https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 73, No. 4, July–August 2025, pp. 2281–2295
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Off-line Estimation of Controlled Markov Chains: Minimaxity and Sample Complexity

Imon Banerjee,[a,*] Harsha Honnappa,[b] Vinayak Rao[c]

[a] Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208; [b] Edwardson School of Industrial Engineering, Purdue University, West Lafayette, Indiana 47907; [c] Department of Statistics, Purdue University, West Lafayette, Indiana 47907

*Corresponding author

Contact: imon.banerjee@northwestern.edu, https://orcid.org/0000-0003-2572-3048 (IB); honnappa@purdue.edu, https://orcid.org/0000-0002-0834-054X (HH); varao@purdue.edu, https://orcid.org/0000-0002-6249-2923 (VR)

Received: January 29, 2023
Revised: November 25, 2023; July 16, 2024; October 24, 2024
Accepted: November 26, 2024
Published Online in Articles in Advance: February 21, 2025

Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2023.0046

Abstract. In this work, we study a natural nonparametric estimator of the transition probability matrices of a finite controlled Markov chain. We consider an off-line setting with a fixed data set of size m, collected using a so-called logging policy. We develop sample complexity bounds for the estimator and establish conditions for minimaxity. Our statistical bounds depend on the logging policy through its mixing properties. We show that achieving a particular statistical risk bound involves a subtle and interesting trade-off between the strength of the mixing properties and the number of samples. We demonstrate the validity of our results under various examples, such as ergodic Markov chains; weakly ergodic inhomogeneous Markov chains; and controlled Markov chains with nonstationary Markov, episodic, and greedy controls. Lastly, we use these sample complexity bounds to establish concomitant ones for off-line evaluation of stationary Markov control policies.

Funding: I. Banerjee was supported in part by the Ross-Lynn fellowship and McLean scholarship at Purdue University. H. Honnappa was partly supported by the National Science Foundation [Grants CAREER/2143752, DMS/1812197 and DMS/2153915]. V. Rao was supported by the National Science Foundation [Grants RI/1816499 and DMS/1812197].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2023.0046.

Keywords: reinforcement learning • controlled Markov chains • stochastic processes • policy evaluation • nonparametric statistics

1. Introduction

This paper presents probably approximately correct (PAC)–style minimax sample complexity results for statistical estimation of transition matrices of discrete-time, finite-state controlled Markov chains (CMCs) (Borkar 1991). We model a controlled Markov chain as a discrete-time pair process $\{X_i, a_i\}$, where $\{a_i\}$ is a sequence of controls and $\{X_i\}$ is the state sequence that, conditioned on a_i, follows a Markov transition kernel. In this paper, we answer the following question: what is the minimum number of samples required to estimate the transition matrices of a discrete-time, finite-state controlled Markov chain to any given degree of precision?

We answer this question by showing that a particular nonparametric estimator (see below) of the transition matrix is minimax optimal. The control sequence can be viewed as generated by a logging or behavior policy. Assuming the policy is stationary Markovian, $\{X_i, a_i\}$ is jointly Markovian (Hernández-Lerma et al. 1991, chapter 2.3). This simplifies the problem to one of estimating the transition kernel of a Markov chain and opens the door to a number of results under suitable ergodicity and mixing assumptions. Some very recent ones include frequentist (Wolfer and Kontorovich 2021) and Bayesian (Banerjee et al. 2021) PAC bounds. On the other hand, logged data in operations research and machine learning tasks such as hospital emergency scheduling (Lee and Lee 2018), minimum system entropy explorations (Mutti et al. 2022), reward machines (Icarte et al. 2018), adversarial Markov games (Wang et al. 2023), and others are often non-Markovian. The following passage from Mutti et al. (2022, p. 2) helps to elucidate the importance of non-Markovian policies: "In this finite-sample formulation non-Markovian strategies are crucial, and we believe they can benefit a significant range of relevant applications. For example, collecting task-specific samples might be costly in some real-world domains, and a pre-trained non-Markovian strategy is essential to guarantee quality exploration even in a single-trial setting … A non-Markovian strategy could exploit the history of interactions to swiftly identify the structure of the environment, then employing the environment-specific optimal strategy thereafter."

Whereas Mutti et al. (2022) focuses on non-Markovian policies in the online setting, Laroche et al. (2022) and Laroche and Tachet Des Combes (2023) observe that non-Markovianity of policies is particularly an issue in the off-line setting in which logging policies can have an arbitrary structure. However, if the logging policy is non-Markovian, finite sample statistical inference results are quite sparse. Consequently, an understanding of the sample complexity of estimating the transition kernel for non-Markovian controls is an important open problem. In this work, all we assume is that the a_i's are adapted and mixing in a sense defined in Assumptions 3 and 4.

1.1. Contributions

We focus on the following nonparametric estimator of the transition matrices. For any state–control–state tuple (s, l, t), define $N_s^{(l)}$ as the number of visits to the (state, control) pair (s, l) and $N_{s,t}^{(l)}$ as the number of transitions from state s to state t under l. Thus, with $\mathbb{1}[\cdot]$ as the indicator function,

$$N_s^{(l)} := \sum_{i=1}^{m} \mathbb{1}[X_i = s, a_i = l]$$
$$N_{s,t}^{(l)} := \sum_{i=1}^{m} \mathbb{1}[X_i = s, X_{i+1} = t, a_i = l]. \tag{1.1}$$

Then, our estimator of the transition probability from state s to t conditioned on control l, $M_{s,t}^{(l)}$ is

$$\hat{M}_{s,t}^{(l)} := \frac{N_{s,t}^{(l)}}{N_s^{(l)}}. \tag{1.2}$$

We prove that this estimator is minimax optimal under suitable conditions.

Although this particular nonparametric estimator of transition matrices has been previously used in the context of model-based reinforcement learning (RL) studies (Mannor and Tsitsiklis 2005, Li et al. 2022a), its statistical guarantees under non-Markovian controls are not known in the literature. This paper fills this obvious gap. The main contributions of this paper can be summarized as follows:

1. Our main result (Theorem 2) shows that the nonparametric estimator is minimax optimal if the number of samples is large enough and identifies an explicit lower bound on the required sample size. Informally, we prove that the sample complexity of estimating the transition matrices in a CMC with d states and k controls is $\Theta(d^2 k)$ if the CMC is geometrically fast mixing. As we argue in Section 4, a geometrically ergodic Markov chain with dk states can be thought of as a special case of a d-state, k-control CMC with stationary controls. Thus, our result (Theorem 2) recovers the optimal sample complexity of estimating Markov chains from Wolfer and Kontorovich (2021) as a special case.

2. We prove in Theorem 1 that the transition probabilities can be estimated even under a weaker mixing assumption than is required for minimaxity (Theorem 2). However, this involves a trade-off, requiring more samples (roughly $\Omega(d^2k^2)$ in place of $\Omega(d^2k)$) to achieve the same level of estimation error.

3. A useful implication of our sample complexity results is that they yield error bounds for off-line policy evaluation (OPE). Theorem 3 evaluates stationary Markov policies from data logged using non-Markovian controls. The resulting sample complexity recovers minimax optimal rates in the literature, which typically assume Markovian logging policies. Furthermore, Theorem 4 demonstrates a sample complexity of estimating the optimal policy under sufficient regularity conditions on the model class. We also demonstrate how to use our theory to derive the sample complexity bounds for learning the demand distribution of an inventory control problem in Proposition 7 (Goldberg et al. 2021).

From a methodological perspective, our analysis reveals two principles that are broadly useful in establishing sample efficiency results for learning CMCs and other controlled stochastic models. First, is a "Goldilocks" principle that no state–control pair must be visited too many or too few times in a single observed sample path. This can be achieved by ensuring that the control sequence is such that the time to return to a particular pair is uniformly bounded over the state–control space (see Assumptions 1 and 2).

Second, the effect of history on the probability of under- or over-visiting any part of the state–control space is controlled by the mixing properties of the control sequence as defined in Assumptions 5 and 7. Roughly speaking, weaker mixing properties imply looser bounds on these probabilities, in turn implying that estimators are possibly sample-inefficient. The bulk of the existing literature on off-line estimation of CMCs focuses on the setting in which the control sequence forms a stationary ergodic Markov sequence and, under this condition, the nonparametric estimator is minimax optimal as implied by Theorem 2 and Proposition 6. However, if the control sequence mixes comparatively slowly (say, polynomially), then Theorem 1 yields a loose sample complexity bound.

As we prove in Sections 4.2 and 4.3, it is relatively straightforward to verify the geometric mixing properties of the control sequences when the controls are Markovian. However, when the controls are non-Markovian, there is no general result to demonstrate geometric mixing. Thus, a practitioner must be cautious of erroneously assuming the logging policy to be Markovian when it is not. If the controls are not Markovian, then one needs $\Omega(d^2k^2)$ samples instead of $\Omega(d^2k)$ to control the probability of large estimation errors on the transition probabilities.

As a final note on the methodological implications, whereas we focus on finite state–control spaces, we believe that these principles and our analysis yield a broad framework for proving sample efficiency results for off-line estimation of CMCs and potentially other controlled stochastic models under more general model assumptions. For instance, if one uses a histogram or a density estimator of a transition kernel on continuous state spaces and compact control spaces, then our results are directly applicable although the optimal sample complexity would depend upon the smoothness properties of the transition function.

1.2. Related Literature

We divide the review of the literature into three parts. In the first part, we place our results in the context of the existing literature on nonparametric estimation for stochastic processes. In the second and third parts, we relate our sample complexity results to existing relevant ones in the literature on off-line RL and system identification, respectively.

1.2.1. Nonparametric Estimation. The foundations of nonparametric estimation (Tsybakov 2009) of finite ergodic Markov chains were laid by Billingsley (1961). Subsequently, Yakowitz (1979) presented an important extension to infinite state spaces with follow-up work on applications to regression (Yakowitz 1989). There is also extensive literature establishing laws of large numbers (LLNs) (Geyer 1998) and central limit theorems (CLTs) (Jones 2004) for a range of time-homogeneous Markov chains. However, somewhat surprisingly, minimax sample complexity bounds for finite ergodic Markov chains were only established recently in Wolfer and Kontorovich (2021). However, barring some results on LLNs and CLTs (Dobrushin 1956a, b; Rosenblatt-Roth 1963, 1964), results on statistical inference for time-inhomogenous Markov chains remain sparse. Furthermore, such properties when the controls are stochastic in nature are even less understood. A crucial complication in our setting is that the state–control pair process need not be Markovian, complicating the application of existing results. Nonetheless, we recover rates similar to those of Wolfer and Kontorovich (2021) as a special case in Section 4.2, demonstrating the generality of our results.

1.2.2. System Identification. The problem in our paper is analogous to system identification in optimal control, (Vidyasagar and Karandikar 2006, Ljung 2010, Tangirala 2018) in which the parameters to be estimated are the transition matrices. There is a growing body of work that revolves around so-called active learning for system identification (RayChaudhuri and Hamey 1996, Chin et al. 2020, Mania et al. 2020). However, in our work, the logging policy does not necessarily aid active learning. Furthermore, the former settings are online in nature, and system identification in the off-line setting traditionally does not involve a controlling policy (Ljung 1987). Our work recognizes the obvious utility of being able to use off-line data.

1.2.3. Model-Based Off-line Reinforcement Learning. Our results are also of importance to model-based off-line RL (Kidambi et al. 2020, Levine et al. 2020, Yu et al. 2020), which is highly relevant to operations and managerial decision-making problems. For instance, data sets on prognosis, diagnosis, and treatment decisions made by physicians have been proposed to be used to train RL agents to potentially identify new (superior) paths to achieving the same (better) outcomes for patients (Shortreed et al. 2011, Liu et al. 2020, Chen et al. 2021, Yu et al. 2021). Analogously, in manufacturing and service operations management settings, as implied by Armony et al. (2015) in the hospital flow setting, data collected using preexisting flow control and routing policies can be mined to discover new/better protocols and policies. Off-line RL is a natural learning framework to achieve this.

