu(\lambda)$ as n increases, therefore $\mu_{g(T)/n}$ lives in the closed interval $[\mu(\theta_a), \mu(\lambda)]$, and we can apply Lemma 19 for the Markov chain that corresponds to the parameter θ_a,

$$\mathbb{P}_{\theta_a}(\bar{Y}_n^a \geq \mu_{g(T)/n}) \leq C_+^a e^{-nD\left(\mu_{g(T)/n} \,\|\, \mu(\theta_a)\right)}.$$

Thus we are left with the task of controlling the sum,

$$\sum_{n=n_0+1}^{\infty} e^{-nD\left(\mu_{g(T)/n} \,\|\, \mu(\theta_a)\right)}.$$

First note that by definition $\mu_{g(T)/n}$ is increasing in n, therefore $D\left(\mu_{g(T)/n} \,\middle\|\, \mu(\theta_a)\right)$ is positive and increasing in n, hence we can perform the following integral bound,

$$\begin{aligned}\sum_{n=n_0+1}^{\infty} e^{-nD\left(\mu_{g(T)/n} \,\|\, \mu(\theta_a)\right)} &\leq \int_{\frac{g(T)}{D(\mu(\theta_a) \,\|\, \mu(\lambda))}}^{\infty} e^{-sD\left(\mu_{g(T)/s} \,\|\, \mu(\theta_a)\right)} ds \\ &= g(T) \int_0^{D(\mu(\theta_a) \,\|\, \mu(\lambda))} \frac{1}{x^2} e^{-\frac{g(T)}{x} D(\mu_x \,\|\, \mu(\theta_a))} dx. \end{aligned} \tag{5.12}$$

The function $\mu \mapsto D\left(\mu \,\|\, \mu(\lambda)\right)$ is convex thus,

$$D\left(\mu \,\|\, \mu(\lambda)\right) \geq D\left(\mu(\theta_a) \,\|\, \mu(\lambda)\right) + \dot{D}\left(\mu(\theta_a) \,\|\, \mu(\lambda)\right)(\mu - \mu(\theta_a)),$$

where $\dot{D}\left(\mu(\theta_a) \,\|\, \mu(\lambda)\right) = \frac{dD(\mu \,\|\, \mu(\lambda))}{d\mu} \,|_{\mu=\mu(\theta_a)}$. Plugging in $\mu = \mu_x \geq \mu(\theta_a)$, for $x \in [0, D\left(\mu(\theta_a) \,\|\, \mu(\lambda)\right)]$, we obtain

$$D\left(\mu(\theta_a) \,\|\, \mu(\lambda)\right) - x \leq \dot{D}\left(\mu(\theta_a) \,\|\, \mu(\lambda)\right)(\mu(\theta_a) - \mu_x). \tag{5.13}$$

From Lemma 8 in [74] we have that,

$$D\left(\mu_x \parallel \mu(\theta_a)\right) \geq \frac{(\mu_x - \mu(\theta_a))^2}{2\sigma^2_{\theta_a,\lambda}}, \tag{5.14}$$

where $\sigma^2_{\theta_a,\lambda} = \sup_{\theta\in[\theta_a,\lambda]} \ddot{\Lambda}(\theta) \in (0,\infty)$.

Combining Equation 5.13 and Equation 5.14 we deduce that,

$$D\left(\mu_x \parallel \mu(\theta_a)\right) \geq \left(\frac{D\left(\mu(\theta_a) \parallel \mu(\lambda)\right) - x}{\sqrt{2}\sigma_{\theta_a,\lambda}\dot{D}\left(\mu(\theta_a) \parallel \mu(\lambda)\right)}\right)^2.$$

Now we use this bound and break the integral in Equation 5.12 in two regions, $I_1 = [0, D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)/2]$ and $I_2 = [D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)/2, D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)]$. In the first region we use the fact that $x \leq D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)/2$ to deduce that,

$$\begin{aligned}\int_{I_1} \frac{1}{x^2} e^{-\frac{g(T)}{x} D(\mu_x \parallel \mu(\theta_a))} dx &\leq \int_{I_1} \frac{1}{x^2} \exp\left\{-\frac{g(T)}{8\sigma^2_{\theta_a,\lambda} x}\left(\frac{D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)}{\dot{D}\left(\mu(\theta_a) \parallel \mu(\lambda)\right)}\right)^2\right\} dx \\ &\leq \frac{8\sigma^2_{\theta_a,\lambda}}{g(T)}\left(\frac{\dot{D}\left(\mu(\theta_a) \parallel \mu(\lambda)\right)}{D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)}\right)^2.\end{aligned}$$

In the second region we use the fact that $D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)/2 \leq x \leq D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)$ to deduce that,

$$\begin{aligned}\int_{I_2} \frac{1}{x^2} e^{-\frac{g(T)}{x} D(\mu_x \parallel \mu(\theta_a))} dx &\leq \int_{I_2} \frac{4\exp\left\{-\frac{(x - D(\mu(\theta_a) \parallel \mu(\lambda)))^2}{2\Sigma_{\theta_a,\lambda}}\right\}}{D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)^2} dx \\ &\leq \int_{-\infty}^{D(\mu(\theta_a) \parallel \mu(\lambda))} \frac{4\exp\left\{-\frac{(x - D(\mu(\theta_a) \parallel \mu(\lambda)))^2}{2\Sigma_{\theta_a,\lambda}}\right\}}{D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)^2} dx \\ &= \frac{2\sqrt{2\pi\sigma^2_{\theta_a,\lambda}}}{\sqrt{g(T)}} \sqrt{\frac{\dot{D}\left(\mu(\theta_a) \parallel \mu(\lambda)\right)^2}{D\left(\mu(\theta_a) \parallel \mu(\lambda)\right)^3}},\end{aligned}$$

where $\Sigma_{\theta_a,\lambda} = \frac{\sigma^2_{\theta_a,\lambda}\dot{D}(\mu(\theta_a) \parallel \mu(\lambda))^2 D(\mu(\theta_a) \parallel \mu(\lambda))}{g(T)}$. □

Lemma 22. *For every arm $a = 1, \ldots, K$,*

$$\mathbb{P}_{\theta_a}\left(\max_{n=\lceil \delta t\rceil,\ldots,t} |\bar{Y}^a_n - \mu(\theta_a)| \geq \epsilon\right) \leq \frac{c\eta^{\delta t}}{1-\eta}, \text{ for } \delta \in (0,1),\ \epsilon > 0, \tag{5.15}$$

where $\eta = \eta(\boldsymbol{\theta}, \epsilon) \in (0,1)$, and $c = c(\boldsymbol{\theta}, \epsilon)$ are constants with respect to t.

Proof. Using the same technique as in the proof of Lemma 19, we have that for any $\theta \geq 0$ and any $\eta \leq 0$,

$$\mathbb{P}_{\theta_a}\left(\max_{n=\lceil \delta t\rceil,\dots,t} |\bar{Y}_n^a - \mu(\theta_a)| \geq \epsilon\right) \leq \sum_{n=\lceil \delta t\rceil}^{\infty} \max_{x,y\in S} \frac{v_\theta^a(y)}{v_\theta^a(x)} e^{-n(\theta(\mu(\theta_a)+\epsilon)-\Lambda_a(\theta))} + \sum_{n=\lceil \delta t\rceil}^{\infty} \max_{x,y\in S} \frac{v_\eta^a(y)}{v_\eta^a(x)} e^{-n(\eta(\mu(\theta_a)-\epsilon)-\Lambda_a(\eta))},$$

where by $\Lambda_a(\theta)$ we denote the log-Perron-Frobenious eigenvalue generated by P_{θ_a}, and similarly by v_θ^a the corresponding right Perron-Frobenius eigenvector.

By picking $\theta = \theta_\epsilon^a$ large enough, and $\eta = \eta_\epsilon^a$ small enough, we can ensure that $\theta(\mu(\theta_a) + \epsilon) - \Lambda_a(\theta) > 0$, and $\eta(\mu(\theta_a) - \epsilon) - \Lambda_a(\eta) > 0$, and so there are constants $\eta = \eta(\boldsymbol{\theta}, \epsilon) \in (0, 1)$ and $c = c(\boldsymbol{\theta}, \epsilon)$, such that for any $a = 1, \dots, K$,

$$\mathbb{P}_{\theta_a}\left(\max_{n=\lceil \delta t\rceil,\dots,t} |\bar{Y}_n^a - \mu(\theta_a)| \geq \epsilon\right) \leq c \sum_{n=\lceil \delta t\rceil}^{\infty} \eta^n \leq \frac{c\eta^{\delta t}}{1-\eta}.$$

□

5.C Analysis of Algorithm 3

As a proxy for the regret we will use the following quantity which involves directly the number of times each arm $a \in \{1, \dots, N\}$ hasn't been played, and the number of times each arm $b \in \{L+1, \dots, K\}$ has been played,

$$\tilde{R}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T) = \sum_{a=1}^{N} (\mu(\theta_a) - \mu(\theta_M))\, \mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}[T - N_a(T)] + \sum_{b=L+1}^{K} (\mu(\theta_M) - \mu(\theta_b))\, \mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}[N_b(T)]. \tag{5.16}$$

For the IID case $\tilde{R}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T) = R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T)$, and in the more general Markovian case $\tilde{R}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T)$ is just a constant term apart from the expected regret $R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T)$. Note that a feature that makes the case of multiple plays more delicate than the case of single plays, even for IID rewards, is the presence of the first summand in Equation 5.16. For this we also need to analyze the number of times each of the best N arms hasn't been played.

Lemma 23.

$$\left|R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T) - \tilde{R}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T)\right| \leq \sum_{a=1}^{K} R_a \cdot \sum_{x\in S} |f(x)|,$$

where $R_a = \mathbb{E}_{\theta_a}\left[\inf\{n \geq 1 : X_{n+1}^a = X_1^a\}\right] < \infty$.

We start the analysis by establishing the relation between the expected regret, Equation 5.1, and its proxy, Equation 5.16. For this we will need the following lemma.

Lemma 24 (Lemma 2.1 in [4])**.** *Let $\{X_n\}_{n\in\mathbb{Z}_{\geq 0}}$ be a Markov chain on a finite state space S, with irreducible transition probability matrix P, stationary distribution π, and initial distribution q. Let $\mathcal{F}_n$ be the σ-field generated by $X_0,\ldots,X_n$. Let τ be a stopping time with respect to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{Z}_{\geq 0}}$ such that $\mathbb{E}[\tau]<\infty$. Define $N(x,n)$ to be the number of visits to state x from time 1 to time n, i.e. $N(x,n)=\sum_{k=1}^{n} I\{X_k=x\}$. Then*

$$|\mathbb{E}[N(x,\tau)]-\pi(x)\,\mathbb{E}[\tau]|\leq R,\ \textit{for } x\in S,$$

where $R=\mathbb{E}[\inf\{n\geq 1: X_{n+1}=X_1\}]<\infty$.

Proof of Lemma 23.
First note that,

$$S_T=\sum_{a=1}^{K}\sum_{x\in S} f(x)N_a(x,N_a(T)).$$

For each $a\in[K]$, using first the triangle inequality, and then Lemma 24 for the stopping time $N_a(T)$, we obtain,

$$\begin{aligned}
&\left|\sum_{x\in S} f(x)(\mathbb{E}^{\boldsymbol{\phi}}_{\boldsymbol{\theta}}[N_a(x,N_a(T))]-\pi_{\theta_a}(x)\,\mathbb{E}^{\boldsymbol{\phi}}_{\boldsymbol{\theta}}[N_a(T)])\right| \\
&\qquad\leq \sum_{x\in S}|f(x)|\left|\mathbb{E}^{\boldsymbol{\phi}}_{\boldsymbol{\theta}}[N_a(x,N_a(T))]-\pi_{\theta_a}(x)\,\mathbb{E}^{\boldsymbol{\phi}}_{\boldsymbol{\theta}}[N_a(T)]\right| \\
&\qquad\leq R_a\cdot\sum_{x\in S}|f(x)|.
\end{aligned}$$

Hence summing over $a\in[K]$, and using the triangle inequality, we see that,

$$\left|S_T-\sum_{a=1}^{K}\mu(\theta_a)\,\mathbb{E}^{\mathcal{A}_\delta}_{\boldsymbol{\theta}}[N_a(T)]\right|\leq\sum_{a=1}^{K}R_a\cdot\sum_{x\in S}|f(x)|.$$

To conclude the proof note that,

$$\begin{aligned}
&T\sum_{a=1}^{M}\mu(\theta_a) - \sum_{a=1}^{K}\mu(\theta_a)\,\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_a(T)] \\
&\quad= \sum_{a=1}^{N}\mu(\theta_a)\,\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[T - N_a(T)] + \mu(\theta_M)(M-N) - \mu(\theta_M)\sum_{a=N+1}^{K}\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_a(T)] \\
&\qquad+ \sum_{b=L+1}^{K}(\mu(\theta_M) - \mu(\theta_b))\,\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_b(T)] \\
&\quad= \sum_{a=1}^{N}(\mu(\theta_a) - \mu(\theta_M))\,\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[T - N_a(T)] + \sum_{b=L+1}^{K}(\mu(\theta_M) - \mu(\theta_b))\,\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_b(T)],
\end{aligned}$$

where in the last equality we used the fact that $\sum_{a=1}^{N}\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_a(T)] + \sum_{a=N+1}^{K}\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_a(T)] = TM$. □

Next we show that Algorithm 3 is well-defined.

Proof of Proposition 4.
Recall that $\sum_{a\in[K]} N_a(t) = tM$, and so there exists an arm a_1 such that $N_{a_1}(t) \geq tM/K$. Then $\sum_{a\in[K]-\{a_1\}} N_a(t) \geq t(M-1)$, and so there exists an arm $a_2 \neq a_1$ such that $N_{a_2}(t) \geq t(M-1)/(K-1)$. Inductively we can see that there exist M distinct arms $a_1, \ldots, a_M$ such that $N_{a_i}(t) \geq t(M-i+1)/(K-i+1) \geq t/K > \delta t$, for $i = 1, \ldots, M$. □

5.C.1 Sketch for the rest of the analysis

Due to Lemma 23, it suffices to upper bound the proxy for the expected regret given in Equation 5.16. Therefore, we can break the analysis in two parts: upper bounding $\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[T - N_a(T)]$, for $a = 1, \ldots, N$, and upper bounding $\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_b(T)]$, for $b = L+1, \ldots, K$.