The model-based setting involves constructing a model for the transition probability matrix and then using it to solve the expected Bellman equation. Notable works in this regard are the trio of papers Li et al. (2022a, b) and Yan et al. (2022), which prove, in the limited setting of discounted or finite-horizon problems under Markovian policies, that the model-based off-line RL is minimax optimal. Our results show that it continues to be an optimal estimator in the non-Markovian regime under suitable mixing conditions.

1.3. Outline

The rough outline of the paper is as follows. In Section 2, we introduce some notation and the concepts from uniform mixing and weak mixing. In Section 3, we construct the empirical estimator $\hat{M}^{(l)}$ for the transition matrix $M^{(l)}$ for any control l and formally introduce our assumptions. We then illustrate the trade-off discussed previously by producing weaker PAC bounds for the estimation error $\sup_l \|\hat{M}^{(l)} - M^{(l)}\|_\infty$ under weaker mixing assumptions and a stronger minimax PAC bound under stronger mixing assumptions. In Section 4, we apply our main result to derive statistical guarantees for various reward-free off-line RL tasks under a range of settings, such as stationary controls, Markov controls, and episodic controls. Finally, in Section 5, we use our estimator to obtain estimation guarantees for learning the value function. We end with a summary and discussion of the open questions in Section 6.

2. Preliminaries

2.1. Notations

Let $\mathbb{N}$ and $\mathbb{R}$ denote the natural and real numbers and the symbol $\lfloor \cdot \rfloor$ the floor function. All random variables

in this paper are defined with respect to a filtered probability sample space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, where $\mathcal{F}$ is a σ-algebra and $\mathbb{F} := \{\mathcal{F}_i\}$, with $\mathcal{F}_i \subset \mathcal{F}$, is a given filtration. Let $\{X_i\}$ represent a discrete-time stochastic process adapted to $\mathbb{F}$ with finite state space χ. Overloading notation, we also denote by $\mathbb{P}(X)$ the law of the random variable X. Let $\mathbb{E}[X]$ be the expectation and $\sigma(X)$ the σ-algebra induced by X. A $d \times d$ matrix Q is a stochastic matrix if the rows of Q add up to one and $Q_{s,t}$ denotes the (s, t)th entry of Q. Let $\mathbb{I}$ be a finite set representing the control space and $\{a_i\}$ represent the (not necessarily Markovian) sequence of controls in which $a_i \in \mathbb{I} \quad \forall i$.

2.2. Definitions

Following Borkar (1991), we define a CMC as an $\mathbb{F}$-adapted pair process $\{(X_i, a_i)\}$, where the process $\{X_i\}$ satisfies

$$\mathbb{P}(X_{i+1} = s_{i+1} \,|\, \mathcal{F}_i) = \mathbb{P}(X_{i+1} = s_{i+1} \,|\, X_i = s_i, a_i = l) =: M^{(l)}_{s_i, s_{i+1}}.$$

Let $\mathbb{M} := \{M^{(1)}, \ldots, M^{(k)}\}$ represent the set of transition probability matrices in which $M^{(l)} = [M^{(l)}_{s,t}]$ for all $s, t \in \chi$ and $l \in \mathbb{I}$. Because $|\mathbb{I}| = k$, the number of possible transition matrices for any given CMC is finite. The control sequence $\{a_i\}$ is assumed to satisfy $a_i \in \mathcal{F}_i$ for each $i \geq 0$ (i.e., $\{a_i\}$ is an adapted sequence of controls). We emphasize that our theory holds even when a_i is non-Markovian and non–time homogenous. Let $\mathcal{M}_{\chi,\mathbb{I}}$ be the class of all probability measures over state–control pairs for a CMC with an initial distribution D_0. This constitutes the class of data-generating measures that we consider. In the case in which $\{a_i\}$ is deterministic, $\{X_i\}$ forms a time inhomogeneous Markov chain, in which the transition matrix changes at time step i according to the control a_i. Observe, in particular, that, if $a_i = f(X_i)$, for some given function $f : \mathcal{S} \to \mathbb{I}$, then $\{a_i\}$ is a Markov control sequence.

Let $\hat{M}^{(l)}_{s,t}$ be the normalized state–control visitation frequencies (defined in Equation (1.2)) for the triplet $(s, l, t) \in \chi \times \mathbb{I} \times \chi$ and $\hat{M}^{(l)}$ be the matrix $[\hat{M}^{(l)}_{s,t}]$. Our objective is to find the sample complexity m_{opt} such that $\mathbb{P}(\sup_l \|\hat{M}^{(l)} - M^{(l)}\| > \varepsilon) < \delta$ whenever $m \geq c_1 m_{opt}$ and there exists no estimator $\tilde{M}$ such that $\mathbb{P}(\sup_l \|\tilde{M}^{(l)} - M^{(l)}\| > \varepsilon) < \delta$ whenever $m \leq c_2 m_{opt}$ (for positive universal constants c_1, c_2). Our findings (Theorem 2) indicate m_{opt} to be roughly of order $\Omega(dk)$. Therefore, the empirical estimator achieves the minimax risk $\mathcal{R}_m$ (as defined below) over the class of data-generating models $\mathcal{M}_{\chi,\mathbb{I}}$ whenever the number of samples is m is of the order $\Omega(dk)$.

Definition 1. Any element $\mathcal{P} \in \mathcal{M}_{\chi,\mathbb{I}}$ has an associated set of transition matrices $\{M^{(1)}, \ldots, M^{(k)}\}$ and conditional distributions over the control $\{a_i\}$ conditional on the history until time i. Then, the minimax risk of an estimator $\hat{\mathbb{M}} := (\hat{M}^{(1)}, \ldots, \hat{M}^{(k)})$ of $\mathbb{M} := (M^{(1)}, \ldots, M^{(k)})$ is given by

$$\mathcal{R}_m = \inf_{\hat{\mathbb{M}}} \sup_{\mathcal{P} \in \mathcal{M}_{\chi \times \mathbb{I}}} \mathbb{P}\left(\sup_{l \in \mathbb{I}} \|\hat{M}^{(l)} - M^{(l)}\|_\infty > \varepsilon \right).$$

Remark 1. Observe that we have defined the conditional probability distributions over a_i implicitly as we never explicitly require it for our analysis.

Intuitively, to enable fast learning, we need bounds on how frequently the CMC visits all the state–control pairs and for how long it retains its past memory. To formalize these notions, we define return times and mixing coefficients of a CMC. For two time points $i < j$, we define the history $\mathcal{H}^j_i$ to be $\mathcal{H}^j_i := \sigma(X_j, a_j, \ldots, X_i, a_i) \subset \mathcal{F}_j$ and sample history $\hbar^j_i \in (\chi \times \mathbb{I})^{(j-i+1)}$ to be a fixed sequence of states and controls $\hbar^j_i := (s_j, l_j, \ldots, s_i, a_i)$. We recursively define the time to return for every pair of states and controls (s, l) as follows.

Definition 2. The first hitting time (s, l) is defined as

$$\tau^{(1)}_{s,l} := \min\{n : (X_n = s, a_n = l), (X_j \neq s, a_j \neq l) \ \forall 0 < j < n\}.$$

When $i \geq 2$, the ith time to return (or return time) of the state–control pair (s, l) is recursively defined as

$$\tau^{(i)}_{s,l} := \min\left\{ n : \left(X_{\sum_{k=1}^{i-1} \tau^{(k)}_{s,l}+n} = s, a_{\sum_{k=1}^{i-1} \tau^{(k)}_{s,l}+n} = l \right), \right.$$
$$\left. (X_j \neq s \cup a_j \neq l) \quad \forall \sum_{k=1}^{i-1} \tau^{(k)}_{s,l} < j < \sum_{k=1}^{i-1} \tau^{(k)}_{s,l} + n \right\}.$$

2.3. Mixing Coefficients

In this section, we define the weak and uniform mixing coefficients and related lemmas. Let $\{(X_i, a_i)\}$ be a CMC. For any $i < j \leq m \in \mathbb{N}$, let $\mathbb{T} \in (\chi \times \mathbb{I})^{m-j+1}$, $s_1, s_2 \in \chi$, and $l_1, l_2 \in \mathbb{I}$. Let $\hbar^{i-1}_0$ be an element of $(\chi \times \mathbb{I})^i$. Define the map $(\mathbb{T}, s_1, s_2, l_1, l_2, \hbar^{i-1}_0) \longmapsto \eta_{i,j}(\mathbb{T}, s_1, s_2, l_1, l_2, \hbar^{i-1}_0)$ as $\eta_{i,j}(\mathbb{T}, s_1, s_2, l_1, l_2, \hbar^{i-1}_0) := |A - B|$, where $A = \mathbb{P}((X_m, a_m, \ldots, X_j, a_j) \in \mathbb{T} \,|\, X_i = s_1, a_i = l_1, \mathcal{H}^{i-1}_0 = \hbar^{i-1}_0)$ and $B = \mathbb{P}((X_m, a_m, \ldots, X_j, a_j) \in \mathbb{T} \,|\, X_i = s_2, a_i = l_2, \mathcal{H}^{i-1}_0 = \hbar^{i-1}_0)$. Then, the weak mixing coefficient $\overline{\eta}_{i,j}$ is

$$\overline{\eta}_{i,j} := \sup_{\substack{\mathbb{T}, s_1, s_2, l_1, l_2, \hbar^{i-1}_0, \\ \mathbb{P}(X_i = s_1, a_i = l_1, \mathcal{H}^{i-1}_0 = \hbar^{i-1}_0) > 0, \\ \mathbb{P}(X_i = s_2, a_i = l_2, \mathcal{H}^{i-1}_0 = \hbar^{i-1}_0) > 0}} \eta_{i,j}(\mathbb{T}, s_1, s_2, l_1, l_2, \hbar^{i-1}_0). \quad (2.1)$$

With $\mathbb{T} \in (\chi \times \mathbb{I})^{m-j+1}$ and $\hbar^i_0$ an element of $(\chi \times \mathbb{I})^{i+1}$ as before, the uniform mixing coefficient is

$$\phi_{i,j} := \sup_{\substack{\mathbb{T}, \hbar^i_0, \\ \mathbb{P}(\mathcal{H}^i_0 = \hbar^i_0) > 0}} |\mathbb{P}((X_m, a_m, \ldots, X_j, a_j) \in \mathbb{T}) - \mathbb{P}((X_m, a_m, \ldots, X_j, a_j) \in \mathbb{T} \,|\, \mathcal{H}^i_0 = \hbar^i_0)|. \quad (2.2)$$

The following lemma relates the two mixing coefficients. Its proof can be found in Online Section I.1.

Lemma 1. *The uniform and weak mixing coefficients in Equations (2.1) and (2.2) satisfy* $\phi_{i,j} \leq \overline{\eta}_{i,j} \leq 2\phi_{i,j}$.

Remark 2. We point out that, as defined, both $\overline{\eta}$ and ϕ are dependent on m. This dependence does not affect the analysis. Therefore, we follow the convention in literature (Kontorovich and Ramanan 2008) and make the dependence of $\overline{\eta}$ on m implicit.

Remark 3. We point the interested reader to the classic text (Bradley 2005, theorem 3.1) for the relationship between the uniform mixing coefficient and the rate of convergence in total variation distance to the stationary distribution.