For the first part, we show in Section 5.C that the expected number of times that an arm $a \in \{1, \ldots, N\}$ hasn't been played, is of the order of $O(\log\log T)$.

Lemma 25. *For every arm $a = 1, \ldots, N$,*

$$\mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}[T - N_a(T)] \leq \frac{4e\gamma^2 NC\left\lceil \frac{2\log\gamma}{\log\frac{1}{\delta}}\right\rceil}{\log\gamma}\log\log T + \gamma^{r_0} + \frac{c\gamma^2\eta^\delta K}{(1-\eta)(1-\eta^\delta)^3},$$

where γ, r_0, η, c and C are constants with respect to T.

For the second part, if $b \in \{L+1, \ldots, K\}$, and $b \in \phi_{t+1}$, then there are three possibilities:

1. $L_t \subseteq [L]$, and $|\bar{Y}_a(t) - \mu(\theta_a)| \geq \epsilon$ for some $a \in L_t$,

2. $L_t \subseteq [L]$, and $|\bar{Y}_a(t) - \mu(\theta_a)| < \epsilon$ for all $a \in L_t$, and $b \in \phi_{t+1}$,

3. $L_t \cap \{L+1, \dots, K\} \neq \emptyset$.

This means that,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}[N_b(T)] \leq M &+ \sum_{t=K}^{T-1} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}\left(L_t \subseteq [L], \text{ and } |\bar{Y}_a(t) - \mu(\theta_a)| \geq \epsilon \text{ for some } a \in L_t\right) \\ &+ \sum_{t=K}^{T-1} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}\left(L_t \subseteq [L], \text{ and } |\bar{Y}_a(t) - \mu(\theta_a)| < \epsilon \text{ for all } a \in L_t, \text{ and } b \in \phi_{t+1}\right) \\ &+ \sum_{t=K}^{T-1} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(L_t \cap \{L+1, \dots, K\} \neq \emptyset),\end{aligned}$$

and we handle each of those three terms separately.

We show that the first term is upper bounded by $O(1)$.

Lemma 26.

$$\sum_{t=K}^{T-1} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}\left(L_t \subseteq [L], \text{ and } |\bar{Y}_a(t) - \mu(\theta_a)| \geq \epsilon \text{ for some } a \in L_t\right) \leq \frac{cL\eta^{\delta K}}{(1-\eta)(1-\eta^{\delta})},$$

where c and η are constant with respect to T.

The second term is of the order of $O(\log T)$, and it is the term that causes the overall logarithmic regret.

Lemma 27.

$$\begin{aligned}&\sum_{t=K}^{T-1} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}\left(L_t \subseteq [L], \text{ and } |\bar{Y}_a(t) - \mu(\theta_a)| < \epsilon \text{ for all } a \in L_t, \text{ and } b \in \phi_{t+1}\right) \\ &\quad\leq \frac{\log T + 3\log\log T}{D\left(\mu(\theta_b) \parallel \mu(\theta_M) - \epsilon\right)} + 1 + 8\sigma^2_{\mu(\theta_a),\mu(\theta_M)-\epsilon}\left(\frac{\dot{D}\left(\mu(\theta_b) \parallel \mu(\theta_M) - \epsilon\right)}{D\left(\mu(\theta_b) \parallel \mu(\theta_M) - \epsilon\right)}\right)^2 \\ &\qquad+ 2\sqrt{2\pi\sigma^2_{\mu(\theta_a),\mu(\theta_M)-\epsilon}}\sqrt{\frac{\dot{D}\left(\mu(\theta_b) \parallel \mu(\theta_M) - \epsilon\right)^2}{D\left(\mu(\theta_b) \parallel \mu(\theta_M) - \epsilon\right)^3}}\left(\sqrt{\log T} + \sqrt{3\log\log T}\right),\end{aligned}$$

where $\sigma^2_{\mu(\theta_a),\mu(\theta_M)-\epsilon}$, and $\dot{D}\left(\mu(\theta_b) \parallel \mu(\theta_M) - \epsilon\right) = \frac{dD(\mu \parallel \mu(\theta_M)-\epsilon)}{d\mu}\,|_{\mu=\mu(\theta_b)}$, are constants with respect to T.

Finally, we show that the third term is upper bounded by $O(\log\log T)$.

Lemma 28.

$$\sum_{t=K}^{T-1} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(L_t \cap \{L+1,\dots,K\} \neq \emptyset) \leq \frac{4e\gamma^2 LC \left\lceil \frac{2\log\gamma}{\log\frac{1}{\delta}} \right\rceil}{\log\gamma} \log\log T + \gamma^{r_0} + \frac{c\gamma^2\eta^\delta K}{(1-\eta)(1-\eta^\delta)^3},$$

where γ, r_0, η, c and C are constants with respect to T.

This concludes the proof of Theorem 15, modulo the four bounds of this subsection which are established in the next subsection.

5.C.2 Proofs for the four bounds

For the rest of the analysis we define the following events which describe good behavior of the sample means and the upper confidence bounds. For $\gamma, r \in \mathbb{Z}_{>1}$ let,

$$A_r = \bigcap_{a\in[K]} \bigcap_{\gamma^{r-1}\leq t\leq\gamma^{r+1}} \left\{ \max_{n=\lceil\delta t\rceil,\dots,t} |\bar{Y}_n^a - \mu(\theta_a)| < \epsilon \right\},$$

$$B_r = \bigcap_{a\in[N]} \bigcap_{\gamma^{r-1}\leq t\leq\gamma^{r+1}} \left\{ \min_{n=1,\dots,\lceil\delta t\rceil-1} U_n^a(t) > \mu(\theta_N) \right\},$$

$$C_r = \bigcap_{a\in[L]} \bigcap_{\gamma^{r-1}\leq t\leq\gamma^{r+1}} \left\{ \min_{n=1,\dots,\lceil\delta t\rceil-1} U_n^a(t) > \mu(\theta_a) \right\}.$$

Indeed, the following bounds, which rely on the concentration results of Section 5.3, suggest that those events will happen with some good probability.

Lemma 29.

$$\mathbb{P}_{\boldsymbol{\theta}}(A_r^c) \leq \frac{cK\eta^{\delta\gamma^{r-1}}}{(1-\eta)(1-\eta^\delta)}, \quad \mathbb{P}_{\boldsymbol{\theta}}(B_r^c) \leq \frac{4eNC \left\lceil \frac{2\log\gamma}{\log\frac{1}{\delta}} \right\rceil}{(r-1)\gamma^{r-1}\log\gamma}, \quad \mathbb{P}_{\boldsymbol{\theta}}(C_r^c) \leq \frac{4eLC \left\lceil \frac{2\log\gamma}{\log\frac{1}{\delta}} \right\rceil}{(r-1)\gamma^{r-1}\log\gamma},$$

where $\eta \in (0,1)$, c and C are constants with respect to r.

Proof. The first bound follows directly from Equation 5.15 and a union bound.

For the second bound, let $p = \left\lceil \frac{2\log\gamma}{\log\frac{1}{\delta}} \right\rceil$, so that $\left\lfloor \frac{\gamma^{r-1}}{\delta^p} \right\rfloor \geq \gamma^{r+1}$. For $i = 0,\dots,p$ let $t_i = \left\lfloor \frac{\gamma^{r-1}}{\delta^i} \right\rfloor$, and define,

$$D_i = \bigcap_{a\in[N]} \left\{ \min_{n=1,\dots,t_i} U_n^a(t) > \mu(\theta_a) \right\}.$$

From Equation 5.10 we see that,

$$\mathbb{P}_{\boldsymbol{\theta}}(D_i^c) \leq \frac{4eN \max_{a\in[N]} C_-^a}{t_i \log t_i} \leq \frac{4eN \max_{a\in[N]} C_-^a}{(r-1)\gamma^{r-1}\log\gamma},$$

where C_-^a is the constant from Lemma 18.

Fix $a \in [N]$, and $\gamma^{r-1} \leq t \leq \gamma^{r+1}$. There exists $i \in \{0, \ldots, p-1\}$ such that $t_i \leq t \leq t_{i+1}$, and so $t_i > \delta t_i - 1 \geq \delta t - 1$, which gives that $t_i \geq \lceil \delta t \rceil - 1$. On D_i, due to Equation 5.8, we have that,

$$\min_{n=1,\ldots,\lceil \delta t \rceil - 1} U_n^a(t) \geq \min_{n=1,\ldots,\lceil \delta t \rceil - 1} U_n^a(t_i) \geq \min_{n=1,\ldots,t_i} U_n^a(t_i) > \mu(\theta_a) \geq \mu(\theta_N).$$

Therefore,

$$\mathbb{P}_{\boldsymbol{\theta}}(B_r^c) \leq \sum_{i=0}^{p-1} \mathbb{P}_{\boldsymbol{\theta}}(D_i^c) \leq \frac{4eNp \max_{a \in [N]} C_-^a}{(r-1)\gamma^{r-1} \log \gamma}.$$

The third bound is established along the same lines. □

In order to establish Lemma 25 we need the following lemma which states that, on $A_r \cap B_r$, an event of sufficiently large probability according to Lemma 29, all the best N arms are played.

Lemma 30 (Lemma 5.3 in [3])**.** *Fix $\gamma \geq \lceil (1 - K\delta)^{-1} \rceil + 2$, and let $r_0 = \lceil \log_\gamma \frac{2K}{1 - K\delta - \gamma^{-1}} \rceil + 2$. For any $r \geq r_0$, on $A_r \cap B_r$ we have that $[N] \subset \phi_{t+1}$ for all $\gamma^r \leq t \leq \gamma^{r+1}$.*

Proof of Lemma 25.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}[T - N_a(T)] &\leq \gamma^{r_0} + \sum_{r=r_0}^{\lceil \log_\gamma (T-1) \rceil - 1} \sum_{\gamma^r \leq t \leq \gamma^{r+1}} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(a \notin \phi_{t+1}) \\
&\leq \gamma^{r_0} + \sum_{r=r_0}^{\lceil \log_\gamma (T-1) \rceil - 1} \sum_{\gamma^r \leq t \leq \gamma^{r+1}} (\mathbb{P}_{\boldsymbol{\theta}}(A_r^c) + \mathbb{P}_{\boldsymbol{\theta}}(B_r^c)) \\
&\leq \gamma^{r_0} + \sum_{r=r_0}^{\lceil \log_\gamma (T-1) \rceil - 1} \left(\frac{cK\gamma^{r+1}\eta^{\delta\gamma^{r-1}}}{(1-\eta)(1-\eta^\delta)} + \frac{4e\gamma^2 NC \left\lceil \frac{2\log\gamma}{\log \frac{1}{\delta}} \right\rceil}{(r-1)\log\gamma} \right),
\end{aligned}$$

where the second inequality follows from Lemma 30, and the third from Lemma 29. Now we use a simple logarithmic upper bound on the harmonic number to obtain,

$$\sum_{r=r_0}^{\lceil \log_\gamma (T-1) \rceil - 1} \frac{1}{r-1} \leq \sum_{r=3}^{\lceil \log_\gamma (T-1) \rceil - 1} \frac{1}{r-1} \leq \log \log_\gamma T \leq \log \log T.$$

Finally, we can upper bound the other summand by a constant, with respect to T, in the following way,

$$\sum_{r=r_0}^{\lceil \log_\gamma (T-1) \rceil - 1} \gamma^{r-1} \eta^{\delta\gamma^{r-1}} \leq \sum_{k=1}^{\infty} k\eta^{\delta k} = \frac{\eta^\delta}{(1-\eta^\delta)^2}.$$

□

Proof of Lemma 26.
Using Equation 5.15 it is straightforward to see that

$$\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}\left(L_t \subseteq [L], \text{ and } |\bar{Y}_a(t) - \mu(\theta_a)| \geq \epsilon \text{ for some } a \in L_t\right) \leq \frac{cL\eta^{\delta t}}{1-\eta},$$

and the conclusion follows by summing the geometric series. □

Proof of Lemma 27.
Assume that $L_t \subseteq [L]$, and $|\bar{Y}_a(t) - \mu(\theta_a)| < \epsilon$ for all $a \in L_t$, and $b \in \phi_{t+1}$. Then it must be the case that $b \equiv t+1 \pmod K$, $b \notin L_t$, and $U_b(t) > \min_{a\in L_t} \bar{Y}_a(t) > \min_{a\in L_t} \mu(\theta_a) - \epsilon \geq \mu(\theta_M) - \epsilon$. This shows that,

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}\left(L_t \subseteq [L], \text{ and } |\bar{Y}_a(t) - \mu(\theta_a)| < \epsilon \text{ for all } a \in L_t, \text{ and } b \in \phi_{t+1}\right)\\ &\quad\leq \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(b \in \phi_{t+1}, \text{ and } U_b(t) > \mu(\theta_M) - \epsilon).\end{aligned}$$

Furthermore,

$$\begin{aligned}&\sum_{t=K}^{T-1} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(b \in \phi_{t+1}, \text{ and } U_b(t) > \mu(\theta_M) - \epsilon)\\ &\quad= \sum_{t=K}^{T-1}\sum_{n=M+1}^{M+T-K} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(\tau_n^b = t+1, \text{ and } U_n^b(t) > \mu(\theta_M) - \epsilon)\\ &\quad\leq \sum_{t=K}^{T-1}\sum_{n=M+1}^{M+T-K} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(\tau_n^b = t+1, \text{ and } U_n^b(T) > \mu(\theta_M) - \epsilon)\\ &\quad= \sum_{n=M+1}^{M+T-K}\sum_{t=K}^{T-1} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(\tau_n^b = t+1, \text{ and } U_n^b(T) > \mu(\theta_M) - \epsilon)\\ &\quad\leq \sum_{n=M+1}^{M+T-K} \mathbb{P}_{\theta_b}(U_n^b(T) > \mu(\theta_M) - \epsilon),\end{aligned}$$

where in the first inequality we used Equation 5.8. Now the conclusion follows from Equation 5.11. □

In order to establish Lemma 28 we need the following lemma which states that, on $A_r \cap C_r$, an event of sufficiently large probability according to Lemma 29, only arms from $\{1, \ldots, L\}$ have been played at least $\lceil \delta t \rceil$ times and have a large sample mean.