3. Empirical Estimation of Transition Probability Matrices

As mentioned in the introduction, our objective is to estimate the transition matrices of the CMC from a single, finite sample path of length $m \gg 1$. Recall $\hat{M}^{(l)}_{s,t}$ from Equation (1.2) and define

$$\hat{M}^{(l)}(s,\cdot) := (\hat{M}^{(l)}_{s,1}, \hat{M}^{(l)}_{s,2}, \ldots, \hat{M}^{(l)}_{s,d}), \text{ and}$$
$$M^{(l)}(s,\cdot) := (M^{(l)}_{s,1}, M^{(l)}_{s,2}, \ldots, M^{(l)}_{s,d}). \tag{3.1}$$

$M^{(l)}(s,\cdot)$ is the sth row of the transition matrix $M^{(l)}$, and $\hat{M}^{(l)}(s,\cdot)$ is the corresponding estimate. Proposition 1 shows the need to control the number of visits to a state–control pair $N^{(l)}_s$ to find theoretical guarantees for $\hat{M}^{(l)}_{s,t}$.

Proposition 1. *Consider a sample* $\{(X_0,a_0),\ldots,(X_m,a_m)\}$ *from a controlled Markov chain. Let* $0 < n_{low,s} < n_{high,s} < m$ *be any two integers. Then, we have*

$$\mathbb{P}(\|\hat{M}^{(l)}(s,\cdot) - M^{(l)}(s,\cdot)\|_1 > \varepsilon, n_{low,s} \le N^{(l)}_s \le n_{high,s}) \le m \exp\left(-\frac{n_{low,s}}{2}\max\left\{0, \varepsilon - \sqrt{\frac{d}{n_{high,s}}}\right\}^2\right). \tag{3.2}$$

The count statistics $N^{(l)}_s$ are well-studied (Billingsley 1961) when the process is a stationary ergodic Markov chain. We list three challenges, moving from Markov chains to controlled Markov chains.

1. Question of aperiodicity: Consider the following three transition probability matrices:

$$M^{(1)} = \begin{bmatrix} 0 & 1 & 0 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \quad M^{(2)} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$
$$M^{(3)} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \end{bmatrix}.$$

It can be verified easily that each of the transition probability matrices is aperiodic (and, in fact, ergodic). However, consider a time-inhomogenous Markov chain on state-space $\{1,2,3\}$, where the transition matrices arrive in a sequence $(M^{(1)}, M^{(2)}, M^{(3)}, M^{(1)}, M^{(2)}, M^{(3)}, \ldots)$. Not only is it periodic if the initial state is one, it is guaranteed to eventually become periodic.

2. Question of irreducibility: Meyn and Tweedie (2012, theorem 13.0.1) show that an aperiodic and irreducible Markov chain admits a stationary distribution, a key consequence of which is Kac's theorem (Meyn and Tweedie 2012, theorem 10.2.2) establishing the finiteness of the return times of every state in a Markov chain. However, such notions do not translate to a controlled Markov chain.

3. Question of mixing: An ergodic Markov chain on a finite state space is uniformly mixing. However, no equivalent result exists for controlled Markov chains.

Our first two assumptions address the questions of aperiodicity and irreducibility by ensuring that no part of the chain is deterministic in nature, and every state–control pair (s,l) is visited sufficiently often.

Assumption 1. *For all times* i, *there exist constants* ζ_1 *and* ζ_2 *and a set* $\mathcal{S}_i \subset \{1,\ldots,d\} \times \mathbb{I}$ *such that*

$$0 < \zeta_2 \le \mathbb{P}[(X_i,a_i) \in \mathcal{S}_i] \le \zeta_1 < 1.$$

Remark 4. If the controlled Markov chain satisfies the previous assumptions on all but a finite number of time points, our results continue to hold by discarding data. However, it would lead to more cumbersome (but very similar) calculations.

Assumption 2 (Return Time). *There exists an integer* $T > 0$ *such that the return time* $\tau^{(i)}_{s,l}$ *satisfies*

$$\sup_{s,l,i} \mathrm{E}[\tau^{(i)}_{s,l} | \mathcal{F}_{\sum_{p=0}^{i-1} \tau^{(p)}_{s,l}}] < T \text{ almost everywhere.}$$

The following lemma on the expected count statistics follows as a consequence of the previous assumptions. The main theorem is proved by controlling deviations around this expectation. Its proof is in Online Section I.3.

Lemma 2. *For any controlled Markov chain that satisfies Assumptions 1 and 2,*

$$\frac{m}{T} - 1 < \mathrm{E}[N^{(l)}_s] \le m \max\{\zeta_1, 1-\zeta_2\}, \tag{3.3}$$

where $\zeta_1, \zeta_2 \in (0,1)$ *are defined in Assumption 1. In particular, if* $m \ge 2T$, *then*

$$\frac{m}{2T} < \mathrm{E}[N^{(l)}_s] \le m \max\{\zeta_1, 1-\zeta_2\}. \tag{3.4}$$

Remark 5. Observe a parallel between this lemma and the minorization property described in texts such as Meyn and Tweedie (2012, chapter 5.1.1), Rosenthal (1995), etc. In particular, when $m = k = 1$ and D_0 is uniform over χ, this lemma recovers the minorization

condition described in Meyn and Tweedie (2012, equation (5.3)) for a uniform measure. Furthermore, taking summation over all l and s in the lower bound, $\sum_{s,l}\frac{m}{2T} < \sum_{s,l}\mathrm{E}[N_s^{(l)}] = \mathrm{E}[\sum_{s,l}N_s^{(l)}] = m$. Therefore, $dk\frac{m}{2T} < m$, which, in turn, implies that

$$T > \frac{dk}{2}. \tag{3.5}$$

Lemma 2 provides upper and lower bounds for $\mathbb{E}[N_s^{(l)}]$. As mentioned, our next objective is to control the deviations of $N_s^{(l)}$ from its expectation. For that, we require the following two assumptions on the decay of $\overline{\eta}$-mixing coefficients of $\{X_i, a_i\}$.

Assumption 3 ($\overline{\eta}$-Mixing). *There exists a constant $\mathbb{C}_\Delta > 1$ independent of m such that*

$$\|\Delta_m\| := \max_{1\le i\le m}(1 + \overline{\eta}_{i,i+1} + \overline{\eta}_{i,i+2} + \dots \overline{\eta}_{i,m}) \le \mathbb{C}_\Delta.$$

Assumption 4 (Exponential $\overline{\eta}$-Mixing). *There exists a constant $\mathbb{C}_\Delta > 1$ independent of m such that*

$$\overline{\eta}_{i,j} \le \exp\left(-(j-i)\log\left(\frac{\mathbb{C}_\Delta}{\mathbb{C}_\Delta - 1}\right)\right).$$

A standard assumption in the off-line RL literature is the finiteness of the clipped concentrability coefficient defined in Li et al. (2022a) as

$$C^*_{clipped} := \max_{i,s,l}\frac{\inf\{\mathbb{P}(X_i = s, a_i^{(o)} = l), d^{-1}\}}{\mathbb{P}(X_i = s, a_i^{(\ell)} = l)},$$

where $a_i^{(o)}$ and $a_i^{(\ell)}$ are controls generated by the optimal and logging policies, respectively. Our Assumptions 1, 2, and 4 are more general: consider a controlled Markov chain in which $l \in \{1,2\}$, $M^{(1)}$ and $M^{(2)}$ are positive stochastic matrices, and $a_i^{(o)} = i \mod 2$ and $a_i^{(\ell)} = (i+1) \mod 2$. It is easy to see that, in this case, $C^*_{clipped} = \infty$. However, without any assumption on the optimal policy, we can still recover a sample complexity of learning the transition matrices (see Online Proposition 9) and the corresponding optimal policy value (Theorem 4).

It is obvious that, if Assumption 4 is satisfied, so is Assumption 3 with the same constant. It also follows from Lemma 1 that, if the $\overline{\eta}$-mixing coefficients satisfy the previous assumptions, so does the ϕ-mixing coefficients with appropriately adjusted constants. Depending on which of the assumptions we make, we have the following increasingly strong concentration inequalities. First, Proposition 2 provides a Hoeffding bound on the tails of the count statistics $N_s^{(l)}$.

Proposition 2. *Consider a sample $\{(X_0,a_0),\dots,(X_m,a_m)\}$ from a controlled Markov chain that satisfies Assumptions 1–3. Let $N_s^{(l)}$ be the number of visits to state–control pair (s, l) as defined in Equation (1.1). Then, for all integers $n_{high,s} > \mathbb{E}[N_s^{(l)}] > n_{low,s}$, we have*

$$\mathbb{P}(N_s^{(l)} \notin [n_{low,s}, n_{high,s}]) \le 2\exp\left(-\frac{(n_{low,s} - \frac{m}{2T})^2}{2m(\mathbb{C}_\Delta)^2}\right) + 2\exp\left(-\frac{(n_{high,s} - m\max\{\zeta_1, 1-\zeta_2\})^2}{2m(\mathbb{C}_\Delta)^2}\right).$$

Proof. The proof of this proposition follows from the fact that

$$\begin{aligned}\mathbb{P}(N_s^{(l)} \notin [n_{low,s}, n_{high,s}]) &= \mathbb{P}(N_s^{(l)} - \mathbb{E}[N_s^{(l)}] < n_{low,s} - \mathbb{E}[N_s^{(l)}]) \\ &\quad + \mathbb{P}(N_s^{(l)} - \mathbb{E}[N_s^{(l)}] > n_{high,s} - \mathbb{E}[N_s^{(l)}]),\end{aligned}$$

and then applying Assumption 3 on Lemma 6 from Online Section B. □

Next, define $\rho_s^{(l)} := \sup_{1\le i\le m}\mathbb{P}(X_i = s, a_i = l)$. Then, under Assumptions 1, 2, and 4, Proposition 3 produces a Bernstein inequality for controlling the tail probability of $N_s^{(l)}$.

Proposition 3. *Consider a sample $\{(X_0,a_0),\dots,(X_m,a_m)\}$ from a controlled Markov chain that satisfies Assumptions 1, 2, and 4. Let $N_s^{(l)}$ be the number of visits to state–control pair (s, l) as defined in Equation (1.1). Then, there exists a positive constant $\mathbb{C}_{pel}$ depending only upon $\mathbb{C}_\Delta$ such that, for all integers $n_{low,s} < \mathbb{E}[N_s^{(l)}] < n_{high,s}$, we have*

$$\mathbb{P}(N_s^{(l)} \notin [n_{low,s}, n_{high,s}]) \le 2\exp\left(-\frac{\mathbb{C}_{pel}\left(n_{low,s} - \frac{m}{2T}\right)^2}{4m\mathbb{C}_\Delta\rho_s^{(l)} + 1 + \left(\frac{m}{2T} - n_{low,s}\right)(\log m)^2}\right) + 2\exp\left(-\frac{\mathbb{C}_{pel}(n_{high,s} - m\zeta^{(\max)})^2}{4m\mathbb{C}_\Delta\rho_s^{(l)} + 1 + (n_{high,s} - m\zeta^{(\max)})(\log m)^2}\right)$$

where $\zeta^{(\max)} := \max\{\zeta_1, 1-\zeta_2\}$.

Although Proposition 3 requires a stronger assumption versus Proposition 2 (geometric versus arithmetic mixing), it provides a tighter concentration that can be used to derive a minimax sample complexity. It is proved similarly to Proposition 2 but by using Lemma 8 (also found in Online Section B) instead of Lemma 6. In many practical examples, $\mathbb{C}_{pel}$ is a universal constant. We discuss this in greater detail in the remark following Lemma 12 in Online Section E.

We can now state our first theorem regarding the sample complexity of estimating the transition probability matrices of a controlled Markov chain.

Theorem 1. *Consider a sample $\{(X_0,a_0),\dots,(X_m,a_m)\}$ from a controlled Markov chain that satisfies Assumptions*

1–3. *Let* $\{\hat{M}^{(l)}_{s,t} : l \in \mathbb{I}\}$ *be the empirical estimators as defined in Equation* (1.2) *with* $\hat{M}^{(l)}$ *being the corresponding estimated transition matrix. There exists a universal constant* $c > 1$ *such that, for any* $\varepsilon > 0$, *and* $\delta \in (0,1)$, *and with* $d = |\chi|$ *and* $k = |\mathbb{I}|$, *if it holds that*

$$m > c \max\left\{ \frac{T}{\varepsilon^2}\log\left(\frac{dkT}{\varepsilon^2\delta}\right), \mathbb{C}_\Delta^2 \max\left\{T^2, \frac{1}{(1-\max\{\zeta_1, 1-\zeta_2\})^2}\right\}\log\left(\frac{dk}{\delta}\right)\right\},$$

and then, the empirical estimator satisfies

$$\mathbb{P}\left(\sup_{l\in\mathbb{I}} \|\hat{M}^{(l)} - M^{(l)}\|_\infty > \varepsilon\right) < \delta. \tag{3.6}$$

As we see in Theorem 1, assuming that the mixing coefficients are summable (Assumption 3) allows us to compute the sample complexity. However, in Theorem 2, we see that, if we further assume the mixing coefficients to be geometrically decaying (Assumption 4), then we have a reduced sample complexity that is also minimax.

3.1. Sketch of Proof of Theorem 1

Step 1: As in Equation (3.1), let

$$\hat{M}^{(l)}(s,\cdot) = (\hat{M}^{(l)}_{s,1}, \hat{M}^{(l)}_{s,2}, \ldots, \hat{M}^{(l)}_{s,d}), \text{ and}$$
$$M^{(l)}(s,\cdot) = (M^{(l)}_{s,1}, M^{(l)}_{s,2}, \ldots, M^{(l)}_{s,d}). \tag{3.7}$$

By an application of the union bound, we get $\mathbb{P}(\sup_{l\in\mathbb{I}} \|\hat{M}^{(l)} - M^{(l)}\|_\infty > \varepsilon) \le \sum_{l\in\mathbb{I}}\sum_{s\in\chi} \mathbb{P}(\|\hat{M}^{(l)}(s,\cdot) - M^{(l)}(s,\cdot)\|_1 > \varepsilon)$.

Step 2: For each $s \in \chi$ and $l \in \mathbb{I}$, we use the law of total probability (Gut and Gut 2005, proposition 4.1) to decompose $\mathbb{P}(\|\hat{M}^{(l)}(s,\cdot) - M^{(l)}(s,\cdot)\|_1 > \varepsilon)$ into a high-probability region and a low-probability region. To be precise, for two integers $n_{high,s}$ and $n_{low,s}$ chosen appropriately by Lemma 2, we write

$$\sum_{s,l} \mathbb{P}(\|\hat{M}^{(l)}(s,\cdot) - M^{(l)}(s,\cdot)\|_1 > \varepsilon)$$
$$\le \sum_{s,l}\sum_{n=n_{low,s}}^{n_{high,s}} \mathbb{P}(\|\hat{M}^{(l)}(s,\cdot) - M^{(l)}(s,\cdot)\|_1 > \varepsilon, N^{(l)}_s = n) + \sum_{s,l} \mathbb{P}(N^{(l)}_s \notin [n_{low,s}, n_{high,s}]).$$

Step 3: In this step, we observe that, if $m > \max\left\{\frac{d}{\varepsilon^2(1+\max\{\zeta_1,1-\zeta_2\})}, \frac{64T}{\varepsilon^2}\log\left(\frac{6dk}{\delta}\right)\right\}$, Proposition 1 gives us $\sum_{s,l}\sum_{n=n_{low,s}}^{n_{high,s}} \mathbb{P}(\|\hat{M}^{(l)}(s,\cdot) - M^{(l)}(s,\cdot)\|_1 > \varepsilon, N^{(l)}_s = n) \le \delta/3$.

Step 4: In this step, we use Proposition 2 to upper bound $\sum_{s,l}\mathbb{P}(N^{(l)}_s \notin [n_{low,s}, n_{high,s}])$:

$$\sum_{s,l} \mathbb{P}(N^{(l)}_s \notin [n_{low,s}, n_{high,s}]) \le dk\left(2\exp\left(-\frac{\left(n_{low,s} - \frac{m}{2T}\right)^2}{2m(\mathbb{C}_\Delta)^2}\right) + 2\exp\left(-\frac{(n_{high,s} - m\max\{\zeta_1, 1-\zeta_2\})^2}{2m(\mathbb{C}_\Delta)^2}\right)\right).$$

It follows that, for a universal constant c, as long as $m > c\max\left\{\mathbb{C}_\Delta^2\log\left(\frac{dk}{\delta}\right)\max\left\{T^2, \frac{1}{(1-\max\{\zeta_1,1-\zeta_2\})^2}\right\}\right\}$, we have $\sum_{s,l}\mathbb{P}(N^{(l)}_s \notin [n_{low,s}, n_{high,s}]) \le 2\delta/3$ completing the sketch (details in Online Appendix G.1).

3.2. Minimax Sample Complexity

In Theorem 2, we show that, under the extra assumption of geometric mixing, our estimator is minimax optimal. Before proceeding to Theorem 2, we introduce some notation. Consider a sample $\{(X_0, a_0), \ldots, (X_m, a_m)\}$ from a controlled Markov chain that satisfies Assumptions 1, 2, and 4. Let $\rho_\star := \sup_{s,l}\sup_{1\le i\le m}\mathbb{P}(X_i = s, a_i = l)$, and with $\mathbb{C}_{pel}$ as in Proposition 3, define

$$\mathbb{C}_\zeta := \frac{8(2\mathbb{C}_\Delta\rho_\star(1-\zeta^{(\max)})^{-2} + (1-\zeta^{(\max)})^{-1})}{\mathbb{C}_{pel}},$$
$$\mathbb{C}_T := \frac{64(\mathbb{C}_\Delta\rho_\star T^2 + 2T)}{\mathbb{C}_{pel}},$$
$$\mathbb{C}_{T,\delta} := \mathbb{C}_T\log\left(\frac{6dk}{\delta}\right), \quad \mathbb{C}_{\zeta,\delta} := \mathbb{C}_\zeta\log\left(\frac{6dk}{\delta}\right).$$

Theorem 2. *Consider the setting of Theorem 1 and suppose that Assumptions 1, 2, and 4 are satisfied and let* $\rho_\star = \max_{s,l}\rho^{(l)}_s$. *Then, there exists a universal constant* $c > 1$ *such that, if*

$$m > c\max\left\{\frac{4d}{\varepsilon^2(1+\max\{\zeta_1,1-\zeta_2\})}, \mathbb{C}_{T,\delta}(\log\mathbb{C}_{T,\delta})^2, \mathbb{C}_{\zeta,\delta}(\log\mathbb{C}_{\zeta,\delta})^2\right\},$$

then the empirical estimator satisfies $\mathbb{P}(\sup_{l\in\mathbb{I}}\|\hat{M}^{(l)} - M^{(l)}\|_\infty > \varepsilon) < \delta$ *for all* $\varepsilon, \delta > 0$ *and is minimax up to* log *and* log log *terms whenever* $0 < \varepsilon < 1/32$.

We point out that this result differs from the previous one by a factor of $\mathbb{C}_\Delta\rho_\star/\mathbb{C}_{pel}$. In Online Section E, we present an example in which $\rho_\star$ is $O(1/T)$ and $\mathbb{C}_{pel}$ is independent of T, therefore assuming exponential mixing improves the sample complexity by a factor of $1/T$ and is minimax optimal.

3.3. Sketch of Proof of Theorem 2

The proof of the theorem is divided into two parts: (1) the sample complexity and (2) the minimaxity. The

proof of sample complexity proceeds similarly to the proof of Theorem 1. The key difference is in step 4, in which, instead of using Proposition 2, we use Proposition 3. The intuition is to use a tighter Chernoff concentration inequality that is available for exponentially mixing random variables instead of a weaker Hoeffding's inequality. This produces a tighter sample complexity that is provably minimax. The details of the first part can be found in Online Section G.3. For this sketch, we focus on the minimaxity. Let $\mathcal{M}_{\chi,\mathbb{I}}$ be the class of all controlled Markov chain on state space χ with control space $\mathbb{I}$. For two collections of stochastic matrices $\mathbb{M}_1 := \{M_1^{(l)}\}_{l\in\mathbb{I}}, \mathbb{M}_2 := \{M_2^{(l)}\}_{l\in\mathbb{I}}$, define $\|\mathbb{M}_1 - \mathbb{M}_2\|_\infty^* := \sup_{l\in\mathbb{I}}\|M_1^{(l)} - M_2^{(l)}\|_\infty$. Observe that the minimax risk satisfies

$$\begin{aligned}\mathcal{R}_m &= \inf_{\hat{\mathbb{M}}} \sup_{\mathcal{P}\in\mathcal{M}_{\chi,\mathbb{I}}} \mathbb{P}(\|\hat{\mathbb{M}} - \mathbb{M}\|_\infty^* > \varepsilon) \\ &\geq \inf_{\hat{\mathbb{M}}} \sup_{\mathcal{P}\in\mathcal{M}'} \mathbb{P}(\|\hat{\mathbb{M}} - \mathbb{M}\|_\infty^* > \varepsilon),\end{aligned}$$

for any subclass of controlled Markov chains $\mathcal{M}' \subset \mathcal{M}_{\chi,\mathbb{I}}$ and any estimation procedure, $\hat{\mathbb{M}}$. The rest of the proof proceeds through two cases by constructing appropriate subclasses $\mathcal{M}'$. The motivation behind these choices are based on the fact that the uniform distribution is the least favorable choice for estimation (Brandwein and Strawderman 1980, Lehmann and Casella 1998, van Eeden 2006, Fourdrinier et al. 2013). We make these examples explicit in Online Section E.

Case I:

$$\left(m < \frac{8d}{\varepsilon^2(1+\max\{\zeta_1, 1-\zeta_2\})}\right).$$

Here, we choose a class of controlled Markov chains with controls distributed uniformly over $\mathbb{I}$ and transition matrices, for vectors $\sigma = \left(\sigma_1, \ldots, \sigma_{\frac{d}{2}}\right) \in \{-1,1\}^{\frac{d}{2}}$, given by

$$M_\sigma = \begin{pmatrix} \frac{1-p_\star}{d} & \cdots & \frac{1-p_\star}{d} & p_\star \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1-p_\star}{d} & \cdots & \frac{1-p_\star}{d} & p_\star \\ \frac{1-p_\star+16\sigma_1\varepsilon}{d} & \cdots & \frac{1-p_\star-16\sigma_{\frac{d}{2}}\varepsilon}{d} & p_\star \end{pmatrix}.$$

We then use Tsybakov's reduction method to lower bound $\inf_{\hat{\mathbb{M}}}\sup_{\mathcal{P}\in\mathcal{M}'}\mathbb{P}(\|\hat{\mathbb{M}} - \mathbb{M}\|_\infty^* > \varepsilon)$ for our chosen subclass of controlled Markov chains.

Case II:

$$m < (2\,\mathbb{C}_{T,\delta}(\log \mathbb{C}_{T,\delta})^2, 2\,\mathbb{C}_{\zeta,\delta}(\log \mathbb{C}_{\zeta,\delta})^2).$$

Step 1: For this case, we set $\mathcal{M}'$ to be a class of controlled Markov chains with controls and transition probability matrices described in Online Section E.

Step 2: We then use Tsybakov's reduction method (Tsybakov 2009, chapter 2.2) to observe that, for any random variable $\mathbb{T}$,

$$\mathcal{R}_m \geq \inf_{\hat{\mathbb{M}}} \sup_{\mathcal{P}\in\mathcal{M}'} \mathbb{P}(\|\hat{\mathbb{M}} - \mathbb{M}\|_\infty^* > \varepsilon \mid \mathbb{T} > m) \times \mathbb{P}(\mathbb{T} > m).$$

$\mathbb{T}$ is chosen to be an appropriate touring time (the time to visit sufficiently many state–control pairs).