Lemma 31 (Lemma 5.3 B in [3])**.** *Fix $\gamma \geq \lceil (1-K\delta)^{-1} \rceil + 2$, and let $r_0 = \lceil \log_\gamma \frac{2K}{1-K\delta-\gamma^{-1}} \rceil + 2$. For any $r \geq r_0$, on $A_r \cap C_r$ we have that $L_t \subseteq [L]$ for all $\gamma^r \leq t \leq \gamma^{r+1}$.*

Proof of Lemma 28.
From Lemma 31 we see that,

$$\sum_{t=K}^{T-1} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(L_t \cap \{L+1,\ldots,K\} \neq \emptyset) \leq \gamma^{r_0} + \sum_{r=r_0}^{\lceil \log_\gamma(T-1)\rceil - 1} \sum_{\gamma^r \leq t \leq \gamma^{r+1}} (\mathbb{P}_{\boldsymbol{\theta}}(A_r^c) + \mathbb{P}_{\boldsymbol{\theta}}(C_r^c)).$$

The rest of the calculations are similar with the proof of Lemma 25. □

Proof of Corollary 3.
In the finite-time regret bound of Theorem 15 we divide by $\log T$, let T go to ∞, and then let ϵ go to 0 in order to get,

$$\limsup_{T\to\infty} \frac{R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T)}{\log T} \leq \sum_{b=L+1}^{K} \frac{\mu(\theta_M) - \mu(\theta_b)}{D\left(\mu(\theta_b) \parallel \mu(\theta_M)\right)}.$$

The conclusion now follows by using the asymptotic lower bound from Equation 5.3. □

Proof of Theorem 16.
The proof of Theorem 16 follows along the lines the proof of Theorem 15, by replacing instances of entries of the right Perron-Frobenius eigenvector $v_\theta(x)$ with one, and is thus omitted. □

5.D General Asymptotic Lower Bound

Recall from Subsection 5.2.1 the general one-parameter family of Markov chains $\{\mathbb{P}_\theta : \theta \in \Theta\}$, where each Markovian probability law $\mathbb{P}_\theta$ is characterized by an initial distribution q_θ and a transition probability matrix P_θ. For this family we assume that,

$$P_\theta \text{ is irreducible for all } \theta \in \Theta. \tag{5.17}$$

$$P_\theta(x,y) > 0 \;\Rightarrow\; P_\lambda(x,y) > 0, \text{ for all } \theta,\lambda \in \Theta,\; x,y \in S. \tag{5.18}$$

$$q_\theta(x) > 0 \;\Rightarrow\; q_\lambda(x), \text{ for all } \theta,\lambda \in \Theta,\; x \in S. \tag{5.19}$$

In general it is not necessary that the parameter space Θ is the whole real line, but it is assumed to satisfy the following denseness condition. For all $\lambda \in \Theta$ and all $\delta > 0$, there exists $\lambda' \in \Theta$ such that,

$$\mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta. \tag{5.20}$$

Furthermore, the Kullback-Leibler divergence rate is assumed to satisfy the following continuity property. For all $\epsilon > 0$, and for all $\theta,\lambda \in \Theta$ such that $\mu(\lambda) > \mu(\theta)$, there exists $\delta > 0$ such that,

$$\mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta \;\Rightarrow\; |D\left(\theta \parallel \lambda\right) - D\left(\theta \parallel \lambda'\right)| < \epsilon. \tag{5.21}$$

An adaptive allocation rule $\boldsymbol{\phi}$ is said to be *uniformly good* if,

$$R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T) = o(T^{\alpha}), \text{ for all } \boldsymbol{\theta} \in \Theta^K, \text{ and all } \alpha > 0.$$

Under those conditions [4] establish the following asymptotic lower bound.

Theorem 17 (Theorem 3.1 from [4])**.** *Assume that the one-parameter family of Markov chains on the finite state space S, together with the reward function $f : S \to \mathbb{R}$, satisfy conditions* (5.17)*,* (5.18)*,* (5.19)*,* (5.20)*, and* (5.21)*. Let $\boldsymbol{\phi}$ be a uniformly good allocation rule. Fix a parameter configuration $\boldsymbol{\theta} \in \Theta^K$, and without loss of generality assume that,*

$$\mu(\theta_1) \geq \ldots \geq \mu(\theta_N) > \mu(\theta_{N+1}) \ldots = \mu(\theta_M) = \ldots = \mu(\theta_L) > \mu(\theta_{L+1}) \geq \ldots \geq \mu(\theta_K).$$

Then for every $b = L+1, \ldots, K$,

$$\frac{1}{D\left(\theta_b \parallel \theta_M\right)} \leq \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}\left[N_b(T)\right]}{\log T}.$$

Consequently,

$$\sum_{b=L+1}^{K} \frac{\mu(\theta_M) - \mu(\theta_b)}{D\left(\theta_b \parallel \theta_M\right)} \leq \liminf_{T\to\infty} \frac{R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T)}{\log T}.$$

Lower bounds on the expected regret of multi-armed bandit problems are established using a change of measure argument, which relies on the adaptive allocation rule being uniformly good. [56] gave the prototypical change of measure argument, for the case of i.i.d. rewards, and [4] extended this technique for the case of Markovian rewards. Here we give an alternative simplified proof using the data processing inequality, an idea introduced in [49, 20] for the i.i.d. case.

We first set up some notation. Denote by $\mathcal{F}_T$ the σ-field generated by the random variables $\phi_1, \ldots, \phi_T, \{X_n^1\}_{n=0}^{N_1(T)}, \ldots, \{X_n^K\}_{n=0}^{N_K(T)}$, and let $\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}} \mid_{\mathcal{F}_T}$ be the restriction of the probability distribution $\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}$ on $\mathcal{F}_T$. For two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ over the same measurable space we define the *Kullback-Leibler divergence* between $\mathbb{P}$ and $\mathbb{Q}$ as

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

where $\frac{d\mathbb{P}}{d\mathbb{Q}}$ denotes the Radon-Nikodym derivative, when $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$. Note that we have used the same notation as for the Kullback-Leibler divergence rate between two Markov chains, but it should be clear from the arguments whether we refer to the divergence or the divergence rate. For $p, q \in [0,1]$, the *binary Kullback-Leibler divergence* is denoted by

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

The following lemma, from [73], will be crucial in establishing the lower bound.