Step 3: We then prove that $\mathbb{P}(\mathbb{T} > m)$ is bounded away from zero as long as $m < 2\,\mathbb{C}_{T,\delta}(\log \mathbb{C}_{T,\delta})$.

Step 4: We then argue that, whenever $\mathbb{T} > m$, there exists a state–control pair s_0, l_0 such that $N_{s_0}^{(l_0)} = 0$.

Step 5: If $N_{s_0}^{(l_0)} = 0$, so is $N_{s_0,t}^{(l_0)} = 0$ for all $t \in \chi$. This proves that there is a uniform error to estimate $M_{s_0,t}^{(l_0)}$, which proves our claim.

4. Applications

We first briefly discuss how Assumptions 2–4, can be reduced to simpler assumptions.

4.1. Reduction of Assumptions

4.1.1. Reduction of Return Times. First, consider the assumption on return times introduced in Assumption 2. We call a sequence of random variables $\{Z_i\}_{i\geq 0}$ a jth order Markov chain if, for all i, the conditional distribution of $Z_\infty, \ldots, Z_i$ satisfies $\mathbb{P}(Z_\infty, \ldots, Z_i | Z_{i-1}, \ldots, Z_0) = \mathbb{P}(Z_\infty, \ldots, Z_i | Z_{i-1}, \ldots, Z_{i-j})$. Observe that, if a_i is jth order Markovian, then so is the paired process (X_i, a_i). For convenience of notation, define $\tau^\dagger := \sum_{p=0}^{i-1}\tau_{s,l}^{(p)}$ and observe that $\tau_{s,l}^{(i)}$ is a function of only $X_{\tau^\dagger+1}, a_{\tau^\dagger+1}, \ldots, X_\infty, a_\infty$. It follows that

$$\sup_{s,l,i} \mathrm{E}[\tau_{s,l}^{(i)} | \mathcal{F}_{\tau^\dagger}] = \sup_{s,l,i} \mathrm{E}[\tau_{s,l}^{(i)} | \mathcal{H}_{\tau^\dagger-j}^{\tau^\dagger}]$$

almost everywhere. Moreover, if a_i is independent of time point i (also called stationary), then we have almost everywhere

$$\sup_{s,l,i} \mathrm{E}[\tau_{s,l}^{(i)} | \mathcal{H}_{\tau^\dagger-j}^{\tau^\dagger}] = \sup_{s,l} \mathrm{E}[\tau_{s,l}^{(1)} | X_0 = s, a_0 = l]. \tag{4.1}$$

4.1.2. Reduction of Mixing Coefficients. Next, we decompose the $\overline{\eta}$-mixing coefficients of the paired process $\{X_i, a_i\}$ into mixing coefficients over states and controls. We motivate this decomposition using two facts:

1. The controls of a controlled Markov chain are often chosen by the user and well behaved.

2. The mixing coefficients of the individual processes can be analyzed more directly.

We begin by defining the γ-mixing coefficients $\gamma_{p,j,i}$ for controls as the following total variation distance:

$$\gamma_{p,j,i} := \sup_{s_p, h_{i+j}^{p-1}, h_0^i} \|\mathbb{P}(a_p | X_p = s_p, \mathcal{H}_{i+j}^{p-1} = h_{i+j}^{p-1}, \mathcal{H}_0^i = h_0^i) - \mathbb{P}(a_p | X_p = s_p, \mathcal{H}_{i+j}^{p-1} = h_{i+j}^{p-1})\|_{TV}, \tag{4.2}$$

where $\mathbb{P}(X_p = s_p, \mathcal{H}_{i+j}^{p-1} = h_{i+j}^{p-1}, \mathcal{H}_0^i = h_0^i) > 0$.

Assumption 5 (Mixing of Controls). *There exists a constant $\mathbb{C} \geq 0$ such that*

$$\sup_{1 \leq i \leq \infty} \sum_{j=1}^{\infty} \sum_{p=i+j+1}^{\infty} \gamma_{p,j,i} \leq \frac{\mathbb{C}}{2}.$$

Remark 6. In the Markovian settings, when the sequence of controls a_i depend only upon X_i, $\gamma_{p,j,i} = 0$ for all p, j, i. In such case, Assumption 5 is satisfied with $\mathbb{C} = 0$. This extends to the case in which a_i depends upon j many past time points. If a_i depend only upon $X_i, a_{i-1}, X_{i-1}, \ldots, a_{i-j+1}, X_{i-j+1}$, then Assumption 5 is satisfied with $\mathbb{C} = j - 1$.

Next, we generalize the Dobrushin coefficients (Dobrushin 1956a, b; Mukhamedov 2013) for inhomogenous Markov chains to the realm of controlled Markov chains. Let $\{(X_0, a_0), \ldots, (X_m, a_m)\}$ be a collected sample. For all integers $j \geq i$, define the mixing coefficient $\overline{\theta}_{i,j}$

$$\overline{\theta}_{i,j} := \sup_{\substack{s_1, s_2 \in \mathbb{X}, l_1, l_2 \in \mathbb{L}, \\ \mathbb{P}(X_i = s_1, a_i = l_1) > 0, \\ \mathbb{P}(X_i = s_2, a_i = l_2) > 0}} \|\mathbb{P}(X_j | X_i = s_1, a_i = l_1) - \mathbb{P}(X_j | X_i = s_2, a_i = l_2)\|_{TV}, \tag{4.3}$$

such that $(s_1, l_1) \neq (s_2, l_2)$. The following assumption on $\overline{\theta}$ controls the mixing of the state process X_i.

Assumption 6 (Mixing of States). *There exists a constant $\mathbb{C}_\theta \geq 0$ such that*

$$\sup_{1 \leq i \leq \infty} \sum_{j=i+1}^{\infty} \overline{\theta}_{i,j} \leq \mathbb{C}_\theta.$$

Note that neither Assumption 5 nor 6 implies the other as the following counterexamples illustrate:

1. Let (X_i, a_i) be an inhomogenous Markov chain for which $\sup_{1 \leq i \leq \infty} \sum_{j=i+1}^{\infty} \overline{\theta}_{i,j} = \infty$. One example of such an inhomogenous Markov chain can be found in Online Lemma 14. However, because the controls are deterministic, every inhomogenous Markov chain satisfies Assumption 5. We prove this fact formally in Online Proposition 9. Therefore, this chain satisfies Assumption 5 but not Assumption 6.

2. For the second counterexample, consider a controlled Markov chain (X_i, a_i) in which the a_i's do not satisfy Assumption 5. Let X_i be independent draws from a uniform distribution over χ. It is easily seen that $\overline{\theta}_{i,j} = 0$ for this example. Therefore, this chain satisfies Assumption 6 but not Assumption 5.

Observe that the previous assumptions on the states and controls imply the summability of the weak mixing coefficients. We formalize it in the following lemma.

Lemma 3. *For any controlled Markov chain that satisfies Assumptions 5 and 6, $\|\Delta_m\| \leq \mathbb{C} + \mathbb{C}_\theta + 1$, where $\|\Delta_m\| = \max_{1 \leq i \leq m}(1 + \overline{\eta}_{i,i+1} + \overline{\eta}_{i,i+2} + \ldots \overline{\eta}_{i,m})$, and $\overline{\eta}_{i,j}$ is as defined in Equation (2.1).*

Remark 7. We remark that Theorem 1 continues to hold under the weaker Assumption 3. However, because all of our examples satisfy Assumptions 5 and 6, we can invoke Lemma 3 to prove that Assumption 3 holds. Next, we state the following two assumptions as stronger versions of Assumptions 5 and 6.

Assumption 7 (Geometric Mixing of Controls). *There exists a constant $\mathbb{C}_\star > 0$ independent of m such that, for all integers $j \geq i$, we have $\sum_{p=i+j+1}^{\infty} \gamma_{p,j,i} \leq e^{-\mathbb{C}_\star(j-i)}$.*

Assumption 8 (Geometric Mixing of States). *There exists a constant $\mathbb{C}_{\theta,\star} > 0$ independent of m such that, for all integers $j \geq i$, we have $\overline{\theta}_{i,j} \leq e^{-\mathbb{C}_{\theta,\star}(j-i)}$.*

We then get the following lemma as a counterpart to Lemma 3.

Lemma 4. *For any controlled Markov chain that satisfies Assumptions 7 and 8, there exists a positive constant $\mathbb{C}_{cof}$ independent of m such that, $\forall$ integers $j \geq i$, we have $\overline{\eta}_{i,j} \leq e^{-\mathbb{C}_{cof}(j-i)}$.*

Remark 8. It can be seen that, if Assumptions 7 and 8 are satisfied, then so are Assumptions 5 and 6 with constants $1/(1 - e^{-\mathbb{C}_\star})$ and $1/(1 - e^{-\mathbb{C}_{\theta,\star}})$, respectively. To simplify notations, we denote $1/(1 - e^{-\mathbb{C}_\star})$ by $\mathbb{C}$ and $1/(1 - e^{-\mathbb{C}_{\theta,\star}})$ by $\mathbb{C}_\theta$, respectively. Finally, observe that Assumptions 7 and 8 provide a sufficient condition for $\overline{\eta}_{i,j}$ to be geometrically decaying uniformly over m.

4.2. Controlled Markov Chains with Stationary Randomized Controls

We say that a CMC has stationary randomized controls if, for any time i, state s, control l, and history h_0^{i-1},

$$\mathbb{P}(a_i = l | X_i = s, \mathcal{H}_0^{i-1} = h_0^{i-1}) = \mathbb{P}(a_i = l | X_i = s) = \mathbb{P}(a_1 = l | X_1 = s). \tag{4.4}$$

In this section, we show that Assumptions 1, 2, 5, and 6 hold for a controlled Markov chain with stationary controls. Writing $P_s^{(l)}$ for $\mathbb{P}(a_1 = l | X_1 = s)$, the transition

probability of the joint state control pair is

$$\begin{aligned}&\mathbb{P}(X_i = t, a_i = l | X_{i-1} = s, a_{i-1} = l') \\ &\quad = \mathbb{P}(X_i = t | X_{i-1} = s, a_{i-1} = l)\mathbb{P}(a_i = l | X_i = t) \\ &\quad = M^{(l)}_{s,t} \times P^{(l)}_t.\end{aligned}$$

The state–control pair is a time-homogeneous Markov chain with transition probabilities given by $M^{(l)}_{s,t} \times P^{(l)}_t$. Our goal is to estimate the transition probabilities $M^{(l)}_{s,t}$. Assume that $M^{(l)}$ is an aperiodic and irreducible (ergodic) transition probability matrix for all $l \in \mathbb{I}$. Then, we have the following proposition.

Proposition 4. *The paired process $\{(X_0, a_0), \ldots, (X_m, a_m)\}$ is a uniformly ergodic Markov chain.*

By verifying the aperiodicity and irreducibility of the paired process, the proof of Proposition 4 follows readily from Meyn and Tweedie (2012, theorem 16.0.2). Let ν be the invariant distribution of this Markov chain with $\nu_{(s,l)}$ being invariant probability corresponding to (s, l). The following proposition proves that $\{(X_i, a_i)\}$ satisfies Assumptions 1, 2, and 4. Its proof can be found in Online Section H.1.

Proposition 5. *Let $\{(X_0, a_0), \ldots, (X_m, a_m)\}$ be a sample from a controlled Markov chain with $d = |\chi|$, $k = |\mathbb{I}|$, and stationary randomized controls. Fix $\varepsilon > 0$, and $\delta \in (0, 1)$. Then, there exists a universal constant $c > 0$ and a constant $\mathbb{C}_\theta > 0$ such Theorem 1 is satisfied with $T = \sup_{s,l} 1/\nu_{s,l}$, $\zeta_2 = P_{min}$ $\zeta_1 = 1 - (k-1)P_{min}$ and $\mathbb{C}_\theta$. Moreover, if $D_0 = \nu$, then $\zeta_1 = \zeta_2 = 1/T$ satisfies Assumption 1.*

4.3. Controlled Markov Chains with Nonstationary Markov Controls

As the next example, we consider a controlled Markov chain with a nonstationary control process. A controlled Markov chain is said to have nonstationary Markov controls if, for any time period i, state s, control l, and sample history $\hbar_0^{i-1}$,

$$\mathbb{P}(a_i = l | X_i = s, \mathcal{H}_0^{i-1} = \hbar_0^{i-1}) = \mathbb{P}(a_i = l | X_i = s).$$

Observe that we allow the law of the control sequence to depend upon the time step i. For convenience, we refer to this as a Markov control. We can write the transition probability of the state-control pair as

$$\begin{aligned}&\mathbb{P}(X_i = t, a_i = l' | X_{i-1} = s, a_{i-1} = l) \\ &\quad = \mathbb{P}(X_i = t | X_{i-1} = s, a_{i-1} = l)\mathbb{P}(a_i = l' | X_i = t) \\ &\quad = M^{(l)}_{s,t} \times P^{(i)}_{t,l'}.\end{aligned}$$

It is straightforward to see that the state–control pair is a time-inhomogeneous Markov chain with transition probabilities given by $M^{(l)}_{s,t} \times P^{(i)}_{s,l'}$. Our goal is to estimate the transition probabilities $\mathbb{P}(X_i = t | X_{i-1} = s, a_{i-1} = l')$. We proceed by making assumptions on the return times of the controls.

Definition 3. Define $\tau^{(i,\star,j)}_{s,l}$ to be the time between the $j-1$th and jth visit to control l after visiting state–control pair s, l for the ith time. For ease of notation, denote $\sum_{k=1}^{i} \tau^{(k)}_{s,l} + \sum_{k=1}^{j-1} \tau^{(i,\star,k)}_{s,l} = \tau_\star$. Then, $\tau^{(i,\star,j)}_{s,l}$ is recursively defined as $\tau^{(i,\star,j)}_{s,l} := \min\{n : (a_{\tau_\star + n} = l), a_j \neq l \ \forall \tau_\star < j < \tau_\star + n\}$.

Next, we make some simplifying assumptions on $\tau^{(i,\star,j)}_{s,l}$ and $M^{(l)}$.

Assumption 9.

1. *For some constant $T_\star > 0$, $\sup_{i \geq 0} \mathbb{E}[\tau^{(i,\star,j)}_{s,l} | \mathcal{F}_{\sum_{p=1}^{i-1} \tau^{(p)}_{s,l} + \sum_{p=1}^{j-1} \tau^{(i,\star,p)}_{s,l}}] < T_\star$ almost everywhere.*

2. *There exists M_{min} and M_{max} such that, for all $s, t \in \chi$ and $l \in \mathbb{I}$*

$$0 < M_{min} \leq M^{(l)}_{s,t} \leq M_{max} < 1. \tag{4.5}$$

The next lemma proves that, under this assumption, $\{(X_i, a_i)\}$ satisfies Assumption 2 and derives T.

Lemma 5. *Under the conditions of Assumption 9 for all $(i, s, l) \in \mathbb{N} \times \chi \times \mathbb{I}$ it holds almost everywhere that*

$$\begin{aligned}&\mathbb{E}[\tau^{(i)}_{s,l} | \mathcal{F}_{\sum_{p=1}^{i-1} \tau^{(p)}_{s,l}}] \\ &< \frac{T_\star M_{max}}{\max\{M_{max}, 1 - M_{min}\}(1 - \max\{M_{max}, 1 - M_{min}\})}.\end{aligned} \tag{4.6}$$

We can now state our main result about the sample complexity of a controlled Markov chain with nonstationary Markov controls. Its proof can be found in Online Section H.3.

Proposition 6. *Let $\{(X_0, a_0), \ldots, (X_m, a_m)\}$ be a sample from a controlled Markov chain with nonstationary Markovian controls satisfying Assumption 9. Fix $\varepsilon > 0$, and $\delta \in (0, 1)$. Then, Assumption 1 holds with $\zeta_1 = M_{max}$, $\zeta_2 = M_{min}$, $T = \frac{T_\star M_{max}}{\max\{M_{max}, 1 - M_{min}\}(1 - \max\{M_{max}, 1 - M_{min}\})}$ and $\mathbb{C}_\theta := \frac{1}{dM_{min}}$.*

We next illustrate how one can use Theorem 1 to derive a sample complexity of learning the demand distribution of an inventory control problem.

4.4. Sample Complexity of Estimating Transitions of a $(\mathbf{s}, \mathbf{S})$-Inventory Control Problem

In this section, we consider estimating the transition probability matrices for a finite state inventory control problem. Here, the Markov state X_i is the inventory at time i, and the controls are such that the inventory is always brought up to a level $\mathbf{S}$ whenever it falls below a level $\mathbf{s}$. Assume that $\mathbf{S} > 2\mathbf{s}$. Then, with $b_i = l \in \{0, \ldots, \mathbf{s}\}$, the demand at time i (having probability $\mathcal{P}(b_i = l) = p_l$). The

system has the following dynamics:

$$X_{i+1} = X_i + (\mathbf{S} - X_i)a_i(X_i) - b_i$$
$$\text{where} \quad a_i(X_i) = \begin{cases} 1 \text{ if } X_i < \mathbf{s} \\ 0 \text{ if } X_i \geq \mathbf{s}. \end{cases}$$

Note that we assume $b_i \in \{0, \dots, \mathbf{s}\}$, resulting in a system without backlog. We do this for simplicity though we can easily relax this with some simple if tedious bookkeeping.

Observe that there are two controls $\{0, 1\}$ and $\mathbb{P}(X_i \geq \mathbf{s}, a_i = 1) = \mathbb{P}(X_i < \mathbf{s}, a_i = 0) = 0$. Thus, we only need to estimate the transition probabilities $\mathbb{P}(X_{i+1} = t | X_i = s, a_i = l)$ if either $\{s < \mathbf{s}, l = 1\}$ or $\{s \geq \mathbf{s}, l = 0\}$. Therefore, the combined state-space for $\{X_i, a_i\}$ is $\{(0,1),(1,1),\dots,(\mathbf{s}-1,1),(\mathbf{s},0),(\mathbf{s}+1,0)\dots,(\mathbf{S},0)\}$. This is a Markov chain with one-step transition probability matrix M, whose $(s,l_1),(t,l_2)$th element $M_{(s,l_1),(t,l_2)} = \mathbb{P}(X_{i+1} = t, a_{i+1} = l_2 | X_i = s, a_i = l_1)$ is

$$M_{(s,l_1),(t,l_2)} = \begin{cases} p_{\mathbf{S}-t} \text{ if } t \in \{\mathbf{S}-\mathbf{s}, \dots \mathbf{S}\}, s \in \{0, \dots, \mathbf{s}-1\}, \\ \qquad\qquad\qquad\qquad l_1 = 1, l_2 = 0 \\ p_{t-\mathbf{s}} \text{ if } s \in \{\mathbf{s}, \dots, \mathbf{S}\}, t \in \{s-\mathbf{s}, \dots, \mathbf{s}\}, \\ \qquad\qquad\qquad\qquad l_1 = 0, l_2 = 0 \\ 0 \text{ otherwise.} \end{cases} \tag{4.7}$$

This leads us to the following proposition whose proof can be found in Online Section H.4.

Proposition 7. *Let $\{(X_i, a_i)\}$ be a $(\mathbf{s}, \mathbf{S})$ inventory process with $\mathbf{S} > 2\mathbf{s}$. Then, Theorem 1 applies with constants $T = (\mathbf{S} - 2\mathbf{s})/M_{min}^{\mathbf{S}-2\mathbf{s}}$, $\mathbb{C}_\Delta = (d/M_{min}^{\mathbf{S}-2\mathbf{s}})^{\lfloor (j-i)/(\mathbf{S}-2\mathbf{s}) \rfloor}$, $\zeta_2 = M_{min}^{\mathbf{S}-2\mathbf{s}}$, and $\zeta_1 = 1 - M_{min}^{\mathbf{S}-2\mathbf{s}}$.*

Remark 9. Observe that assuming $\mathbf{S} > 2\mathbf{s}$ entails a loss of generality but lets us reuse our earlier results. This assumption can be further generalized with more calculations, and we leave it to the interested reader.

The examples in Sections 4.2 and 4.4 can also be viewed as stationary Markov chains, and therefore, the sample complexity results may also be recovered using the analysis in Wolfer and Kontorovich (2021). However, a naive application of Wolfer and Kontorovich (2021) to more general CMCs will yield suboptimal sample complexity results. Furthermore, Wolfer and Kontorovich (2021) can only be applied to stationary, ergodic chains, whereas our theory is applicable in much greater generality as highlighted by the examples in Section 4.3 and Online Appendix D. This also opens up the interesting possibility of designing controls to estimate the transition matrices faster than would be possible by a stationary ergodic Markov chain. The following section provides one such example.

4.5. Example: Designing Controls for Faster Learning of Transition Matrices

Let $M^{(1)}$ and $M^{(2)}$ be two stationary ergodic transition matrices. We assume that $M^{(1)}$ and $M^{(2)}$ have states t_1 and t_2 that are difficult to reach in comparison with other states. Formally, we assume the following.

Assumption 10. *Let $t_1 \neq t_2$ be states in χ, and $\iota_1 := \sum_{s \in \chi} M^{(1)}_{s,t_1}$, and $\iota_2 := \sum_{s \in \chi} M^{(2)}_{s,t_2}$. We assume $\exists M_{min} < 1$ such that $M^{(l)}_{s,t} \geq M_{min} > 8 \min\{\iota_1, \iota_2\}$ for any $s \in \chi$ and any $(t,l) \notin \{(t_1,1),(t_2,2)\}$.*

Remark 10. We remark that this assumption simplifies the calculations, and the result holds true much more generally as demonstrated by the empirical findings.

For such transition matrices, the following proposition demonstrates that (modulo Assumption 10) the sample complexity of individually estimating such transition matrices can be larger than estimating them simultaneously as a CMC with a predesigned control sequence. Its proof can be found in Online Section H.5.

Proposition 8. *Let $m^{(1)}$ and $m^{(2)}$ be, respectively, the sample complexity (ignoring* log *terms) of learning the transition probabilities of a Markov chain with the transition probability matrix $M^{(1)}$ and $M^{(2)}$. Then, one can construct a controlled Markov chain with transition matrices $M^{(1)}$ and $M^{(2)}$ with deterministic controls $\{a_i\}$ such that the sample complexity (ignoring* log *terms) of learning this controlled Markov chain $m^{(c)}$ satisfies $m^{(c)} < (m^{(1)} + m^{(2)})/2$.*

Remark 11. From the explicit expressions (barring log-terms)

$$m^{(1)} = \max\left\{\frac{1}{\iota_1 \varepsilon^2}\log\left(\frac{d}{\iota_1 \varepsilon^2 \delta}\right), \frac{1}{((d-2)M_{min}\iota_1)^2}\right\}$$
$$m^{(2)} = \max\left\{\frac{1}{\iota_2 \varepsilon^2}\log\left(\frac{d}{\iota_2 \varepsilon^2 \delta}\right), \frac{2}{((d-2)M_{min}\iota_2)^2}\right\}$$
$$m^{(c)} = \max\left\{\frac{4}{M_{min}\varepsilon^2}\log\left(\frac{4d}{M_{min}\varepsilon^2\delta}\right), \frac{32}{(d-2)^2 M_{min}^4}\right\},$$

which is found in the proof of Proposition 8, one can see that $m^{(c)}$ is independent of either ι_1 or ι_2. The rest of the proof follows by straightforward algebra.