Lemma 32 (Lemma 1 in [73]). *Let $\boldsymbol{\theta}, \boldsymbol{\lambda} \in \Theta^K$ be two parameter configurations. Let τ be a stopping time with respect to $(\mathcal{F}_t)_{t\in\mathbb{Z}_{>0}}$, with $\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[\tau]$, $\mathbb{E}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}[\tau] < \infty$. Then*

$$D\left(\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}\mid_{\mathcal{F}_\tau} \Big\| \mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}\mid_{\mathcal{F}_\tau}\right) \leq \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_a(\tau)] D\left(\theta_a \parallel \lambda_a\right) + \sum_{a=1}^{K} D\left(q_{\theta_a} \parallel q_{\lambda_a}\right) + \sum_{a=1}^{K} R_{\theta_a} \sum_{x,y} \pi_{\theta_a}(x) P_{\theta_a}(x,y) \left|\log \frac{P_{\theta_a}(x,y)}{P_{\lambda_a}(x,y)}\right|,$$

where $R_{\theta_a} = \mathbb{E}_{\theta_a}\left[\inf\{n \geq 1 : X_{n+1}^a = X_1^a\}\right] < \infty$, the first summand is finite due to (5.19)*, and the second summand is finite due to* (5.18)*.*

Proof of Theorem 17.
Fix $b \in \{L+1, \dots, K\}$, and $\epsilon > 0$. Due to Equation 5.20 and Equation 5.21, there exists $\lambda \in \Theta$ such that

$$\mu(\theta_M) < \mu(\lambda), \text{ and } |D\left(\theta_b \parallel \theta_M\right) - D\left(\theta_b \parallel \lambda\right)| < \epsilon.$$

We consider the parameter configuration $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$ given by,

$$\lambda_a = \begin{cases} \theta_a, & \text{if } a \neq b, \\ \lambda, & \text{if } a = b. \end{cases}$$

Using Lemma 32 we obtain,

$$D\left(\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}\mid_{\mathcal{F}_T} \Big\| \mathbb{P}_{\boldsymbol{\lambda}}^{\boldsymbol{\phi}}\mid_{\mathcal{F}_T}\right) \leq D\left(q_{\theta_b} \parallel q_\lambda\right) + R_{\theta_b} D\left(\theta_b \parallel \lambda\right) + \mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}\left[N_b(T)\right] D\left(\theta_b \parallel \lambda\right).$$

From the data processing inequality, see the book of [21], we have that for any event $\mathcal{E} \in \mathcal{F}_T$,

$$D_2\left(\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(\mathcal{E}) \Big\| \mathbb{P}_{\boldsymbol{\lambda}}^{\boldsymbol{\phi}}(\mathcal{E})\right) \leq D\left(\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}\mid_{\mathcal{F}_T} \Big\| \mathbb{P}_{\boldsymbol{\lambda}}^{\boldsymbol{\phi}}\mid_{\mathcal{F}_T}\right).$$

We select $\mathcal{E} = \{N_b(T) \geq \sqrt{T}\}$. Then using Markov's inequality, and the fact that $\boldsymbol{\phi}$ is uniformly good we obtain for any $\alpha > 0$,

$$\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(\mathcal{E}) \leq \frac{\mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}[N_b(T)]}{\sqrt{T}} = \frac{o(T^\alpha)}{\sqrt{T}}, \quad \mathbb{P}_{\boldsymbol{\lambda}}^{\boldsymbol{\phi}}(\mathcal{E}^c) \leq \frac{\mathbb{E}_{\boldsymbol{\lambda}}^{\boldsymbol{\phi}}[T - N_b(T)]}{T - \sqrt{T}} = \frac{o(T^\alpha)}{T - \sqrt{T}}.$$

Using those two inequalities we see that,

$$\liminf_{T\to\infty} \frac{D_2\left(\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(\mathcal{E}) \Big\| \mathbb{P}_{\boldsymbol{\lambda}}^{\boldsymbol{\phi}}(\mathcal{E})\right)}{\log T} = \liminf_{T\to\infty} \frac{\log \frac{1}{\mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}(\mathcal{E}^c)}}{\log T} \geq \lim_{T\to\infty} \frac{\log \frac{T-\sqrt{T}}{o(T^\alpha)}}{\log T} = 1.$$

Therefore,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}[N^b(T)]}{\log T} \geq \frac{1}{D\left(\theta_b \parallel \lambda\right)} \geq \frac{1}{D\left(\theta_b \parallel \theta_M\right) + \epsilon},$$

and the first part of Theorem 17 follows by letting ϵ go to 0. The second part follows from Lemma 23, and Equation 5.16. □

Bibliography

[1] R. Agrawal. "Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem". In: *Adv. in Appl. Probab.* 27.4 (1995), pp. 1054–1078. ISSN: 0001-8678.

[2] D. Aldous and J. Fill. *Reversible Markov Chains and Random Walks on Graphs*. 2002.

[3] V. Anantharam, P. Varaiya, and J. Walrand. "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. I. I.I.D. rewards". In: *IEEE Trans. Automat. Control* 32.11 (1987), pp. 968–976. ISSN: 0018-9286.

[4] V. Anantharam, P. Varaiya, and J. Walrand. "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. II. Markovian rewards". In: *IEEE Trans. Automat. Control* 32.11 (1987), pp. 977–982. ISSN: 0018-9286.

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-time Analysis of the Multiarmed Bandit Problem". In: *Mach. Learn.* 47.2-3 (May 2002), pp. 235–256. ISSN: 0885-6125.

[6] J. Backhoff, D. Bartl, M. Beiglböck, and M. Eder. "Adapted Wasserstein distances and stability in mathematical finance". In: *Finance and Stochastics* 24.3 (2020), pp. 601–632.

[7] J. Backhoff, M. Beiglböck, M. Eder, and A. Pichler. "Fundamental properties of process distances". In: *Stochastic Processes and their Applications* 130.9 (2020), pp. 5575–5591. ISSN: 0304-4149.

[8] J. Backhoff, M. Beiglböck, Y. Lin, and A. Zalashko. "Causal transport in discrete time and applications". In: *SIAM J. Optim.* 27.4 (2017), pp. 2528–2562. ISSN: 1052-6234.

[9] S. Balaji and S. P. Meyn. "Multiplicative ergodicity and large deviations for an irreducible Markov chain". In: *Stochastic Process. Appl.* 90.1 (2000), pp. 123–144. ISSN: 0304-4149.

[10] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Society for Industrial and Applied Mathematics, 1994.

[11] D. P. Bertsekas. "Monotone mappings with application in dynamic programming". In: *SIAM J. Control Optim.* 15.3 (1977), pp. 438–464. ISSN: 0363-0129.

[12] D. Blackwell. "Discounted dynamic programming". In: *Ann. Math. Statist.* 36 (1965), pp. 226–235. ISSN: 0003-4851.

[13] E. Bolthausen and U. Schmock. "On the maximum entropy principle for uniformly ergodic Markov chains". In: *Stochastic Process. Appl.* 33.1 (1989), pp. 1–27. ISSN: 0304-4149.

[14] L. D. Brown. *Fundamentals of statistical exponential families with applications in statistical decision theory*. Vol. 9. Institute of Mathematical Statistics Lecture Notes—Monograph Series. Institute of Mathematical Statistics, Hayward, CA, 1986, pp. x+283. ISBN: 0-940600-10-2.

[15] S. Bubeck and N. Cesa-Bianchi. "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems". In: *Foundations and Trends in Machine Learning* 5.1 (2012), pp. 1–122. ISSN: 1935-8237.

[16] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. "Kullback-Leibler upper confidence bounds for optimal sequential allocation". In: *Ann. Statist.* 41.3 (2013), pp. 1516–1541. ISSN: 0090-5364.

[17] J.-R. Chazottes, P. Collet, C. Külske, and F. Redig. "Concentration inequalities for random fields via coupling". In: *Probab. Theory Related Fields* 137.1-2 (2007), pp. 201–225. ISSN: 0178-8051.

[18] H. Chernoff. "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations". In: *Ann. Math. Statistics* 23 (1952), pp. 493–507. ISSN: 0003-4851.

[19] K. L. Chung. *Markov chains with stationary transition probabilities*. Die Grundlehren der mathematischen Wissenschaften, Bd. 104. Springer-Verlag, Berlin-Göttingen-Heidelberg, 1960, pp. x+278.

[20] R. Combes and A. Proutiere. *Unimodal Bandits without Smoothness*. 2014. arXiv: `1406.7447 [cs.LG]`.

[21] T. M. Cover and J. A. Thomas. *Elements of information theory*. Second. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006, pp. xxiv+748.

[22] I. Csiszár, T. M. Cover, and B. S. Choi. "Conditional limit theorems under Markov conditioning". In: *IEEE Trans. Inform. Theory* 33.6 (1987), pp. 788–801. ISSN: 0018-9448.

[23] L. D. Davisson, G. Longo, and A. Sgarro. "The error exponent for the noiseless encoding of finite ergodic Markov sources". In: *IEEE Trans. Inform. Theory* 27.4 (1981), pp. 431–438. ISSN: 0018-9448.

[24] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Second. Vol. 38. Applications of Mathematics (New York). Springer-Verlag, New York, 1998, pp. xvi+396. ISBN: 0-387-98406-2.

[25] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001, pp. xii+208. ISBN: 0-387-95117-2.

[26] I. H. Dinwoodie. "A probability inequality for the occupation measure of a reversible Markov chain". In: *Ann. Appl. Probab.* 5.1 (1995), pp. 37–43. ISSN: 1050-5164.

[27] M. D. Donsker and S. R. S. Varadhan. "Asymptotic evaluation of certain Markov process expectations for large time. I. II". In: *Comm. Pure Appl. Math.* 28 (1975), 1–47, ibid. 28 (1975), 279–301. ISSN: 0010-3640.

[28] R. Douc, E. Moulines, J. Olsson, and R. van Handel. "Consistency of the maximum likelihood estimator for general hidden Markov models". In: *Ann. Statist.* 39.1 (2011), pp. 474–513. ISSN: 0090-5364.

[29] R. Durrett. *Probability: Theory and Examples*. Fifth. Cambridge University Press, Cambridge, 2019.

[30] R. Durrett. *Probability: theory and examples*. Fourth. Vol. 31. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010, pp. x+428. ISBN: 978-0-521-76539-8.

[31] R. S. Ellis. "Large deviations for a general class of random vectors". In: *Ann. Probab.* 12.1 (1984), pp. 1–12. ISSN: 0091-1798.

[32] F. Esscher. "On the probability function in the collective theory of risk". In: *Scandinavian Actuarial Journal* 1932.3 (1932), pp. 175–195.

[33] E. Even-Dar, S. Mannor, and Y. Mansour. "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems". In: *J. Mach. Learn. Res.* 7 (2006), pp. 1079–1105. ISSN: 1532-4435.

[34] J. Fan, B. Jiang, and Q. Sun. *Hoeffding's lemma for Markov Chains and its applications to statistical learning*. 2018. eprint: `arXiv:1802.00211`.

[35] A. Garivier and F. Leonardi. "Context tree selection: A unifying view". In: *Stochastic Processes and their Applications* 121.11 (2011), pp. 2488–2506. ISSN: 0304-4149.

[36] A. Garivier and O. Cappé. "The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond". In: *Proceedings of the 24th Annual Conference on Learning Theory*. Ed. by S. M. Kakade and U. von Luxburg. Vol. 19. Proceedings of Machine Learning Research. Budapest, Hungary: PMLR, June 2011, pp. 359–376.

[37] A. Garivier and E. Kaufmann. "Optimal best arm identification with fixed confidence". In: *Proceedings of the 29th Conference On Learning Theory* 49 (Jan. 2016), pp. 1–30.

[38] J. Gertner. "On large deviations from an invariant measure". In: *Teor. Verojatnost. i Primenen.* 22.1 (1977), pp. 27–42. ISSN: 0040-361x.

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

[40] P. W. Glynn and D. Ormoneit. "Hoeffding's inequality for uniformly ergodic Markov chains". In: *Statist. Probab. Lett.* 56.2 (2002), pp. 143–146. ISSN: 0167-7152.

[41] S. Goldstein. "Maximal coupling". In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 46.2 (1979), pp. 193–204. ISSN: 0044-3719.

[42] D. Griffeath. "A maximal coupling for Markov chains". In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 31 (1975), pp. 95–106.

[43] M. Hayashi and S. Watanabe. "Information geometry approach to parameter estimation in Markov chains". In: *Ann. Statist.* 44.4 (2016), pp. 1495–1535. ISSN: 0090-5364.

[44] W. Hoeffding. "Probability inequalities for sums of bounded random variables". In: *J. Amer. Statist. Assoc.* 58 (1963), pp. 13–30. ISSN: 0162-1459.

[45] R. A. Horn and C. R. Johnson. *Matrix analysis*. Second. Cambridge University Press, Cambridge, 2013, pp. xviii+643. ISBN: 978-0-521-54823-6.

[46] K. G. Jamieson, M. Malloy, R. D. Nowak, and S. Bubeck. "lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits." In: *COLT*. Vol. 35. JMLR Workshop and Conference Proceedings. 2014, pp. 423–439.

[47] M. Jerrum, A. Sinclair, and E. Vigoda. "A polynomial-time approximation algorithm for the permanent of a matrix with non-negative entries". In: *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*. ACM, New York, 2001, pp. 712–721.

[48] N. Kahale. "Large deviation bounds for Markov chains". In: *Combin. Probab. Comput.* 6.4 (1997), pp. 465–474. ISSN: 0963-5483.

[49] E. Kaufmann, O. Cappé, and A. Garivier. "On the Complexity of Best-arm Identification in Multi-armed Bandit Models". In: *J. Mach. Learn. Res.* 17.1 (Jan. 2016), pp. 1–42. ISSN: 1532-4435.

[50] E. Kaufmann and W. Koolen. *Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals*. 2018. arXiv: `1811.11419 [stat.ML]`.

[51] J. Komiyama, J. Honda, and H. Nakagawa. "Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-Armed Bandit Problem with Multiple Plays". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, 2015, pp. 1152–1161.

[52] A. Kontorovich and M. Raginsky. "Concentration of Measure Without Independence: A Unified Approach Via the Martingale Method". In: *Convexity and Concentration*. Ed. by E. Carlen, M. Madiman, and E. M. Werner. New York, NY: Springer New York, 2017, pp. 183–210.

[53] A. Kontorovich and K. Ramanan. "Concentration inequalities for dependent random variables via the martingale method". In: *Ann. Probab.* 36.6 (2008), pp. 2126–2158. ISSN: 0091-1798.