As a numerical illustration, consider a controlled Markov chain with corresponding transition matrices $M^{(1)}$ and $M^{(2)}$ that take the form of Equation (4.7), determined solely by two (deterministic) probability vectors $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$. We can interpret this as corresponding to a time-inhomogeneous $(\mathbf{S}, \mathbf{s})$ inventory system, wherein the demand distribution changes as a function of a low price $\mathbf{p}^{(1)}$ and a high price $\mathbf{p}^{(2)}$ control set by the inventory manager.

Now, setting $(\mathbf{S}, \mathbf{s}) = (6, 2)$, $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ can be constructed as 3×1 dimensional vectors as follows:

generate the low price vector $\mathbf{p}^{(1)}$ by sampling a 3×1 random vector of independent uniform $[0,1]$ marginals, divide the first coordinate by 1,000, and renormalize the vector to sum up to one. The high price vector is constructed similarly by multiplying the first coordinate by 100. Roughly speaking, $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ correspond to zero sales having 0.1% and 90% probabilities.

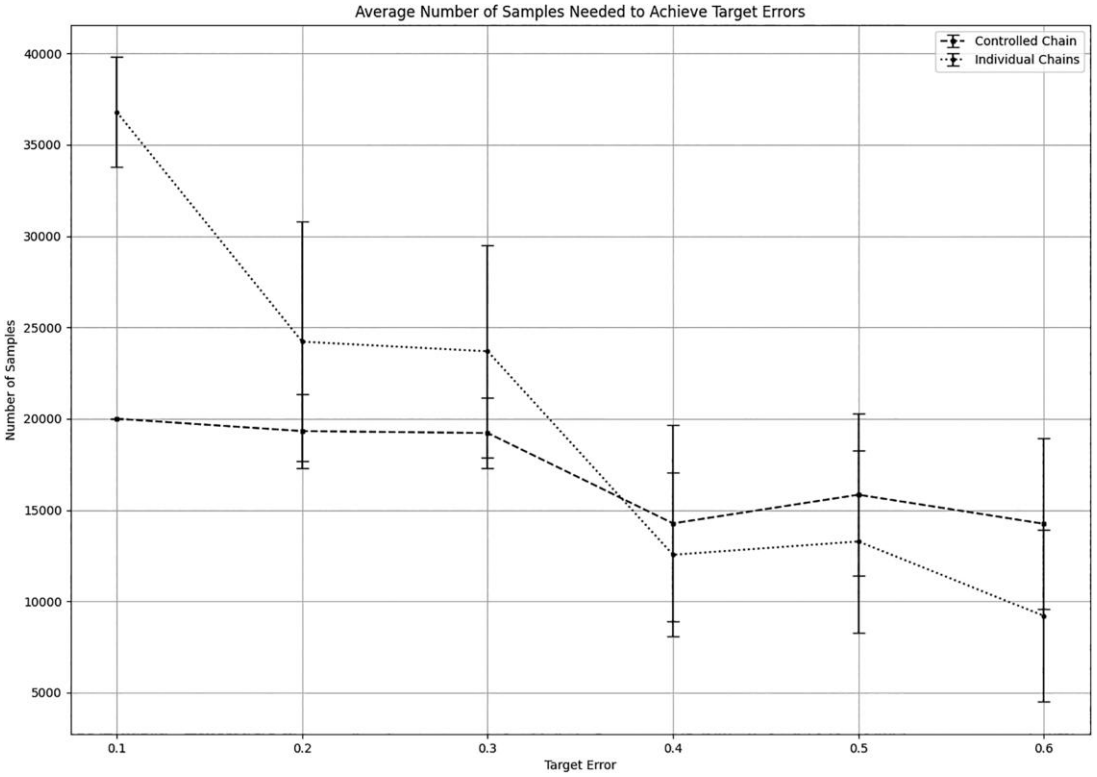
We calculate the number of samples required to achieve a target error when the price/control sequence alternates between the low price $\mathbf{p}^{(1)}$ and the high $\mathbf{p}^{(2)}$ and compare it with the average number of samples required to achieve the same error for the individual chains. Figure 1 displays the results with the error bars resulting from 100 repetitions. The dotted curve represents $(m^{(1)} + m^{(2)})/2$, where $m^{(1)}$ is the number of samples required by the first Markov chain to hit a target error and $m^{(2)}$ is the same for the second Markov chain. The dashed curve represents the samples required by the controlled chain to hit the target error $m^{(c)}$. As we target smaller and smaller errors, the average number of samples required to achieve a certain error for the individual Markov chains exceeds the number of samples required to learn the controlled Markov chain. This empirically validates our claim.

5. Sample Complexity of Policy Evaluation and Optimal Policy Recovery

Optimal policy recovery (OPR) is a key problem in offline RL settings, wherein one wishes to identify the policy that maximizes the value function given a sequence of states and controls generated by an unknown logging policy and unknown transition probabilities. An allied problem is OPE, in which we use the state–control sequence to estimate the value function of an arbitrary policy. In this section, we demonstrate how our previous results can be used to provide a high-probability bound for recovering the optimal policy. We first use Theorem 1 to provide a high-probability bound for OPE and then extend that to a corresponding result for OPR.

Let $\Delta(\mathbb{I})$ denote the probability simplex on the control space, and suppose $\pi : \mathcal{X} \to \Delta(\mathbb{I})$ is a given stationary stochastic policy. Let the matrices M and Π represent the probabilities of the next state and action, respectively, given the current state and action. Thus, $M = [M^{(1)}, \ldots, M^{(k)}]^T$, and $\Pi = [\pi, \pi, \ldots, \pi]$ are $dk \times d$ dimensional matrices. Let $\tilde{g} := (\tilde{g}(x,a) : (x,a) \in \mathcal{X} \times \mathbb{I})$ be a $dk \times 1$ vector for which element $s + k(l-1)$ denotes the cost

Figure 1. The Number of Samples Required to Achieve a Target Error for Two Individual Markov Chains as Described in Section 4.5 vs. Learning Them Together as a Controlled Markov Chain



Note. As we target smaller and smaller error, the controlled Markov chain requires fewer samples and also has smaller variation compared with learning the chains individually.

associated with the state s and control l. Then, the per-stage expected reward function $g := (g(x) : x \in \chi)$ is a $d \times 1$ vector in which $g(x) = \sum_{a \in \mathbb{I}} \pi(x,a)\tilde{g}(x,a)$.

For a known discount factor $0 < \alpha_{dis} < 1$, the value function $V_\Pi := (V_\Pi(x) : x \in \mathcal{X}) \in \mathbb{R}^d$, obtained by solving the Bellman equation (Bertsekas 2011), is given by

$$V_\Pi = (I - \alpha_{dis}\Pi^T M)^{-1} g.$$

Substituting M by $\hat{M}$, we obtain the plug-in estimate $\hat{V}_\Pi = (I - \alpha_{dis}\Pi^T \hat{M})^{-1} g$. The next theorem provides a sample complexity bound on estimating the value function V.

Theorem 3. *Let $\{(X_0,a_0),\dots,(X_m,a_m)\}$ be a sample from a controlled Markov chain with stationary randomized controls. Assume that, for some $T > 0$, $\mathbb{P}(X_i = s, a_i = l) > T^{-1}$ for all s, l, i. Then, there exists a universal constant $c > 1$ such that $\mathbb{P}(\|\hat{V}_\Pi - V_\Pi\|_\infty > \varepsilon) < \delta$ if*

$$m > c \max\left\{\frac{T_\alpha}{\varepsilon^2}\log\left(\frac{dkT_\alpha}{\varepsilon^2\delta}\right), \mathbb{C}_\theta^2 \max\left\{T^2, \frac{1}{(1-\max\{\zeta_1,1-\zeta_2\})^2}\right\}\log\left(\frac{dk}{\delta}\right)\right\},$$

where $T_\alpha = \|g\|_1^2 d\alpha_{dis}^2 T/(1-\alpha_{dis})^4$, $\zeta_1 = \zeta_2 = T^{-1}$, and $\mathbb{C}_\theta = T/dk$.

The proof of this theorem can be found in Online Section G.4.

Remark 12. The assumption $\mathbb{P}(X_i = s, a_i = l) > T^{-1}$ can be relaxed with appropriate assumptions on the return time of X_i, a_i. In practice, this would translate to an assumption on the logging policy.

Next, one can use any method to find the optimal policy, for instance, policy iteration (Bertsekas 2011, chapter 1) or policy gradient (Sutton and Barto 2018, chapter 13). Let Π_{opt} and $\hat{\Pi}_{opt}$ denote the optimal policies for maximizing the reward function for the true and estimated transition matrices M and $\hat{M}$, respectively. The following theorem provides a sample complexity of recovering the optimal value. Its proof can be found in Online Section G.5.

Theorem 4. *Under the conditions of Theorem 3, we have $\|V_{\Pi_{opt}} - V_{\hat{\Pi}_{opt}}\|_\infty \le \frac{d\sqrt{d}\,\alpha_{dis}}{(1-\alpha_{dis})^2}\|g\|_1\varepsilon$ with probability at least $1-\delta$ if*

$$m > c \max\left\{\frac{T_\alpha}{\varepsilon^2}\log\left(\frac{dkT_\alpha}{\varepsilon^2\delta}\right), \mathbb{C}_\theta^2 \max\left\{T^2, \frac{1}{(1-\max\{\zeta_1,1-\zeta_2\})^2}\right\}\log\left(\frac{dk}{\delta}\right)\right\}. \quad (5.1)$$

Remark 13. We can make a further assumption that, for any $\varepsilon < 1/1.5$, and policy Π' such that $\|\Pi_{opt} - \Pi'\|_\infty > \varepsilon$, the value functions corresponding to Π_{opt} and Π' satisfy $\inf_{s\in\chi}\{V_{\Pi_{opt}}(s) - V_{\Pi'}(s)\} > 2\varepsilon$. Then, it trivially follows that $\|\Pi_{opt} - \hat{\Pi}_{opt}\|_\infty < \varepsilon$, giving us a theoretical guarantee for recovering the optimal policy with high probability.

One can compare the dependence of the sample complexity (Equation (5.1)) on the discount factor α_{dis} to those in Li et al. (2022a). Our dependence is $\alpha_{dis}^2(1-\alpha_{dis})^{-4}$, which is better than $(1-\alpha_{dis})^{-3}$ in Li et al. (2022a) when $\alpha_{dis} < 0.618$. However, neither achieves the lower bound $(1-\alpha_{dis})^{-2}$ given by Agarwal et al. (2020).

6. Conclusions

In this paper, we derive exact rates of convergence for the empirical estimator of the transition probability matrices of a controlled Markov chain and use these to derive the sample complexity of achieving a desired estimation error. We tease out the exact effect of the mixing coefficients of the states and controls on the sample complexity and provide conditions under which the empirical estimator is minimax optimal. We use our sample complexity results in a number of examples, including error bounds for the value function and the optimal policy estimated from data. Below, we highlight three possible topics for future research.

6.1. Countable State Spaces

As an obvious extension to our work, consider the problem of countably infinite state and control spaces. Some work in this regard can be found in Wolfer and Kontorovich (2021) in which state spaces are countably infinite, but there are no easy extensions to the setting with countably infinite control space.

6.2. Uncountably Infinite State Space and Finite Controls

We have also found no result that derives the minimax sample complexity of estimating the transition probability distribution of a controlled Markov chain on an uncountably infinite state space. The histogram estimator is the obvious counterpart to this work. Indeed, recent studies (Sart 2014, Löffler and Picard 2021) demonstrate promising properties of the histogram in estimating the transition functions of a continuous Markov chain. But, to the best of our knowledge, the techniques do not translate to uncountable control spaces, and optimally estimating the transition probability distribution remains an open question.