[54] I. Kontoyiannis, L. A. Lastras-Montaño, and S. P. Meyn. "Exponential Bounds and Stopping Rules for MCMC and General Markov Chains". In: *Proceedings of the 1st International Conference on Performance Evaluation Methodolgies and Tools*. valuetools '06. Pisa, Italy: ACM, 2006. ISBN: 1-59593-504-5.

[55] I. Kontoyiannis and S. P. Meyn. "Spectral theory and limit theorems for geometrically ergodic Markov processes". In: *Ann. Appl. Probab.* 13.1 (2003), pp. 304–362. ISSN: 1050-5164.

[56] T. L. Lai and H. Robbins. "Asymptotically efficient adaptive allocation rules". In: *Adv. in Appl. Math.* 6.1 (1985), pp. 4–22. ISSN: 0196-8858.

[57] R. Lassalle. *Causal transference plans and their Monge-Kantorovich problems*. 2015. arXiv: `1303.6925 [math.PR]`.

[58] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. 2019.

[59] C. A. León and F. Perron. "Optimal Hoeffding bounds for discrete reversible Markov chains". In: *Ann. Appl. Probab.* 14.2 (2004), pp. 958–970. ISSN: 1050-5164.

[60] P. Lezaud. "Chernoff-type bound for finite Markov chains". In: *Ann. Appl. Probab.* 8.3 (1998), pp. 849–867. ISSN: 1050-5164.

[61] O.-A. Maillard, R. Munos, and G. Stoltz. "A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences". In: *Proceedings of the 24th Annual Conference on Learning Theory*. Ed. by S. M. Kakade and U. von Luxburg. Vol. 19. Proceedings of Machine Learning Research. Budapest, Hungary: PMLR, June 2011, pp. 497–514.

[62] S. Mannor and J. N. Tsitsiklis. "The sample complexity of exploration in the multi-armed bandit problem". In: *J. Mach. Learn. Res.* 5 (2004), pp. 623–648. ISSN: 1532-4435.

[63] K. Marton. "Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration". In: *Ann. Probab.* 24.2 (1996), pp. 857–866. ISSN: 0091-1798.

[64] K. Marton. "A measure concentration inequality for contracting Markov chains". In: *Geom. Funct. Anal.* 6.3 (1996), pp. 556–571. ISSN: 1016-443X.

[65] K. Marton. "Measure concentration and strong mixing". In: *Studia Sci. Math. Hungar.* 40.1-2 (2003), pp. 95–113. ISSN: 0081-6906.

[66] K. Marton. "Measure concentration for a class of random processes". In: *Probab. Theory Related Fields* 110.3 (1998), pp. 427–439. ISSN: 0178-8051.

[67] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.

[68] B. Miasojedow. "Hoeffding's inequalities for geometrically ergodic Markov chains on general state space". In: *Statist. Probab. Lett.* 87 (2014), pp. 115–120. ISSN: 0167-7152.

[69] H. D. Miller. "A convexity property in the theory of random variables defined on a finite Markov chain". In: *Ann. Math. Statist.* 32 (1961), pp. 1260–1270. ISSN: 0003-4851.

[70] V. Moulos. "A Hoeffding Inequality for Finite State Markov Chains and its Applications to Markovian Bandits". In: *2020 IEEE International Symposium on Information Theory (ISIT)*. 2020, pp. 2777–2782.

[71] V. Moulos. *Bicausal Optimal Transport for Markov Chains via Dynamic Programming*. 2020. arXiv: 2010.06831.

[72] V. Moulos. "Finite-Time Analysis of Round-Robin Kullback-Leibler Upper Confidence Bounds for Optimal Adaptive Allocation with Multiple Plays and Markovian Rewards". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[73] V. Moulos. "Optimal Best Markovian Arm Identification with Fixed Confidence". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019, pp. 5605–5614.

[74] V. Moulos and V. Anantharam. *Optimal Chernoff and Hoeffding Bounds for Finite State Markov Chains*. 2019. eprint: 1907.04467.

[75] 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)*. Okinawa, Japan, Nov. 2005, pp. 1091–1095.

[76] R. Ortner, D. Ryabko, P. Auer, and M. Rémi. "Regret Bounds for Restless Markov Bandits." In: *Algorithmic Learning Theory (ALT)*. 2012.

[77] D. Paulin. "Concentration inequalities for Markov chains by Marton couplings and spectral methods". In: *Electron. J. Probab.* 20 (2015), no. 79, 32. ISSN: 1083-6489.

[78] G. C. Pflug. "Version-Independence and Nested Distributions in Multistage Stochastic Optimization". In: *SIAM Journal on Optimization* 20.3 (2010), pp. 1406–1420.

[79] G. C. Pflug and A. Pichler. "A Distance For Multistage Stochastic Optimization Models". In: *SIAM Journal on Optimization* 22.1 (2012), pp. 1–23.

[80] J. W. Pitman. "On coupling of Markov chains". In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 35.4 (1976), pp. 315–322.

[81] S. Rao. "A Hoeffding inequality for Markov chains". In: *Electron. Commun. Probab.* 24 (2019), Paper No. 14, 11. ISSN: 1083-589X.

[82] J. S. Rosenthal. "Faithful couplings of Markov chains: now equals forever". In: *Adv. in Appl. Math.* 18.3 (1997), pp. 372–381. ISSN: 0196-8858.

[83] P.-M. Samson. "Concentration of measure inequalities for Markov chains and Φ-mixing processes". In: *Ann. Probab.* 28.1 (2000), pp. 416–461. ISSN: 0091-1798.

[84] A. Slivkins. "Introduction to Multi-Armed Bandits". In: *Foundations and Trends® in Machine Learning* 12.1-2 (2019), pp. 1–286. ISSN: 1935-8237.

[85] R. E. Strauch. "Negative dynamic programming". In: *Ann. Math. Statist.* 37 (1966), pp. 871–890. ISSN: 0003-4851.

[86] D. W. Stroock. *An introduction to Markov processes*. Second. Vol. 230. Graduate Texts in Mathematics. Springer, Heidelberg, 2014, pp. xviii+203.

[87] C. Tekin and M. Liu. "Online algorithms for the multi-armed bandit problem with Markovian rewards". In: *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Sept. 2010, pp. 1675–1682.

[88] C. Tekin and M. Liu. "Online Learning of Rested and Restless Bandits". In: *IEEE Trans. Inf. Theor.* 58.8 (Aug. 2012), pp. 5588–5611. ISSN: 0018-9448.

[89] H. Thorisson. *Coupling, stationarity, and regeneration*. Probability and its Applications (New York). Springer-Verlag, New York, 2000, pp. xiv+517. ISBN: 0-387-98779-7.

[90] C. Villani. *Optimal Transport: Old and new*. Vol. 338. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 2009, pp. xxii+973. ISBN: 978-3-540-71049-3.

[91] J. Ville. *Étude critique de la notion de collectif*. NUMDAM, 1939, p. 116.

[92] S. Watanabe and M. Hayashi. "Finite-length analysis on tail probability for Markov chain and application to simple hypothesis testing". In: *Ann. Appl. Probab.* 27.2 (2017), pp. 811–845. ISSN: 1050-5164.