6.3. Learning in the Presence of Weaker Mixing or Adversarial Controls

Although the strong mixing properties of the controls is a sufficient condition for the estimability of the transition matrices, it may not be a realistic assumption when the system dynamics are weakly mixing or adversarial

(Pinto et al. 2017). System identification under the presence of an adversary remains an interesting question that was addressed in a recent paper (Showkatbakhsh et al. 2016) using strong linearization assumptions and exponential computation times. However, this is well beyond the scope of the current work and is a direction for future study.

6.4. Instance-Dependent and Data-Dependent Learning

Our bounds are derived over an entire class of CMC models. Indeed, so long as a specific model conforms to the assumptions we have imposed over this class, our bounds are applicable. This reflects extant analyses of off-line RL (Li et al. 2022a, b). On the other hand, recent work in online RL seeks to establish instance-dependent bounds for specific models (for example, Zanette and Brunskill 2019, Mou et al. 2021, Khamaru et al. 2022). Establishing such bounds for off-line CMC identification is also an interesting open problem. Khamaru et al. (2022) also emphasize that inferential theory for RL should be data-dependent, for instance, allowing for the computation of data-dependent confidence intervals. This is an important future direction for our work as well.

Acknowledgments

I. Banerjee thanks Anamitra Chaudhuri for numerous insightful discussions and comments throughout the duration of this project. The authors thank the anonymous reviewers for their insightful comments and especially for pointing them toward the interesting application detailed in Section 4.5. Work was completed while Imon Banerjee was at the Department of Statistics, Purdue University.

References

Agarwal A, Kakade S, Yang LF (2020) Model-based reinforcement learning with a generative model is minimax optimal. *Proc. Thirty Third Conf. Learn. Theory* (PMLR, New York), 67–83.

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.

Banerjee I, Rao VA, Honnappa H (2021) PAC-Bayes bounds on variational tempered posteriors for Markov models. *Entropy* 23(3):313.

Bertsekas DP (2011) *Dynamic Programming and Optimal Control*, 3rd ed., vol. II (Athena Scientific, Belmont, MA).

Billingsley P (1961) Statistical methods in Markov chains. *Ann. Math. Statist.* 32(1):12–40.

Borkar VS (1991) *Topics in Controlled Markov Chains* (Longman, Harlow, UK).

Bradley RC (2005) Basic properties of strong mixing conditions: A survey and some open questions. *Probab. Surveys* 2:107–144.

Brandwein AC, Strawderman WE (1980) Minimax estimation of location parameters for spherically symmetric distributions with concave loss. *Ann. Statist.* 8(2):279–284.

Chen IY, Joshi S, Ghassemi M, Ranganath R (2021) Probabilistic machine learning for healthcare. *Annual Rev. Biomedical Data Sci.* 4:393–415.

Chin R, Maass AI, Ulapane N, Manzie C, Shames I, Nešić D, Rowe JE, Nakada H (2020) Active learning for linear parameter-varying system identification. *IFAC-PapersOnLine* 53(2):989–994.

Dobrushin RL (1956a) Central limit theorem for nonstationary Markov chains. I. *Theory Probab. Appl.* 1(1):65–80.

Dobrushin RL (1956b) Central limit theorem for nonstationary Markov chains. II. *Theory Probab. Appl.* 1(4):329–383.

Fourdrinier D, Mezoued F, Strawderman WE (2013) Bayes minimax estimation under power priors of location parameters for a wide class of spherically symmetric distributions. *Electronic J. Statist.* 7:717–741.

Geyer CJ (1998) Markov chain Monte Carlo lecture notes. Course Notes, Spring Quarter 80, Minneapolis.

Goldberg DA, Reiman MI, Wang Q (2021) A survey of recent progress in the asymptotic analysis of inventory systems. *Production Oper. Management* 30(6):1718–1750.

Gut A, Gut A (2005) *Probability: A Graduate Course*, vol. 5 (Springer, New York).

Hernández-Lerma O, Montes-de Oca R, Cavazos-Cadena R (1991) Recurrence conditions for Markov decision processes with Borel state space: A survey. *Ann. Oper. Res.* 28(1):29–46.

Icarte RT, Klassen T, Valenzano R, McIlraith S (2018) Using reward machines for high-level task specification and decomposition in reinforcement learning. *Proc. 35th Internat. Conf. Machine Learn.* (PMLR, New York), 2107–2116.

Jones GL (2004) On the Markov chain central limit theorem. *Probab. Surveys* 1:299–320.

Khamaru K, Xia E, Wainwright MJ, Jordan MI (2022) Instance-dependent confidence and early stopping for reinforcement learning. Preprint, submitted January 1, https://arxiv.org/abs/2201.08536.

Kidambi R, Rajeswaran A, Netrapalli P, Joachims T (2020) Morel: Model-based offline reinforcement learning. *Adv. Neural Inform. Processing Systems* 33:21810–21823.

Kontorovich LA, Ramanan K (2008) Concentration inequalities for dependent random variables via the martingale method. *Ann. Probab.* 36(6):2126–2158.

Laroche R, Tachet Des Combes R (2023) On the occupancy measure of non-Markovian policies in continuous MDPs. Krause A, Brunskill E, Cho K, Engelhardt B, Sabato, S Scarlett J, eds. *Proc. 40th Internat. Conf. Machine Learn.*, vol. 202 (PMLR, New York), 18548–18562.

Laroche R, Combes RTD, Buckman J (2022) Non-Markovian policies occupancy measures. Preprint, submitted May 27, https://arxiv.org/abs/2205.13950.

Lee HR, Lee T (2018) Markov decision process model for patient admission decision at an emergency department under a surge demand. *Flexible Services Manufacturing J.* 30(1):98–122.

Lehmann EL, Casella G (1998) *Theory of Point Estimation*, 2nd ed. (Springer, New York), 150.

Levine S, Kumar A, Tucker G, Fu J (2020) Offline reinforcement learning: Tutorial, review, and perspectives on open problems. Preprint, submitted May 4, https://arxiv.org/abs/2005.01643.

Li Y, Wang R, Yang LF (2022b) Settling the horizon-dependence of sample complexity in reinforcement learning. *2021 IEEE 62nd Annual Sympos. Foundations Computer Sci.* (IEEE, Piscataway, NJ), 965–976.

Li G, Shi L, Chen Y, Chi Y, Wei Y (2022a) Settling the sample complexity of model-based offline reinforcement learning. Preprint, submitted April 11, https://arxiv.org/abs/2204.05275.

Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng M (2020) Reinforcement learning for clinical decision support in critical care: Comprehensive review. *J. Medical Internet Res.* 22(7):e18477.

Ljung L (1987) *System Identification: Theory for the User* (Prentice-Hall, Englewood Cliffs, NJ).

Ljung L (2010) Perspectives on system identification. *Annual Rev. Control* 34(1):1–12.

Löffler M, Picard A (2021) Spectral thresholding for the estimation of Markov chain transition operators. *Electronic J. Statist.* 15(2):6281–6310.

Mania H, Jordan MI, Recht B (2020) Active learning for nonlinear system identification with guarantees. Preprint, submitted June 18, https://arxiv.org/abs/2006.10277.

Mannor S, Tsitsiklis JN (2005) On the empirical state-action frequencies in Markov decision processes under general policies. *Math. Oper. Res.* 30(3):545–561.

Meyn SP, Tweedie RL (2012) *Markov Chains and Stochastic Stability* (Springer Science & Business Media, London).

Mou W, Pananjady A, Wainwright MJ, Bartlett PL (2021) Optimal and instance-dependent guarantees for Markovian linear stochastic approximation. Preprint, submitted December 23, https://arxiv.org/abs/2112.12770.

Mukhamedov F (2013) The Dobrushin ergodicity coefficient and the ergodicity of noncommutative Markov chains. *J. Math. Anal. Appl.* 408(1):364–373.

Mutti M, De Santi R, Restelli M (2022) The importance of non-Markovianity in maximum state entropy exploration. Preprint, submitted February 7, https://arxiv.org/abs/2202.03060.

Pinto L, Davidson J, Sukthankar R, Gupta A (2017) Robust adversarial reinforcement learning. *Internat. Conf. Machine Learn.* (PMLR, New York), 2817–2826.

RayChaudhuri T, Hamey LG (1996) Active learning for nonlinear system identification and control. *IFAC Proc. Volumes* 29(1):2592–2596.

Rosenblatt-Roth M (1963) Some theorems concerning the law of large numbers for non-homogeneous Markoff chains. *Zeitschrift Wahrscheinlichkeitstheorie Verwandte Gebiete* 1(5):433–445.

Rosenblatt-Roth M (1964) Some theorems concerning the strong law of large numbers for non-homogeneous Markov chains. *Ann. Math. Statist.* 35(2):566–576.

Rosenthal JS (1995) Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* 90(430):558–566.

Sart M (2014) Estimation of the transition density of a Markov chain. *Annales l'IHP Probabilités Statistiques* 50(3):1028–1068.

Shortreed SM, Laber E, Lizotte DJ, Stroup TS, Pineau J, Murphy SA (2011) Informing sequential clinical decision-making through reinforcement learning: An empirical study. *Machine Learn.* 84(1): 109–136.

Showkatbakhsh M, Tabuada P, Diggavi S (2016) System identification in the presence of adversarial outputs. *2016 IEEE 55th Conf. Decision Control* (IEEE, Piscataway, NJ), 7177–7182.

Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).

Tangirala AK (2018) *Principles of System Identification: Theory and Practice* (CRC Press, Boca Raton, FL).

Tsybakov AB (2009) *Introduction to Nonparametric Estimation* (Springer, New York).

van Eeden C (2006) Minimax estimators and their admissibility. *Restricted Parameter Space Estimation Problems: Admissibility and Minimaxity Properties* (Springer, New York), 33–67.

Vidyasagar M, Karandikar RL (2006) A learning theory approach to system identification and stochastic adaptive control. *Probabilistic and Randomized Methods for Design Under Uncertainty* (Springer, London), 265–302.

Wang S, Si N, Blanchet J, Zhou Z (2023) On the foundation of distributionally robust reinforcement learning. Preprint, submitted November 15, https://arxiv.org/abs/2311.09018.

Wolfer G, Kontorovich A (2021) Statistical estimation of ergodic Markov chain kernel over discrete state space. *Bernoulli* 27(1):532–553.

Yakowitz S (1979) Nonparametric estimation of Markov transition functions. *Ann. Statist.* 7(3):671–679.

Yakowitz S (1989) Nonparametric density and regression estimation for Markov sequences without mixing assumptions. *J. Multivariate Anal.* 30(1):124–136.

Yan Y, Li G, Chen Y, Fan J (2022) Model-based reinforcement learning is minimax-optimal for offline zero-sum Markov games. Preprint, submitted June 8, https://arxiv.org/abs/2206.04044.

Yu C, Liu J, Nemati S, Yin G (2021) Reinforcement learning in healthcare: A survey. *ACM Comput. Surveys* 55(1):1–36.

Yu T, Thomas G, Yu L, Ermon S, Zou JY, Levine S, Finn C, Ma T (2020) MOPO: Model-based offline policy optimization. *Adv. Neural Inform. Processing Systems* 33:14129–14142.

Zanette A, Brunskill E (2019) Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *Internat. Conf. Machine Learn.* (PMLR, New York), 7304–7312.

Imon Banerjee is an Industrial Engineering and Management Sciences Alumni Fellow at Northwestern University. His current research encompasses mathematical statistics and stochastic processes with applications in reinforcement learning. More broadly, he is interested in exploring the theoretical aspects of machine learning using tools from applied probability.

Harsha Honnappa is an associate professor of industrial engineering at Purdue University. His research interests as an applied probabilist encompass stochastic modeling, optimization, and control with applications to machine learning, simulation, and statistical inference.

Vinayak Rao is an associate professor in the Department of Statistics at Purdue University. His research interests include methodological, computational, and theoretical aspects of Bayesian and nonparametric statistics.